Abstract

A striking finding in experimental philosophy is that attributions of intentionality are sometimes influenced by seemingly external moral considerations. The most notable case of this phenomenon involves the intentionality of causing side-effects; causing a good side-effect is judged as unintentional while causing a bad side-effect is judged as intentional (Knobe, 2003a). I present a number of experimental manipulations among participants considering these types of cases that demonstrate the effect of increased reflective thinking. In the experiments presented, increased reflection leads participants to judge that both good and bad side-effect cases are unintentional. I argue that if our goal is conceptual analysis of ‘intentionality,’ increased reflection could lead to a clearer understanding of that concept. That is, if we are concerned with the ordinary concept of intentionality, the side-effect effect might be partly the result of insufficient reflection and conceptual clarity.

Keywords: intention; intentionality; side-effect effect

Introduction

The folk concept of intentionality has been a recent topic of importance in the philosophy of action. A striking finding is that attributions of intentionality are sometimes influenced by seemingly external moral considerations (Knobe 2003a; Nichols, 2004; Leslie, Knobe & Cohen, 2006). The most notable case of this phenomenon involves the intentionality of causing side-effects. Side-effects are typically thought of as events or outcomes whose consequences are known but unintended.\(^1\) Broadly, the pattern is that causing certain side-effects is judged as more intentional when the result is bad rather than good (Knobe, 2003a). This phenomenon is sometimes colloquially called the ‘Knobe effect,’ but here I will refer to it as the side-effect effect.

Here I survey a number of experimental manipulations among participants considering the side-effect effect cases, which demonstrate the effect of font differences on intentionality attributions in side-effect cases, as well as in the Cognitive Reflection Task (Frederick, 2005). I demonstrate that a font manipulation, by decreasing participant reading fluency, causes increased reflection and lower ratings of intentionality in certain side-effect cases. I argue this study suggests the side-effect asymmetry may not be as troubling as initially thought.

In the study presented, the font (increased reflection) manipulation causes participants to increasingly judge that both good and bad side-effect cases are unintentional. In a survey of other experiments, increased reflection and deliberative thinking lead to this same pattern of judgments about side-effect cases. I argue that if our goal is conceptual analysis of ‘intentionality,’ increased reflection could lead to a clearer understanding of that concept. That is, if we are concerned with the ordinary concept of intentionality, the side-effect effect might be partly the result of insufficient reflection and conceptual clarity.

The Side-Effect Effect

Knobe (2003a) has shown that when judging whether an action’s side effect is intentional, ordinary people’s attributions of intentionality are affected by the moral valence of the side effect. Often, bad side effects are judged as being brought about intentionally, while good side effects are judged as unintentional. This effect has been shown across a number of different scenarios (Knobe, 2003b; Knobe, 2004a; Knobe, 2004b; Knobe, 2006; Mele & Cushman, 2007; Cushman & Mele, 2008), as well as in different cultures (Knobe & Burra, 2006). One of the best, and certainly the most cited, pair of cases demonstrating the side-effect effect is the harm/help chairman case pair from Knobe’s original paper (Knobe, 2003a):

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 the 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.

In the harm case, most participants judge that the chairman did intentionally harm the environment, but in the help case, most participants tend to judge that the chairman...
did not intentionally help the environment (Knobe, 2003a).²

A plausible interpretation of the Knobe results is that the participants’ reports provide insight into ordinary people’s concept of ‘intentionally.’³ If we take the Knobe studies as indicators of ordinary people’s concept of ‘intentionally,’ a first thought might be that the side-effect effect is quite surprising; who knew that seemingly external moral considerations had an influence on how we perceive the intentionality of actions? Knobe has offered a conceptual competence model (Knobe, 2006), arguing the differences in intentionality attributions between good and bad consequence side-effect cases reflect the ordinary concept of intentionality. That is, the differences in intentionality attributions based on moral valence are not irrational; the concept of intention has a (surprising) normative component.

