Forthcoming in *Mind and Language.*

Philosophy's New Challenge: Experiments and Intentional Action[†]

N. Ángel Pinillos, Nick Smith, G. Shyam Nair, Peter Marchetto, Cecilea Mun[*]

ABSTRACT: Experimental philosophers have gathered impressive evidence for the surprising conclusion that philosophers' intuitions are out of step with those of the folk. As a result, many argue that philosophers' intuitions are unreliable. Focusing on the Knobe Effect, a leading finding of experimental philosophy, we defend traditional philosophy against this conclusion. Our key premise relies on experiments we conducted which indicate that judgments of the folk elicited under higher quality cognitive or epistemic conditions are more likely to resemble those of the philosopher. We end by showing how our experimental findings can help us better understand the Knobe Effect.

1. Introduction

Experimental philosophers have recently uncovered an impressive body of evidence for the surprising conclusion that philosophers' intuitions about knowledge, morality, intentional action, reference and other foundational notions are out of step with those of the public. Since these intuitions play an important role in philosophical theorizing and are also assumed to be widely shared, experimental philosophers have taken these results to call into doubt the reliability of our intuitions and thereby challenge the very foundations of traditional philosophy. In this paper, we put forward an empirical method

[†] We would like to thank the following people for their valuable help in this project: Tom Blackson, David Chalmers, Stewart Cohen, Edward Cokely, Peter de Marneffe, Max Deutsch, Adam Feltz, Shane Frederick, Mike Gifford, Carrie Gillon, Steve Goldinger, Daniel Kahneman, Andy Khoury, Joshua Knobe, Edouard Machery, Jennifer Nagel, Shaun Nichols, Michelle Saint, Jonathan Schaffer, Tamar Szabo Gendler, Jonathan Weinberg, and an anonymous referee from *Mind and Language*. We also want to thank participants at the AHRC 2009 Methodology Conference at the University of St. Andrews where parts of this paper were presented, as well as members of the National Endowment for the Humanities 2009 Summer Institute on Experimental Philosophy at the University of Utah where parts of this paper were discussed. Finally, Mr. Nair would like to thank the USC Provost's PhD Fellowship for support during part of this project.

[*] All authors at Arizona State University except Mr. Nair who is at University of Southern California and Mr. Marchetto who is at University of Massachusetts, Amherst.

for assessing the soundness of this criticism, and report on experiments we carried out that implement this method. As a case study, we look closely at one of the most famous experimental philosophy results, the Knobe Effect, and conclude that it gives us little reason to be pessimistic about the reliability of intuitions or about the prospects of traditional philosophy. Two further points fall out of our study. First, we can rule out Joshua Knobe's theoretical account of the effect. Second, we give some reasons to think that the effect is sensitive to factors concerning two distinct cognitive processes known in psychology as System 1 and System 2.

2. Experiments and Traditional Philosophy

It is difficult to imagine a serious account of some philosophical concept that is wholly indifferent to our intuitions. An ethical theory which predicts that murder is always a good thing is implausible from the get-go. It runs up against intuitions to the effect that murder in certain cases is bad. An epistemic theory which says 'we know nothing' is difficult to accept because it conflicts with intuitions to the effect that we know many ordinary propositions (e.g. 'I exist', '2+2=4' etc.). These counter-examples are so compelling that they are nearly enough to discredit the theories in question.[1]

This modest piece of reasoning displays philosophical methodology in action. Indeed, many philosophers believe that intuitions carry great evidentiary weight. Saul Kripke (1980, p. 42) suggests at least this much:

> …Of course some philosophers think that something's having intuitive content is very inconclusive evidence in favor of it…I really don't know in

[1] Of course, the exact role of intuitions in traditional philosophy is a controversial issue. See, for example, Ernest Sosa (2005, 2007) and Timothy Williamson (2007). Although most of this section describes one plausible way of understanding their role, the main points of this paper can be established under much weaker assumptions (see the last paragraph of this section for more details).

a way what more conclusive evidence one can have about anything ultimately…

One doesn't have to accept Kripke's (perhaps strong) remarks concerning evidence and intuition to recognize that appealing to intuitions is standard practice in philosophy.

A question that has not been seriously investigated until recently is whether or not the intuitions that philosophers call upon to develop their theories are shared by the general public.[2] In the last few years, however, some theorists known as 'experimental philosophers' have attempted to answer this question. They have gathered a fascinating body of evidence which suggests that philosophers' intuitions concerning knowledge, intentional action, reference, morality and other central notions are distinct from the intuitions held by the general public, or in some cases certain socio-economic or cultural groups.[3]

Some theorists take these results to show that contemporary philosophical practice must change. One line of criticism is that if philosophy aims to say something of relevance to humanity, and not just to philosophers, it must take into account the intuitions of the general public-- and this requires adopting whatever empirical methods are needed to carry out this project in a scientifically respectable manner.[4] Another more

[2] This question is especially pertinent given that philosophers often carry out their investigations in the 'armchair', or by consulting only their peers (and hardly ever by directly consulting the public).

[3] There is a large body of literature on each of these topics. Here is a very small sampling of it. Knowledge: Alexander and Weinberg (2007); Bengson *et al,.* (2008); Intentional Action: see the citations found in this paper; Reference: Machery *et al.,* (2004); Mallon *et al.*, forthcoming; Morality: Haidt (2001); Cushman, Hauser, and Young (2006). For differences that arise among different socio-economic or cultural groups, see Nichols, Stich and Weinberg (2003); Weinberg, Nichols and Stich (2001).

[4] On this way of thinking, intuitions should still be used in philosophical theorizing but we must ensure that they are genuine folk intuitions (gathered using empirical methods). Most notably, Knobe's, (e.g. 2003a; 2003b) aim is to give an account of intentional action, and he thinks that experimental philosophy can greatly aid this project. Bengson *et al.* (2008) think the evidence they gathered from the folk provides prima facie evidence for a radical intellectualist analysis of know-how. Knobe and Prinz (2008) collected data they think is relevant for understanding the folk concept of consciousness. See Appiah (2008) and Nahmias (2006) for a general discussion along these lines.

radical line of criticism is that intuitions are unstable to the point of being unreliable and should, in many cases, no longer count as serious evidence in the development of philosophical theories (e.g. Weinberg, Stich, and Nichols, 2001). Either way, it is thought that the results of experimental philosophy are of great significance to the future of philosophy.[5] A key assumption driving both of these criticisms is a type of skepticism about philosophers' intuitions: the experimental results show that the evidentiary quality of philosophers' intuitions is lower than previously thought.

We suspect, however, that many philosophers would not grant that this skepticism about intuitions is something that falls out of the results of experimental philosophy. They may hold that in the cases where the philosopher's intuition is incompatible with what the folk say in certain experimental conditions, the folk are probably mistaken. Consider a claim made by George Bealer (1999, p. 202):

> Many philosophers enjoy the pastime of 'intuition bashing,' and in support of it they are fond of invoking the empirical findings of cognitive psychologists. Although these studies evidently bear on 'intuition' in a less discriminating use of the term (e.g., as a term of uncritical belief), they tell us little about intuition in the relevant sense.