It may help to consider an alternative view of the side-effect effect through Chomsky’s (1965) performance/competence distinction. Knobe’s conceptual competence model views the side-effect effect as part of the competent speaker’s concept of intention, but others (e.g., Alicke, 2008) view the asymmetry between good and bad side effect cases more like a bias, or performance error, rather than as a reflection of conceptual competence. Here I will argue for an interpretation closer to this second view.

**Experiments**

Psychological work has shown that processing fluency, the ease with which material is processed, influences various types of judgment (see Alter & Oppenheimer (2009) for review). Specifically, increasing cognitive disfluency has been shown to increase reflective or deliberative modes of cognitive processing and decrease decision biases (Alter et al., 2007). I hypothesized that decreasing the fluency of the hypothetical intentionality attribution scenarios would reduce attributions of intentionality in both good and bad side-effect cases. Specifically, a font manipulation designed to decrease fluency would increase reflection, resulting in lower levels of intentionality attribution. To test this hypothesis I conducted three experiments. The first two tested a font manipulation on intentionality attributions and the third tested the font manipulation on a measure of reflection, the Cognitive Reflection Test.

**Experiment 1: Font and Intentionality Attributions**

Experiment 1 tested whether font differences change attributions of intentionality in the original Knobe cases (Knobe 2003a). The prediction is that decreasing the fluency of the intentionality attribution question by changing font size and boldness unexpectedly will cause participants to reflect more thoroughly on the question and result in lower intentionality attributions.

**Methods** 79 undergraduate students (Mage = 20.3, 46 male) were approached at University bus stops to complete a paper survey. Participants were randomly assigned to either the help or harm condition crossed with a bold or control font condition. Each participant received one paper survey with one scenario and corresponding questions. Participants then completed a brief demographic questionnaire.

The help/harm vignettes were those used by Knobe (2003a). Participants in the help condition were asked to rate from 1 (do not agree) to 7 (agree) whether or not they agree with the following statement:

The chairman intentionally helped the environment.

In the harm condition, participants rated agreement with:

The chairman intentionally harmed the environment.

In control conditions, this agreement statement was printed in 12 point font, the same as the surrounding text. In bold conditions, the agreement statement was printed in larger point, bold font.

**Results** A 2(font: bold, control) x 2(condition: harm, help) ANOVA indicated a main effect of font on intentionally ratings; bold condition ratings (M=3.43, SD=2.24) were lower than the control condition ratings (M=4.08, SD=2.39), F(1, 79) = 4.54, p < .05, see Figure 1 below.

Ratings for the control harm (M=6.11, SD=.99) were higher than for bold harm (M=5.00, SD=1.94), and ratings for control help (M=2.15, SD=1.56) were higher than bold help (M=1.85, SD=1.13). The font effect was significant within the harm condition, p < .05, but not within the help condition, p = .49. There was no condition x font interaction, F(1, 79) = 3.20, p = .23.

**Figure 1: Mean Intentionality Ratings in Experiment 1**
Discussion These results suggest that decreasing fluency through a font effect manipulation causes lower attributions of intentionality in the harm and help side-effect cases. Though the font effect was not significant within the help condition it is possible that the agreement scale caused a floor effect; the help condition ratings may have been too low already for the font manipulation to lower them significantly more.

Experiment 2: Font, Order, and Intentionality Attributions

Experiment 2 sought to replicate Experiment 1 in a larger online study and to use a test measure that could avoid the possible floor effect in Experiment 1. The design also provided a test of possible order effects between help and harm scenarios.

Methods 90 participants ($M_{age} = 28.5, 47$ male) were recruited through Amazon Mechanical Turk to complete an online study. Participants were randomly divided into bold or control conditions and received both the help and harm scenario and corresponding questions in a randomly determined order. The harm condition questions were as follows:

Please tell us whether or not you agree with the following statement:

The chairman intentionally harmed the environment.

Agree Disagree

How confident are you in your answer to the previous question?