Bealer distinguishes between 'intuition' as uncritical belief, and 'intuition' understood as a mental state one can have in 'a higher quality cognitive condition'. The typical results of experimental philosophers would then mainly concern the former, whereas traditional philosophy is concerned with the latter. Timothy Williamson (2007, p. 7) makes a related point:

[5] These two criticisms correspond roughly to the two main camps experimental philosophers fall into when it comes to saying how their work is significant for philosophy. Joshua Alexander and Jonathan Weinberg (2006) call the first position the *proper foundation view* and the second position the *restrictionist view*.

> The method of conducting opinion polls among non-philosophers is not very much more likely to be the best way of answering philosophical questions than the method of conducting opinion polls among non-physicists is to be the best way of answering physical questions…

Although Williamson does not discuss intuitions here, the passage is suggestive of the following: the responses elicited from the folk in the conditions created by experimental philosophers (usually surveys) should carry minimal weight in philosophy. Furthermore, philosophers (in virtue of being experts) enjoy a relative epistemic advantage over non-specialists.

These ideas then point to the notion that the reported conflicts between the philosopher and the public do not warrant skepticism about the intuitions employed by philosophers. This is because philosophers are in a better epistemic position than the folk when the latter are put in the typical experimental conditions. We call this general response to experimental philosophy 'The Immunity Objection'. The name suggests that traditional philosophers are largely immune from the sorts of criticisms raised by experimental philosophers.[6]

We believe that the Immunity Objection, suitably interpreted, has great merit. However, we do not believe that it is enough for philosophers to simply state the objection and go back to business as usual. One who accepts the objection must be

[6] The Immunity Objection can take various forms. Swain *et al.* (2008), for example, consider a version in which the philosopher's epistemic privilege comes from the fact that they are more reflective. In contrast, the experimental philosopher, by using surveys, is only calling for quick intuitions that are less reflective. Matthew Liao (2008) considers a similar objection he calls 'The Argument From Robust Intuitions'. Other authors that have criticized experimental philosophy on similar grounds include Antti Kauppinen (2007) and Ernest Sosa (2007). But these aren't the only ways to understand the Immunity Objection. For example, it might be thought that philosophers display higher levels of intelligence or that they are better at not being distracted by irrelevant details. The term 'Immunity Objection' is intended to cover this family of objections.

committed to a significant claim that is empirically testable: *in the cases of conflict discussed by experimental philosophers, judgments of the folk delivered under sufficiently better epistemic conditions are more likely to match the intuitions of traditional philosophers.* In this paper, we test the efficacy of the Immunity Objection against skeptical conclusions drawn from the Knobe Effect. The experiments we conducted give strong evidence for the soundness of the Immunity Objection applied in this domain.

A caveat is in order before we continue. We follow the literature in calling the following two types of judgments 'intuitions': judgments elicited from the folk in the typical experimental conditions and judgments philosophers make in response to thought experiments in the typical 'armchair' conditions. Thus, we will be using 'judgment' and 'intuition' in these contexts interchangeably. This should not be understood as a commitment on our part to a substantive claim about the nature of intuitions (must they be quick and unreflective or must they be the product of careful deliberation?). However, it is worth pointing out that one of the take-home messages from this paper is that in the cases of interest where the folk and the philosophers disagree, they may be arriving at their judgments through distinct psychological processes.

3. The Knobe Effect

In a well-known experiment, Joshua Knobe (2003b) gave subjects the following vignette which we call '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.

Knobe discovered that 82% of subjects given this vignette in a survey responded that the chairman intentionally harmed the environment. But now consider a parallel scenario, which we call 'Help', where the side effect on the environment is positive:

> 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.

When Knobe gave Help to a different set of subjects, he found that only 23% of them said that the chairman intentionally helped the environment. These results are robust. Our own experiments confirm this finding (see Appendix A). It has also been replicated for speakers of different languages (Knobe and Burra, 2006) and for young children (Leslie *et al.*, 2006). We understand 'The Knobe Effect' to be the fact that agents placed in the experimental conditions typical of these studies give the asymmetric responses we have seen to the Harm and Help vignettes.[7] The Knobe Effect suggests that competent judgments about whether someone intentionally brought about a side effect depend on the perceived moral quality of the action (including its side effects), and this is the

[7] This accords with Nichols and Ulatowski's (2007) use of 'Knobe Effect'.

conclusion that Knobe himself wants to draw.[8]

These results are surprising.[9] Many philosophers feel strongly that moral considerations do not bear on competent judgments of intentional action in the way that the Knobe Effect suggests.[10] Here then is a case where experimental philosophers have found a conflict between philosophers' judgments and what the folk say under certain conditions. Accordingly, experimental philosophers will want to draw the skeptical conclusions. And this is where the Immunity Objection comes in: due to the philosopher's epistemic privilege, her intuitions hold more evidentiary weight in getting at the concept of intentional action than those judgments elicited from the folk (in the typical experiments).

To test whether the Immunity Objection is effective in this case we conducted a number of experiments aimed to see whether the following hypothesis is true:

General Hypothesis: Subjects in better epistemic conditions are less likely to display the Knobe Effect.

If this hypothesis is correct, it is strong evidence that the Immunity Objection is a sound response to skeptical conclusions motivated by the Knobe Effect. To evaluate the General Hypothesis, we formed three more specific and testable hypotheses:

Intelligence Hypothesis: Subjects with higher general intelligence are less

[8] We follow the literature in speaking of the environmental impact described in the vignettes as being 'side effects'. But we are not convinced this is the best way of describing the situation if a side effect is supposed to not also be a means for bringing about the desired outcome. This is because it is possible to interpret the vignettes so that the environmental effect was a means for bringing about profits. For a discussion of this idea see Cushman *et al.* (2008).

[9] Although see Harman (1976) for a dissenting opinion.

[10] At least this much has been suggested by Knobe (2003b) and it is substantiated by the fact that philosophers find the Knobe Effect surprising. Note that we are only saying that the philosopher's judgments about the Harm and Help scenario are symmetric. We do not make a commitment about the particular pattern of responses that generates the symmetry. Moreover, we don't even assume that philosophers make uniform judgments about these cases, only that they are symmetric.

likely to display the Knobe Effect.

Awareness Hypothesis: Subjects aware that their initial response to a question may be mistaken are less likely to display the Knobe Effect.

Further Information Hypothesis: Subjects allowed to see both Help and Harm scenarios before giving answers are less likely to display the Knobe Effect.

It is natural to think that the truth of these claims supports the general hypothesis. With respect to the Intelligence Hypothesis, it is plausible that having higher general intelligence gives one, ceteris paribus, an epistemic advantage over similarly situated individuals with lower general intelligence. With respect to the Awareness Hypothesis, it is often the case that one's knee-jerk reaction to a question is wrong, and being made aware of this puts one in a better epistemic situation. Simply put, one will reason more carefully. With respect to the Further Information Hypothesis, a person who is made aware of further relevant information concerning an issue is generally in a better epistemic position than someone equally positioned who does not possess that extra piece of information.

It turns out that our experiments support all the three hypotheses, and hence the General Hypothesis. The contrast between what the folk say in the typical experiments and what philosophers say is due to the relatively low quality of the experimental subjects' epistemic position. As a consequence, the Immunity Objection applied to the Knobe Effect is successful. In other words, the Knobe Effect does not give us good reason to be skeptical about intuitions and in particular, philosophers' intuitions.

4. The Experiments

We conducted three experiments to test the General Hypothesis. But before we get to the experiments, we discuss what should count as a situation in which the Knobe Effect is reduced for people under better epistemic conditions. As mentioned earlier, the Knobe Effect is the great asymmetry found between the percentage of people who say 'Yes, the chairman intentionally harmed the environment' in response to the Harm vignette and the percentage of people who say 'Yes, the chairman intentionally helped the environment' in response to the Help vignette when these people are tested in the conditions created by the experimental philosopher. A high percentage of subjects give the first answer while a low percentage give the second answer. Suppose then that you divide a sample of the population into two groups such that one of them enjoys an epistemic advantage over the other. If you find that the subjects in the high epistemic quality group exhibit less asymmetry in their answers than those in the low epistemic quality group, then here is a case where the subjects in the better epistemic condition are less likely to exhibit the Knobe Effect. To be clear then, when we say in the context of this paper that subjects in a higher quality condition are less likely to exhibit the Knobe Effect, we should be understood as making a comparative claim: we are saying that they exhibit less asymmetry compared to subjects in relatively lower quality conditions.

To carry out these experiments, we surveyed 1,094 undergraduate students enrolled at Arizona State University between June and October of 2008. The surveys were conducted in classrooms. Every class that we surveyed, with the exception of one, was not an upper level philosophy class. In the case of the exception, the class is well-known to be populated mostly by people who have not taken other philosophy courses. The

surveys were administered by the authors of this paper.

4.1 The Intelligence Hypothesis: Subjects With Higher General Intelligence are Less Likely to Display The Knobe Effect.

In order to test the Intelligence Hypothesis, we needed a way of measuring general intelligence. We used Shane Frederick's (2005) Cognitive Reflection Test. We predicted that groups that do better on this test were less likely to display the Knobe Effect. Two questions arise: what is the Cognitive Reflection Test (CRT)? And why did we use it? Regarding the first, the CRT consists of three, quick questions:

> (1) A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost? ____ cents
>
> (2) If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? _____ minutes
>
> (3) In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? _____ days

Consider (1). At first glance, the answer seems obvious. In fact, most participants have the immediate but mistaken impression that the solution is 10 cents, even when they go on and give the correct response. When participants reflect carefully (by double-checking their answer for example), they realize that their initial response is mistaken and settle on the correct answer (5 cents). Thus, 'catching [the] error is tantamount to solving the problem.' (Frederick, 2005, p. 27) All three questions were specifically designed in this way: the seemingly obvious answer is wrong, but can be easily corrected upon minimal

reflection.[11]

So, why did we use this test? Frederick's studies of over 3,000 participants revealed highly significant correlations between the CRT and many other measures of general intelligence (SAT verbal, SAT math, ACT, the Need for Cognition test, WPT, etc). Studies were conducted at multiple universities of differing SAT averages—the CRT averages tend to reflect these differences. Furthermore, the test is ideal for quick surveys since it can be taken in about two minutes. Ultimately, the CRT is suitable for our purposes since we only wish to separate participants into very broad groups of general intelligence (rather than more narrow groups, which would require something more like an IQ test).[12]

4.1.1 Intelligence Experiment

Method

The participants were 870 undergraduate students (873 originally, but 3 failed to complete both parts of the survey and so their partial answers were not included). They were split into four groups, each of which received a two-page survey. Participants were verbally instructed by those administering the surveys not to go back to the first sheet after they looked at the second sheet. At the bottom of the first page were instructions

[11] Explaining why he chose these questions, Frederick (2005, pp. 27-28) writes, 'The proposition that the three CRT problems generate an incorrect "intuitive" answer is supported by several facts. First, among all the possible wrong answers people could give, the posited intuitive answers (10, 100 and 24) dominate. Second, even among those responding correctly, the wrong answer was often considered first, as is apparent from introspection, verbal reports and scribbles in the margin (for example, 10 cents was often crossed out next to 5 cents, but never the other way around). Third, when asked to judge problem difficulty (by estimating the proportion of other respondents who would correctly solve them), respondents who missed the problems thought they were easier than the respondents who solved them…Fourth, respondents do much better on analogous problems that invite more computation. For example, respondents miss the "bat and ball" problem far more often than they miss the "banana and bagel" problem: "A banana and a bagel cost 37 cents. The banana costs 13 cents more than the bagel. How much does the bagel cost?'.

[12] Joshua Knobe suggested to us that the CRT may be understood as a measure of 'reflectiveness' and that this, in turn, may be seen as a feature of intelligence. We do not take a stance on this issue. It is enough for our purposes that the CRT correlates (in a highly significant way) with other measures of intelligence.

that they had to finish before they turned the page. And at the bottom of the second page were instructions to turn the test over once they completed it. The administrators monitored the students to ensure the instructions were followed.

Group 1 (N = 199) received the Harm vignette on the first page and then the CRT on the second page. Group 2 (N = 207) received the Help vignette and then the CRT. Group 3 (N = 234) received the CRT and then the Harm vignette, while Group 4 (N = 230) received the CRT and then the Help vignette. We will refer to the first two groups as the 'CRT-After' ordering and the last two groups as the 'CRT-Before' ordering. The Harm/Help vignettes were immediately followed by this question:

Which of the following do you agree with most?

Those receiving the Harm vignette were given the following answer choices:

(A) The chairman intentionally harmed the environment.

(B) The chairman did not intentionally harm the environment.

Those receiving the Help vignette received similar answer options except with 'help' replacing 'harm' in each.

General Intelligence Measure

There were no statistically significant differences found in the CRT scores between the CRT-Before group ($M = 0.77$, $SD = 1.00$) and the CRT-After group ($M = 0.85$, $SD = 1.10$), $t(868) = -1.10$, $p = .272$. In each group, 45% of the participants answered at least one question correctly, $\chi^2(1, N = 870) < .001$, $p > .999$, Cramer's $V = .002$. Thus, the evidence suggests that in the CRT-After ordering, the subjects' thinking about the Harm/Help vignettes did not affect the answers they gave on the CRT immediately after. It is fair to conclude then that the CRT-After ordering surveys accurately split

participants into groups of general intelligence.

Results

We used the CRT-After ordering to test the Intelligence Hypothesis.[13] As predicted, general intelligence (as measured by the CRT) reduced the Knobe Effect. In particular, the Yes Harm judgment decreased in groups of greater general intelligence.[14] In the Harm scenario, of the 107 participants who scored 0 on the CRT, 79% of them answered Yes Harm. But of the 92 participants who scored 1 or greater, only 65% gave that answer. A chi-square test showed this decrease to be significant, $\chi^2(1, N = 199) = 4.367$, $p = .037$. We point out, however, that no such increase or decrease was found for judgments in the Help scenario, $\chi^2(1, N = 207) = .308$, $p = .579$. See table 1 for a summary. Note that that we will often refer to a score of 1 or greater on the CRT as '1+'.

Table 1 (CRT-After)

	Yes Harm	Yes Help
CRT 0	79%	14%
CRT 1+	65%	18%

Discussion

If we compare the group of those who scored 0 on the CRT and those who scored 1 or more, we found that the second group displayed less asymmetry. This supports the Intelligence Hypothesis: those who display higher general intelligence are less likely to exhibit the Knobe Effect. More details on this experiment can be found in Appendix B.