(not at all confident) 1 2 3 4 5 6 7 (completely confident)

As in Experiment 1, bold condition surveys had a larger, bolder agreement statement (“The chairman intentionally harmed [helped] the environment”) and help condition vignettes and questions replaced all mentions of ‘harm’ with ‘help.’ Participants then received three comprehension questions and a demographic questionnaire. I calculated an intentionality attribution ‘score’ for each participant’s responses by scoring ‘agree’ as 1 and disagree as -1 and multiplying by the corresponding confidence answer (see Starmans and Friedman (2012) for similar scoring methodology).

Results A 2(font: bold, control) x 2(order: harm first, help first) ANOVA showed a main effect of order on the harm score; when harm was presented first the harm score was higher ($M=4.10, SD=4.80$) than when harm was second ($M=1.46, SD=6.23$), $F(1, 90) = 8.76, p < .01$. This is consistent with Feltz and Cokely’s (2007) reported order effect; intentionality ratings were lower for the harm case when it was presented second (after the help case), but no effect is reported for the help case when it is presented second.

There was a significant difference between bold and control condition help score ratings when help was presented first; bold (help first) help scores were lower ($M=-3.75, SD = 4.78$) than control (help first) help scores, $p < .05$. No other font effects were found between pairs of help and harm scores presented in the same order.

Discussion These results provide further support for the hypothesis and indicate that the font effect is not exclusive to the harm question; the help score when presented first was lowered by the font manipulation. This experiment also revealed that both font and order effect intentionality attributions. Order can also be seen as a measure of increasing reflection; a scenario presented second is presumably no less reflectively considered than a scenario presented first. When the harm case was presented second, intentionality ratings were lower, but there was no corresponding effect for the help case, indicating reflection induced by order lowers intentionality attributions in side-effect cases.

Experiment 3: Font Effects and Intentionality Attributions

There are a number of possible explanations for the results of Experiments 1 and 2. The favored hypothesis is that the font manipulation causes greater levels of reflection and this reflection pushes participants towards a negative answer to intentionality attribution questions. A plausible competing hypothesis is that the font manipulation just makes participants more skeptical and therefore less likely to agree with any statement.

To distinguish between these two hypotheses and demonstrate the font effect is one that increases reflection, I conducted one further experiment. I ran a similar study to Experiment 1 but with questions with objectively true answers that can be obtained through careful reflection. The experiment used the Cognitive Reflection Test (“CRT,” Frederick, 2005). The CRT uses questions that have objective correct answers, but also wrong answers that come easily to mind:

1. A bat and ball cost $1.10. The bat costs $1.00 more than the ball. How much does the ball cost?

2. It takes 5 machines 5 minutes to make 5 widgets. How long would it take 100 machines to make 100 widgets?

3. In a lake there is a patch of lily pads. Every day, the patch doubles in size and it takes 48 days for the patch to cover the entire lake. How many days would it take for the patch to cover half the lake?

For instance, the answer to question 1 is not $0.10, though this answer often comes to mind. Instead the correct answer
is $0.05. The CRT is correlated with performance on intelligence tests and is a predictor of performance on heuristics and biases tasks (Toplak et al. 2011). We might best understand performance on the CRT as a measure of deliberative ability or reflection. Success on the CRT typically consists in overcoming the temptation to give the obvious, but wrong answer.

The reflection hypothesis predicts that the disfluent font manipulation should cause participants to reflect more and thus pick correct answers more often with higher levels of confidence on the CRT. The skeptical hypothesis, on the other hand, predicts that the font manipulation would cause lower levels of agreement and confidence with any proposed answer to CRT questions, whether they are right or wrong.

Methods 72 participants were recruited on Amazon Mechanical Turk (Mage = 31.5, 42 male). Participants were randomly divided into control or bold font conditions, and received two Cognitive Reflection Tasks in one of two orders; participants either received surveys with the correct answers or intuitive but incorrect answers. For instance, question 1 from the control survey in the correct answer condition read:

1. A bat and ball cost $1.10. The bat costs $1.00 more than the ball.

Please tell us whether or not you agree with the following statement:

The ball costs 5 cents.

Agree Do not agree

Participants also rated their confidence in their answers on a scale from 1 (not at all confident) to 7 (completely confident).