[13] We do not use the CRT-Before ordering to test the Intelligence hypothesis because getting questions right on the CRT affects how one responds to the Help/Harm vignettes. We will see below that this is the crucial result of the Awareness Experiment.

[14] We code the 'The chairman intentionally harmed the environment' answer as 'Yes Harm'. We also use 'No Harm', 'Yes Help' and 'No Help' for the other answers.

4.2 The Awareness Hypothesis: Subjects Aware That Their Initial Response to a Question May Be Mistaken Are Less Likely to Display The Knobe Effect.

Recall that each CRT question is designed in such a way that one's first pass judgment is mistaken. Furthermore, most participants who get the correct answer have the initial mistaken intuitions of 10 cents, 100 minutes, or 24 days (see footnote 11). Suppose then that you have just taken the CRT and gotten some questions right. It is plausible that at this point, you are now made aware that your first pass judgment to problems may very well be mistaken. Accordingly, if you are then immediately given another question, this awareness may then play a role in how you answer that question. It is natural to think that this gives one an epistemic advantage. Being made so aware makes one, for the moment, into a more careful thinker. Hence, we predicted that those who get questions right on the CRT *before* they are given the Harm/Help vignette are less likely to display asymmetric answers than those who get questions right but take the CRT *after* they are given the Harm/Help vignette.[15]

4.2.1 Awareness Experiment

Method

Here, we look at the subjects who took the CRT before they were given the Harm/Help vignettes (Groups 3 and 4) and got at least one question right. The control are those who got at least one question right in the CRT-After condition (Groups 1 and 2). Since the control is composed of people with the same level of intelligence as measured by the CRT, we rule out level of intelligence as a confounding variable. We refer the reader to the 'Method' section of Experiment 1 for more details on the groups.

Results

[15] As far as we know, we are the first to use the CRT to *prompt* more careful reasoning.

We begin by looking at those who score a 1+ on the CRT in the CRT-After condition. 65% of them answered Yes Harm to the Harm vignette. Now, for those who score a 1+ in the CRT-Before condition this percentage is 63%. There is no statistically significant difference between these figures, $\chi^2(1, N=201)=.02$, $p=.893$.

The interesting result comes when we look at the Yes Help numbers. In the CRT-After condition, 18% of those who scored a 1+ on the CRT responded Yes Help to the Help vignette. But this percentage is much greater for those who scored a 1+ in the CRT-Before condition. The percentage increases to 34% (See Table 2). This difference is statistically significant, $\chi^2(1, N=192)=5.62$, $p=.018$.[16]

Table 2 (CRT-Before vs. CRT-After for CRT 1+)

	Yes Harm	Yes Help
CRT-After (1+ Score)	65%	18%
CRT-Before (1+ Score)	63%	34%

Discussion

Looking at table 2 we see that there is a reduction of asymmetry for the subjects who took the CRT before answering the vignettes and who got at least one question right. But before we conclude that this supports the Awareness Hypothesis, it is worth checking to see that this data is not better explained by another hypothesis.

It might be thought that the reduced asymmetry is due to the idea that performing well on the test does not increase 'awareness' but simply gets people to want to give the

[16] The comparison can also be made between the groups restricted to those who score a 2 or better on the CRT. Here are the results. In the CRT-After condition, 69% said Yes Harm while 15% said Yes Help. In the CRT-Before condition, however, 63% said Yes Harm while 45% said Yes Help. The difference between the Yes Harm responses is not statistically significant, $\chi^2(1, N=115)=.1$, $p=.748$. But the differences between the Yes Help answers is statistically significant, $\chi^2(1, N=101)=9.5$, $p=.002$. This adds further support for the Awareness Hypothesis.

opposite response from their initial 'gut reaction' answer. This idea fails since if this were the case, we would see that the increase in Yes Help answers is matched by an equal decrease in Yes Harm answers. But we saw that the Yes Harm answer does not significantly change between the groups being compared.

Now we saw that getting questions right on the CRT makes one aware that their initial 'gut feeling' responses to questions might be wrong. We conclude that this 'awareness' is what explains the lowered asymmetry. Hence, the data presented here supports the Awareness Hypothesis. [17] [18]

[17] It is worth comparing the pattern of Help/Harm answers between groups separated by CRT scores in the CRT-Before ordering. 14% of those who scored a zero on the CRT said Yes Help while 34% of those who scored one or higher gave that answer. Hence, participants who answered at least one question correctly on the CRT-Before ordering were more likely to say Yes Help. This outcome is statistically significant $\chi^2(1, N = 230) = 16.494$, $p < .001$. Recall from the Intelligence Hypothesis discussion, however, that there was no statistically significant difference between the figures for the corresponding groups in the CRT-After ordering. This indicates that the difference in the CRT-Before condition is due to more than just an increase in intelligence--it is due to 'awareness'. Furthermore, each point increase in the CRT (in the CRT-Before ordering for those scoring 1 or higher) corresponds to a remarkable increase in the percentage of Yes Help answers: 24% of those who scored a 1, 38% of those who scored a 2, and 56% of those who scored a 3 said Yes Help! These results provide strong evidence that even among participants scoring at least 1 on the CRT there is an association between CRT score and Help answer, $\chi^2(2, N = 101) = 6.323)$, $p = .042$. This was also statistically significant. Note that since there is no such step-wise increase in Yes Help answers in the CRT-After ordering (see Appendix B) we believe that the increases are due to "awareness" and that this is evidence that "awareness" is gradable.

[18] A referee points out that some of the data collected from this experiment could also count against the Immunity Objection. In the CRT-Before condition, 89% of those who score a zero in the CRT answer 'No Help' while only 44% of those who got a perfect score in the CRT gave that answer. Assuming that philosophers tend to give the 'No Help' answer, it looks like the less smart and less aware folk give answers that are more closely aligned with what philosophers say than their smarter and more aware peers. The referee suggests that this result can count against the Immunity Objection. Here are some responses. First, it strikes us that discovering in our sea of data a pattern of answers that goes against the Immunity Objection is not enough to discredit it (we are not suggesting that the referee has this in mind, however). For we give in this paper plenty of other evidence in its favor. Moreover, this is evidence predicted by antecedently motivated hypotheses. Second, we must distinguish between a general form of the Immunity Objection and a more specific one applied to the Knobe Effect. In this paper, we are interested in testing the specific one, which predicts that subjects in better epistemic conditions are less likely to exhibit the Knobe Effect asymmetry. It is clear that the data the referee points to is not evidence against this hypothesis. But is it evidence against the more general one? The more general hypothesis is committed to the claim that the cases of conflict (found by experimental philosophers) between ordinary folk and philosophers are due to the epistemic advantage of the philosopher. The data, however, does not seem to constitute evidence against this, since presumably there is no conflict between the ordinary folk and the philosopher in their answers to the Help scenarios (they both say 'No Help'). Moreover, the point of the Immunity Objection is to address data that have been used as evidence by theorists to draw certain well motivated conclusions of

To end this section, we would like to share a surprising discovery we stumbled upon when examining our numbers. If we look at the subjects that score perfectly on the CRT in the CRT-Before condition, we find an amazing pattern of answers. 65% of them say Yes Harm in response to the Harm scenario, while 56% say Yes Help in response to the Help scenario. This constitutes a dramatic reduction from the asymmetry found in the original Knobe studies. In fact, there is no statistically significant difference in the percentage of 'Yes' answers between subjects given the Help and Harm vignettes, $\chi^2(1, N = 41) = .094$, $p = .759$, Cramer's $V = .098$. For this group, the Knobe Effect has all but disappeared! What this result indicates is that those that are maximally 'aware' and maximally intelligent (as measured by the CRT) do not exhibit the Knobe Effect. We believe that this result taken on its own is very strong evidence for the soundness of the Immunity Objection applied to the Knobe Effect.[19]

4.3 The Further Information Hypothesis: Subjects that are Allowed to See Both Help and Harm Scenarios before Giving Answers are Less Likely to Display The Knobe Effect.

For this experiment, before we asked subjects to give their judgments on whether the chairman intentionally helped/harmed the environment, we gave them both the Harm and Help vignettes. By contrast, in Knobe's original studies, subjects were only given one story before being asked the corresponding question. The main difference then in our experiment is that we gave subjects more pertinent information before they had to make

---

philosophical interests. The data discussed in this footnote is not of this type. Nonetheless, it might still be relevant to those wishing to give a positive accounts of the concept 'intentionally'.

[19] At the time this manuscript was being prepared, we became aware of the interesting results in Cokely and Feltz (2009). They found no correlation between CRT and the 'Knobe Effect'. There were some important differences between their probes and ours that we cannot fully discuss here. We point out, however, that their sample size was smaller (97 subjects) compared to ours (870 subjects for the CRT in the before and after conditions). Hence, we were able to detect subtle differences their probes could not.

up their minds. In general, giving subjects further relevant information will allow them to make a more informed judgment. In short, it will put them in a better epistemic position. We predicted that subjects who participate in this experiment were less likely to display the Knobe Effect.[20]

At this point we need to be more specific about what it means to have a reduction of the Knobe Effect in this condition. This requires a bit of set up so we ask the reader to bear with us. As mentioned before, in the original Knobe study 82% of subjects answer Yes Harm when presented the Harm vignette by itself. But 77% of subjects (from a different group) answer No Help when presented with the Help vignette by itself. But suppose we want to know the percentage of people who will have *both* of the following dispositions: The disposition to say Yes Harm in response to the Harm vignette given by itself, and the disposition to say No Help in response to the Help vignette given by itself. Simply put, we want to know what percentage of the population has the dispositions to make the Knobe asymmetric judgments. How can we determine the answer to this question?

It is difficult to carry out a direct experiment to determine this percentage. If we gave people both scenarios, one after the other, we run the risk that the reaction to the second scenario might be affected by their first answer. However, from the original Knobe numbers, we can determine the *minimum* percentage of people who have the dispositions to make the asymmetric judgments. Here's an informal explanation of how we computed this number. In what follows, we use the nearly identical figures we got

[20] Nichols and Ulatowski (2007) report on an experiment where subjects are also presented with both vignettes. One difference between their experiment and ours is that (in theirs) the subjects are asked for their reaction right after they read each vignette where there is no possibility that they can change their answer to the first vignette after reading the second.

when we replicated Knobe's numbers (because our sample size is larger, and is drawn from the same population as our experiments). These are 78% for Yes Harm (N=81) and 18% for Yes Help (N=82). See Appendix A.

Suppose we are given a group of 100 randomly selected people. According to our results and idealizing, 78 of them are disposed to say Yes Harm when presented with the Harm vignette by itself. We know that *at most* 18 out of that 78 are also disposed to say Yes Help when presented with the Help vignette by itself (this is because 18 out of the total 100 are so disposed). This means that *at least* 60 (78 minus 18) out of that 78 are instead disposed to say No Help when presented with the Help vignette by itself. So, at least 60 have both dispositions. Converting to percentages, at least 60% of the population has both dispositions. That is, at least 60% of the population has the disposition to give the Knobe asymmetric answer. We leave the formal derivation of the result for Appendix C.[21]

We can now give a gloss to our prediction that the Knobe Effect will be reduced for a group in which both Help and Harm vignettes are given at the same time. This claim is made true if the percentage of people in that group who answer 'Yes Harm *and* No Help' is less than 60%. This counts as a reduction of the Knobe Effect.

4.3.1 Further Information Experiment

Method

The participants were 221 undergraduate students. Each received one sheet which contained both the Harm and the Help scenarios (ordering was counterbalanced) and a single question with four answer options:

[21] If we use Knobe's original data the minimum percentage population would be 59%.

(A) The first chairman intentionally harmed the environment and the second chairman intentionally helped the environment.

(B) The first chairman did not intentionally harm the environment and the second chairman did not intentionally help the environment.

(C) The first chairman intentionally harmed the environment, but the second chairman did not intentionally help the environment.

(D) The second chairman intentionally helped the environment, but the first chairman did not intentionally harm the environment.

The answer choices were suitably modified for the second ordering. The instructions also made it explicit that the companies in the two scenarios were unrelated.

Results

As predicted, adding more information reduces the Knobe Effect. Only 48% of participants chose the Knobe asymmetric response in this condition, compared to the (at least) 60% of people who are disposed to give these answers if they were given the scenarios on their own. (There was no ordering effect for the Knobe response--on both orderings it was 48%). Using the 60% number as the null hypothesis, the result is highly statistically significant, χ^2 (1, N = 221) = 17.422, $p < .001$.

Discussion

We just saw a reduction of the Knobe Effect among subjects who were allowed to see both scenarios before giving the judgments about the chairman. The number goes down from 60% to 48%. We emphasize that 60% is the most conservative number that is mathematically possible. In reality, the reduction may be much greater. We conclude then that there is good reason to think that the Further Information hypothesis is true. Being

presented with both scenarios does lead to a reduction of the Knobe Effect. [22]

4.4 General Discussion

We saw that all the three hypotheses were empirically supported. The first experiment showed that subjects with higher general intelligence are less likely to exhibit the Knobe Effect. The second experiment showed that subjects that display 'awareness' are also less likely to display the effect. Finally, the Further Information experiment showed that when subjects are given both Help and Harm scenarios, they are less likely to give the asymmetric 'Knobe' response. Each of these results supports the General Hypothesis: that agents in better epistemic positions are less likely to exhibit the Knobe Effect. As previously discussed, a reduction of the effect reflects a pattern of answers that align more closely with philosophical judgment. We conclude that our data gives strong support for the Immunity Objection as applied to the sort of skepticism about philosophers' intuitions that may arise when presented with the Knobe Effect. Before we continue, however, we would like to address some worries about our results.

Objection: The experiments all show that subjects in better epistemic conditions are less likely to display the Knobe Effect asymmetry. But it is not clear that this is enough to vindicate the Immunity Objection. If large majorities of subjects in the epistemically advantaged groups still display the asymmetry (although to a lesser degree than their

[22] There is another way one might interpret what it means to have a reduction of the Knobe Effect in this condition. We could compare the percentage of Yes Harm answers when given both vignettes (which would be reflected by giving the (a) or the (c) answer in the multiple choice options above) to the percentage of Yes Harm answers when given only one vignette (and make a similar comparison with the 'Help' answers). On this way, as well as the one discussed in the main body of the paper, we found a significant reduction of the Knobe Effect. 57% of the 221 subjects given both vignettes said Yes Harm, this number is significantly different from the 78% Yes Harm out of the 81 subjects given only the Harm vignette, $\chi^2(1, N = 302) = 10.470$, $p = .001$. There was no such association for the Yes Help answers, $\chi^2(1, N = 303) = 1.125$, $p = .289$. This way of understanding the reduction is, however, complicated by the fact that there is likely a slight ordering effect (depending on which vignette is given first) for the 'non-Knobe' answers (e.g. 'Yes Harm and Yes Help' etc.--see Appendix D). This ordering effect is interesting but we do not pursue it here.

peers in the disadvantaged groups), then it looks like the difference in intuitions between the philosophers and the folk is not solely due to the former's presumed epistemic advantage. The difference might be due to something else (for example, bias due to philosophical training). And if so, the Immunity Objection is not vindicated.