Bold surveys had the agreement statements (e.g. The ball costs 5 cents) bolded and printed in larger font and the incorrect answer condition surveys contained the statements: (i) The ball costs 10 cents, (ii) It would take 100 machines 100 minutes to make 100 widgets, and (iii) It would take 1 day for the patch to cover half the lake. These answers, though incorrect often come to mind when people first encounter the CRT.

A total was calculated for each participant by scoring a correct response as 1 and an incorrect response as -1 and summing these. A score was calculated for each question for each participant by scoring a correct response as 1 and an incorrect response as -1 and multiplying by the answer to the corresponding confidence question. In the correct answer condition, ‘agree’ is scored as 1 and in the incorrect answer condition ‘agree’ is scored as -1.

Results A 2(font: bold, control) x 2(condition: correct answer, incorrect answer) ANOVA, using total number of correct CRT responses and CRT score as dependent variables indicated a main effect of font on the number of overall correct responses, p < .01. Participants answered 49% of the questions correctly in the control condition and 65% in the bold condition. There was also a significant difference between the overall bold and control scores; bold scores (M = 1.90, SD = 6.11) were higher than control scores (M = .04, SD = 6.13), p < .01.

Interestingly, the differences between bold and control conditions were significant within only the correct answer condition. In the correct answer condition, there was a significant difference in the overall number of correct responses in bold and control conditions. Participants agreed with the correct responses 53% of the time in the control condition and 73% of the time in the bold condition. Fisher’s Exact Test, p < .05. There was also a significant difference between the bold and control average score within the correct answer condition, (Bold M=2.94, SD =5.78, Control M=.54, SD =6.14, p = .05).

There were only marginal differences between bold and control conditions within the incorrect answer condition. Participants in the control condition agreed with correct responses 45% of the time and participants in the bold condition agreed with correct responses 57% of the time, Fisher’s Exact Test, p = .08. The mean CRT score in the control was -.62 (SD=6.08) while the mean CRT score in the bold condition was .87 (SD=6.27), p = .08. See Figure 2 below.

Discussion These findings indicate that the font manipulation increased reflection, as measured by the number of correct answers and the correct answer score on the CRT. That the effect was stronger in the correct answer condition than in the incorrect answer condition suggests a certain limitation of the font effect as applied to the CRT; for some, increased reflection may be enough to recognize a correct answer but not enough to produce a correct answer.

Notably, these results indicate the failure of the skeptical hypothesis; the font effect was significant in the correct answer condition. In this condition, higher levels of agreement, not disagreement, with the statement counts as an increase in correct responding. The skeptical hypothesis would predict lower levels of agreement in both the correct and incorrect answer conditions.
Some might think a difference in font should be irrelevant to performance on the CRT. But if we take the CRT to measure reflection, the seemingly irrelevant disfluent font manipulation increases reflection, as evidenced by improved CRT scores. I offer this as an explanation for the role of the font manipulation in the intentionality attribution cases in Experiments 1 and 2; there the font manipulation also causes greater reflection, and this reflection pushes participants towards lower intentionality attributions in both harm and help cases. It is worth noting that Pinillos et al. (2011) found that better performance on the CRT lessened the harm/help intentionality asymmetry in these same harm and help cases. Importantly, Pinillos et al. (2011) found the greater reduction comes in harm ratings. Those performing well on the CRT attributed intentionality 14% less in the harm case than those performing poorly. The increase in intentionality attributions in the help cases was only 4% for those who performed well on the CRT.

**General Discussion**

The experiments presented here represent a number of ways in which attributions of intentionality can be altered in both good and bad consequence side-effect cases. What these experiments together suggest, though, is that while participants vary in their attributions of intentionality, those attributing less intentionality in the harm and help cases seem to employ a more reflectively produced concept of intentionality. That is, the ordinary concept of intentionality reported under a range of reflection inducing and tracking conditions – from different questions orderings to fluency changes – is one in which both harm and help cases are judged as intentional less often. In no experiment have we seen increased reflection cause a significant increase in attributions of intentionality for either the harm or help case.