Response: Before we get to the heart of the objection, we want to make some preliminary points. Since we discovered statistically significant reductions in asymmetry between the relevant groups of subjects, we have found evidence in favor of the Immunity Objection. We agree that (in certain cases) discoveries of larger reductions would constitute stronger evidence. But (of course) the fact that there might be better evidence available should not take away from the significance of our results. Furthermore, it is unreasonable to expect that subjects in the epistemically superior groups will give the same answers as philosophers. We should not suppose, for example, that those who score 1 or greater on the CRT in our experiments are on equal epistemic footing as the philosophers. So defenders of the Immunity Objection should not expect the asymmetries to completely vanish for our experimental groups (groups in the better epistemic conditions).

Now, let us address the heart of the objection. Given the reported reductions in asymmetry, we can still wonder about how much support we have actually found for the Immunity Objection (as applied to the Knobe Effect). We believe that the results reported in this paper provide strong support for it. Using the simple mathematical methods discussed in the third experiment and in Appendix C, we can see how in each of our three probes, the experimental groups exhibit reduced asymmetries to the point where we should doubt that anything like a large majority of their members exhibit the Knobe

Effect. Let us turn to this. In the Intelligence Experiment, out of those who scored a 1+ on the CRT in the CRT-After condition, 65% of them say Yes Harm and 18% say Yes Help. From this, the strongest thing we can conclude about the pervasiveness of the Knobe Effect is that at least 46% of them (65 minus 18) and at most 65% exhibit the Knobe Effect (have the disposition to say Yes Harm and the disposition to say No Help). Similarly, in the Awareness Experiment, out of those who score a 1+ in the CRT-Before condition, we can only conclude that at least 29% and at most 63% of them exhibit the Knobe Effect. Finally, in the Further Information Experiment, exactly 48% of those receiving the 'further' information exhibit the Knobe Effect. In sum, we see that in all our conditions, the epistemically privileged groups exhibit reduced asymmetries to the point where we cannot even establish that the majority of their members exhibit the Knobe Effect. In the case of the third experiment, and in the highest scoring group from the second experiment, we can say for sure that the majority do not exhibit the Knobe Effect. In light of this information, we doubt that anything like a large majority of subjects in our epistemically superior groups display the Knobe Effect. It is fair to conclude then that our results give strong evidence for the Immunity Objection as applied to the Knobe Effect.

5. Implications for Understanding the Concept of Intentional Action.

The asymmetry in the responses to the Help and Harm scenarios elicited in the original Knobe experiment suggested to some that the folk concept of intentional action is intimately connected to assessments (conscious or not) of the moral quality of the action including its side effects. After all, the only salient difference between the Harm and Help scenarios is whether the environmental effect is good or bad. It might be thought that a

competent deployment of the concept of intentional action would have it that the chairman in the Harm scenario intentionally harmed the environment, while the chairman in the Help scenario did not intentionally help the environment. That is, the asymmetry discovered by Knobe does not reflect random performance factors and instead accurately reflects our competence with the concept 'Intentionally'. [23] If we add to this position the further idea that there is only one concept 'Intentionally', we have what we call the 'One Concept' theory. Nichols and Ulatowski (2007) and Machery (2003) attribute to Knobe a position approximating this view, and we have confirmed with Knobe (personal communication) that this is what he believes.[24]

The One Concept view requires that the responses to the Harm and Help scenarios which reflect competency with 'Intentionally' are Yes Harm and No Help respectively.[25] Our study gives evidence that the One Concept view is mistaken. If the asymmetric responses are those that reflect competency, then people in better epistemic conditions should not exhibit less asymmetry. The results presented in this paper, however, show the opposite result. People in better epistemic conditions do exhibit less asymmetry.

6. System 1/System 2 Considerations.

Dual-process theory, a widely accepted view in cognitive psychology, maintains a

[23] See Nichols and Ulatowski, 2007 for a discussion of the evidence that the Knobe Effect is not due to performance errors.

[24] We take the one concept view to be incompatible with the idea that the 'intentionally' concept is context sensitive (i.e. the extension of the concept varies with the context of thought). Similarly, we take it to be incompatible with the idea that the concept is relativistic in the sense that its extension varies with some 'interesting' parameter (neither the time nor the world parameters) in the circumstance of evaluation. Nichols and Ulatowski (2007) and Cushman and Mele (2008) explicitly reject the one concept view.

[25] This assumes that the vignettes are not so severely underdetermined so that any of the available responses are potentially correct (because the stories could be completed in various ways). We suspect that Knobe would not accept that there is underdetermination since it would seriously undermine the conclusions he draws from his experiments: the asymmetry in responses would say more about how people are likely to fill in the story, than about their concept of intentional action.

distinction between two types of cognitive processing known as System 1 and System 2.[26] [27] Daniel Kahneman (2003, p. 698) highlights some of the differences:

> The operations of System 1 are typically fast, automatic, effortless, associative, implicit (not available to introspection), and often emotionally charged; they are also governed by habit and are therefore difficult to control or modify. The operations of System 2 are slower, serial, effortful, more likely to be consciously monitored and deliberately controlled; they are also relatively flexible and potentially rule governed.

Distinguishing these two types of processes has been very useful in understanding various cognitive processes including those connected to judgment, decision-making, and reasoning.[28] We believe that the experiments presented in this paper are best explained by appealing to dual-process theory. Due to space limitations, the following remarks are speculative. We take ourselves to be merely suggesting that research along these lines may prove fruitful.

The first experiment showed that responses on the Harm vignette are correlated with general intelligence. How might this finding connect with dual-process theory? Evans (2008, p.262) states that, 'One of the stronger bases for dual-systems theory is the

---

[26] See Evans, (2003, 2008) for excellent overviews of dual-process theories. Also, it's worth mentioning that these theories, although popular, are not uncontroversial. See Gigerenzer and Regier (1996), Osman (2004), and Keren and Schul (forthcoming) for criticisms.

[27] It is important to note that the distinction we are appealing to is simply between two kinds of processes, which is a much weaker claim than there being two systems in a more robust sense of the term. Kahneman and Frederick write, '[The terms System 1 and System 2] may suggest the image of autonomous homunculi, but such a meaning is not intended. We use *systems* as a label for collections of processes that are distinguished by their speed, controllability, and the contents on which they operate' (2002, p.51).

[28] For a sampling of the literature, see Epstein (1994); Sloman (1999); Stanovich and West (1998); Liebermann, (2000, 2007); Nisbett *et al.* (2001); Kahneman and Frederick, (2002, 2005); Evans (2003, 2006, 2008). Also note that these theorists do not all use the terms System 1/System 2. For instance, Epstein (1994) uses the names *experiential* and *rational* systems, and Sloman (1996) uses *associative* and *rule-based* systems.

evidence that “controlled” cognitive processing correlates with individual differences in general intelligence and working memory capacity, whereas “automatic” processing does not'. Evans here ties higher general intelligence with use of ‘controlled’ systems, which in turn are normally associated with System 2 processes. If this is right, then there is good reason to think that participants of greater general intelligence are less likely to say Yes Harm because they are relying more on System 2 processes. Hence, the reduction of the asymmetry from the first experiment may be due, in part, to System 2 considerations.

In the second experiment, we found that increased 'awareness' resulted in a greater number of participants saying Yes Help. One plausible line of thinking is that getting at least one correct answer on the CRT is what psychologists call *disfluent.* That is, doing well on the CRT causes a 'metacognitive' feeling that one's cognitive processes are not operating smoothly. This is because getting a question right on the CRT amounts to realizing that one's initial answer was mistaken. This disfluency then results in the activation of System 2 processing. Recent work suggests that there is precedence for this idea. Alter et al. (2007), for example, give evidence that disfluency tends to cause greater System 2 activation. Thus, dual-process theory looks like a promising way to help explain the results of the second experiment.

Concerning the third experiment, we believe that presenting agents with both vignettes (and letting them see the range of multiple choice answers), pushes them to think more carefully before giving the final judgment. If we compare this with the original Knobe experiments (where subjects were given only one vignette followed by just two answer options), it is plausible that subjects there were less careful in their reasoning. Since, as we mentioned earlier, thinking more carefully is characteristic of

System 2 processes, the reduction of the asymmetry may be due to increased activation of that system.[29]

We conclude then that one's reaction to the Help or Harm vignettes should be understood in terms of dual-process theory. If this line of reasoning is correct, then insofar as research into the Knobe Effect is not sensitive to System 1/System 2 considerations, one should be hesitant before drawing conclusions about the concept of intentional action or general skepticism about philosophical practice.[30]

7. Conclusion

We began with a discussion of the experimental philosopher's challenge to traditional philosophy. In particular, we considered the idea that the experimental results give us reason to be skeptical about the intuitions that philosophers use in theorizing, and as a consequence traditional philosophy should change its methods. We then considered the Immunity Objection: this skepticism is not warranted because the judgments of philosophers are made in better epistemic conditions than those elicited from the folk in the typical experiments. Focusing on the Knobe Effect, we presented empirical evidence

[29] One possible line of research worth noting is the similarity between our third experiment and what's known as 'framing effects' in the psychological literature. See Tversky and Kahneman (1974) for an early discussion of the most famous framing effect, the Asian Disease example.

[30] We warn against a simplistic understanding of our experiments and their connection to System 2. We do not want to say that if a group that is in a higher quality epistemic condition (and hence displays an increase in System 2 activation) is more prone to give a certain answer to a vignette, then that answer is the correct or the competent response to the vignette. This conclusion assumes that (a) there is but one concept 'Intentionally' and (b) that the vignettes are not radically undetermined so that they admit of many plausible interpretations (which may yield different competent/correct answers). In fact, at least one author of this paper believes that both of these assumptions are very likely false and that System 2 activation has the effect of making people 'aware' (in some sense) that one or both of these assumptions are in fact false. We can imagine keen subjects having the following reaction to the vignettes 'well, the answer depends on what you mean by "intentionally" and it also depends on whether the "side effect" is a *means* for bringing about profit'. For these and other reasons, we do not give a positive account of the Knobe Effect in this paper. The results of this paper have led us to believe that a full account of the Effect is much more complicated than previously thought.

that the Immunity Objection is effective. We also concluded that a certain explanation of the Knobe Effect attributed to Joshua Knobe is likely mistaken, and that reactions to the Harm and Help vignette may be subject to System 1/System 2 considerations. On a broader note, we hope to have shown how, through the use of the sorts of empirical methods displayed in this paper, we may better understand the connection between experimental and traditional philosophy.

Appendix A. Replication of the Knobe Effect

Help Vignette (N=82)
Yes: 15
Yes: 18.29%

Harm Vignette (N=81)
Yes: 63
Yes: 77.77%

Appendix B. CRT-Before and CRT-After Results

CRT Followed by Harm Vignette

Total (N=234)	YES (N=163)	NO (N=71)
CRT 0 (N=125)	94 (75%)	31
CRT 1 (N=48)	30 (63%)	18
CRT 2 (N=38)	24 (63%)	14
CRT 3 (N=23)	15 (65%)	8
CRT 1+(N=109)	69 (63%)	40
CRT 2+(N=61)	39 (64%)	22

CRT Followed by Help Vignette

Total (N=230)	YES (N=48)	NO (N=182)
CRT 0 (N=129)	14 (11%)	115

CRT 1 (N=54)	13 (24%)	41
CRT 2 (N=29)	11 (38%)	18
CRT 3 (N=18)	10 (56%)	8
CRT 1+ (N=101)	34 (34%)	67
CRT 2+ (N=47)	21 (45%)	26

Harm Vignette Followed by CRT

Total (N=199)	YES (N=145)	NO (N=54)
CRT 0 (N=107)	85 (79%)	22
CRT 1 (N=38)	23 (61%)	15
CRT 2 (N=31)	22 (71%)	9
CRT 3 (N=23)	15 (65%)	8
CRT 1+(N=92)	60 (65%)	32
CRT 2+(N=54)	37 (69%)	17

Help Vignette Followed by CRT

Total (N=207)	YES (N=32)	NO (N=175)
CRT 0 (N=116)	16 (14%)	100
CRT 1 (N=37)	8 (22%)	29
CRT 2 (N=22)	2 (9%)	20
CRT 3 (N=23)	6 (19%)	26
CRT 1+ (N=91)	16 (18%)	75
CRT 2+ (N=54)	8 (15%)	46

Appendix C. 60% "Minimum" Computation

Let 'Joe Sixpack' be the name of a randomly selected person. Let P be the proposition that Joe Sixpack has the disposition to answer YES HARM when being presented with the Harm vignette on its own. Let Q be the proposition that Joe Sixpack has the disposition to answer YES HELP when presented with the Help vignette on its own. We assume that ~P is the proposition that Joe Sixpack has the disposition to answers NO HARM in the relevant conditions, and that ~Q is the proposition that Joe Sixpack has the disposition to answer NO HELP in the relevant conditions.

We show that the probability of (P and ~Q) is at least 60%. We can understand this number as the probability that a person exhibits the full Knobe Effect under the experimental conditions of the original Knobe experiment.

Proof:

1. "Additivity" Axiom: Pr(XvY) = Pr(X) + Pr(Y) when X and Y are logically incompatible.

2.Replacing (P and Q) for X and (P and ~Q) for Y, we get:

*Pr[(P and Q) v (P and ~Q)]= Pr(P and Q) + Pr(P and ~Q).

3. Since P is equivalent to [(P and Q) v (P and ~Q)], we can substitute in * to get:

**Pr(P)=Pr(P and Q) + Pr(P and ~Q)

4. We established from our replication of the Knobe Result that Pr(P) is .78. Furthermore, Pr(Q) is .18. So Pr(P and Q) is at most .18.