One might object here that just because reflection leads to different intentionality attributions does not imply that reflection leads to preferable intentionality attributions. It is quite reasonable to avoid the assumption that reflective practices are necessarily good practices. A number of studies have shown that reflection does not always lead to greater satisfaction (e.g. Dijksterhuis & van Olden (2006), Wilson & Schooler (1991)). One response to this objection is to recall the domain and context of the current research. The goal is to analyze the ordinary concept of ‘intentionality.’ Many might think that as this practice typically proceeds in philosophy, thorough reflection is good practice.

Some may find this response unsatisfactory; philosophers do not unanimously endorse reflection as always preferable to (unreflective) intuition. Aristotle suggests that the person who has mastered virtue will not need to deliberate (or reflect) for long, and Bernard Williams (1981) criticizes some moral agents for having “one thought too many,” deliberating too thoroughly about what should be a clear moral choice. If this reflection-as-preferable response is unsatisfactory, we could weaken the claim about intentionality; rather than claiming the reflective concept of intentionality represents a concept unmarred by performance errors, we may claim that there exist two concepts of intentionality. The ‘reflective intentionality’ concept represents the concept in which moral considerations do not impact intentionality judgments (as strongly), while the ‘non-reflective intentionality’ concept represents the side-effect effect laden type of intentionality. The reflective manipulations here simply encourage employment of the ‘reflective intentionality’ concept. It may come as no surprise that philosophers, who are used to running through mental comparisons, thought experiments, and reflections, sometimes seem to prefer the ‘reflective intentionality’ concept and are sometimes surprised by the ordinary employment of the ‘non-reflective intentionality’ concept and the prevalence of the side-effect effect.

**Conclusion**

There are further questions to ask about the concept of intentionality and side-effect cases. There may be multiple concepts of intentionality, perhaps context sensitive ones. The argument here can remain neutral on these issues. What is central is that in more reflective responses to the side-effect effect cases, seemingly external moral considerations have less of an influence on the concept of intentionality. The side-effect effect might be seen as a performance error or simply an employment of a ‘non-reflective intentionality’ concept. Interpreted in either way, the side-effect effect has different implications for philosophy than currently thought. Under the first interpretation, if we view the side-effect effect as a performance error, we then have reason to question the use of the Knobe (2003a) Harm case findings in a theory of intentional action. Under this interpretation, the experiments here suggest judgments in the Harm case are partially a result of insufficient reflection. Harm case intentionality attributions do not accurately reflect the competent language user’s concept of intentionality.

There is an alternative interpretation. We might instead recognize the existence of multiple concepts of intentionality. The experiments presented here suggest two of these concepts might differ based on the level of reflection with which people use to employ the concept. Under this interpretation, we have a ‘non-reflective intentionality’ and a ‘reflective intentionality.’ The non-reflective intentionality corresponds to the concept developed in Knobe’s conceptual competence model, while the reflective intentionality corresponds to the concept of intentionality that people seem to have when they think more reflectively about intention – one in which both Harm/Help side-effect actions are judged as unintentional.

This second interpretation may seem like a less dangerous one to take, as it does not require rejection of the conceptual competence model; instead, it integrates that model into the theory of the non-reflective intentionality concept. Yet, this interpretation comes with two noteworthy costs. In
admitting that what philosophers thought was a single concept of intentionality is not a single concept, the interpretation opens the door to eliminative materialist-style objections. A second worrying issue is that level of reflection is a fact about the concept-user, rather than part of the internal logical structure of a concept. It is unusual for facts about concept-users to divide concepts, and allowing this type of division may raise problems for the methodology of intuition-based philosophy and conceptual analysis. In addition to analyzing concepts based on their internal structure, we must also take as relevant to the concept facts about the mental state (e.g. level of reflection) of the concept user. The costs of this second interpretation appear too high and too far-reaching. Thus, of the two interpretations presented I favor the former.

Acknowledgments

Great thanks to Regina Rini for comments on a previous version of this paper and to Joshua Knobe and the audience of Topoi’s Conference on Intentions for helpful discussion.

References