5. From 4 and ** we deduce that Pr(P and ~Q) is at least .60 Hence the likelihood that a person exhibits the Knobe effect (in the intended sense) is at a minimum 60%.

Appendix D. "Further Information" Experiment Data

"Further Information" Survey Results (by ordering)

	Harm First Order (N=110)	Help First Order (N=111)
Yes Harm, Yes Help	4	15
No Harm, No Help	48	39
Yes Harm, No Help (Knobe)	53	53
No Harm, Yes Help	5	4

References

Alexander, J., and Weinberg, J. 2007: Analytic epistemology and experimental philosophy. *Philosophy Compass*, 2, 56-80.

Alter, A. L., Oppenheimer, D.M., Epley, N., and Eyre, R.N. 2007: Overcoming intuition: metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General,* 136, 569–576.

Appiah, K.A. 2008: *Experiments in Ethics. The Mary Flexner Lectures Series of Bryn Mawr College*. Cambridge: Harvard University Press.

Bealer, G. 1999: Intuition and the autonomy of philosophy. In M. DePaul and W. Ramsey (eds) *Rethinking Intuition: The Psychology of Intuition and Its Role In Philosophical Inquiry.* Lanham, Maryland: Rowman and Littlefield Publishers Inc.

Bengson, J., Moffett, M. and Wright, J. 2007: Know-how and concept possession. *Philosophical Studies*, 136, 31-57.

Cokely, E.T. and Feltz, A. 2009. Individual differences, judgment biases, and Theory-of-Mind: Deconstructing the intentional action side effect asymmetry. *Journal of Research in Personality*. Volume 43, Issue 1, 18-24.

Cushman, F., Young, L., and Hauser, M. 2006: The role of conscious reasoning and

intuition in moral judgment: testing three principles of harm. *Psychological Science,* 17, 1082-1089.

Cushman, F. and Mele, A. 2008: Intentional action: two and half folk concepts. In J. Knobe and S. Nichols (eds) *Experimental Philosophy*. New York: Oxford University Press.

Epstein, S. 1994: Integration of the cognitive and psychodynamic unconscious. *American Psychologist*, 49, 709-724.

Evans, J. 2003: In two minds: dual process accounts of reasoning. *Trends in Cognitive Sciences,* 7, 454–459.

Evans, J. 2006: The heuristic-analytic theory of reasoning: extension and evaluation. *Psychonomic Bulletin and Review,* 13, 378–395.

Evans, J. 2008: Dual-processing accounts of reasoning, judgment, and social cognition". *Annual Review of Psychology*, 59: 255-278.

Frederick, S. 2005: Cognitive reflection and decision making. *The Journal of Economic Perspectives*, 19, 25-42.

Gigerenzer, G. and Regier, T. 1996: How do we tell an association from a rule?

Comment on Sloman (1996). *Psychological Bulletin*, 119, 23-26

Haidt, J. 2001: The emotional dog and its rational tail. *Psychological Review,* 108, 814-834.

Harman, G. 1976: Practical reasoning. *Review of Metaphysics*, 29, 431-63.

Kahneman D, and Frederick, S. 2002: Representativeness revisited: attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, and D. Kahneman (eds) *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge, UK: Cambridge University Press.

Kahneman, D. 2003: A perspective on judgment and choice: mapping bounded rationality. American Psychologist, 58, 697-720

Kahneman D, and Frederick, S. 2005: A model of heuristic judgment. In K. Holyoak and R.G. Morrison (eds) *The Cambridge Handbook of Thinking and Reasoning*. Cambridge, UK: Cambridge University Press.

Kauppinen, A. 2007: The rise and fall of experimental philosophy. *Philosophical Explorations*, 10, 95-118.

Keren, G., and Schul, Y. (Forthcoming): Two is not always better than one: A critical

evaluation of two system theories. *Perspectives on Psychological Science*.

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. 2006: The concept of intentional action: a case study in the uses of folk psychology. *Philosophical Studies*, 130, 203-231.

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

Kripke, S., 1980: *Naming and Necessity*. Harvard University Press.

Leslie, A., Knobe, J. and Cohen, A. 2006: Acting intentionally and the side-effect effect: 'theory of mind' and moral judgment. *Psychological Science*, 17, 421-427.

Liao, M. 2008: A defense of intuitions. *Philosophical Studies*, 140, 247-262.

Lieberman, M.D. 2000: Intuition: a social cognitive neuroscience approach. *Psychological Bulletin*, 126, 109-137.

Lieberman, M.D. 2007: The X- and C-systems: the neural basis of automatic and controlled social cognition. In E. Harmon-Jones and P. Winkelman (eds) *Fundamentals of Social Neuroscience*. New York: Guilford.

Livengood, J. and Machery, E. 2007: The folk probably don't think what you think they think: experiments about causation by absence. *Midwest Studies in Philosophy*, 21, 107-127.

Machery, E. 2008: The folk concept of intentional action: philosophical and experimental issues. *Mind & Language*, 23, 165-189.

Machery, E., Mallon, R., Nichols, S., and Stich, S. 2004: Semantics, cross-cultural style. *Cognition* 92, B1-B12.

Mallon, R., Machery, E., Nichols, S., and Stich, S. 2009: Against arguments from reference. *Philosophy and Phenomenological Research* 79:2, 332-356.

Nahmias, E., Morris, S., Nadelhoffer, T., and Turner, J. 2006: Is incompatibilism intuitive?. *Philosophy and Phenomenological Research* 73: 28-53

Nichols, S., Stich, S., Weinberg, J. 2003: Meta-skepticism: meditations in ethno-epistemology. In S. Luper (ed) *The Skeptics: Contemporary Essays*. Burlington, VT: Ashgate.

Nichols, S., and Ulatowski, J. 2007: Intuitions and individual differences: the Knobe Effect revisited. *Mind and Language*, 22, 346-365.

Nisbett R., Peng, K., Choi, I., Norenzayan, A. 2001: Culture and systems of thought: holistic vs. analytic cognition. *Psychological Review,* 108, 291–310.

Osman, M. 2004: An evaluation of dual-process theories of reasoning. *Psychonomic Bulletin & Review*, 11, 988-1010.

Sloman S.A. 1996: The empirical case for two systems of reasoning. *Psychological Bulletin,* 119, 3–22.

Sosa, E. 2005: A defense of the use of intuitions in philosophy. In M. Bishop and D. Murphy (eds.) *Stich and his Critics.* Blackwell Publishers.

Sosa, E. 2007: Experimental Philosophy and Philosophical Intuition. *Philosophical Studies,* 132(1), 99-107.

Stanovich, K., and West, R. 1998: Individual differences in rational thought. *Journal*

of Experimental Psychology, 127, 161-188.

Swain, S., Alexander, J., and Weinberg J. (2008): The instability of philosophical intuitions: running hot and cold on truetemp". *Philosophy and Phenomenological Research*, 76, 138 – 155.

Tversky, A. and Kahneman, D. 1974: Judgment under uncertainty: heuristics and biases. *Science*, 185, 1124-1131.

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

Weinberg, J., Alexander, J., and Gonnerman, C. manuscript: Unstable intuitions and need for cognition: how being thoughtful sometimes just means being wrong in a different way.

Williamson, T. 2007: *The Philosophy of Philosophy*. Blackwell Publishing.