LEBS
Vol. 14 No. 2 (2023) 37–42
Letters on Evolutionary Behavioral Science

Is the Knobe Effect due to Error Management? A Functional Approach to the Side-Effect Effect

Ryo Oda *

Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya 466-8555, Japan

*Author for correspondence (oda.ryo@nitech.ac.jp)

The Knobe effect (side-effect effect) is a tendency to ascribe intentionality in cases of negative, but not positive, side effects. From an adaptive point of view, one possible function of attributing intentionality is to make it easier to predict the actor's future actions. When the cost of false negatives (i.e., missing an existing intentionality) is high, the need to deal with future actions of the perpetrator increases, which would lead to an increase in the degree of intentionality attribution. This study uses the "lieutenant scenario" to examine whether increasing the severity of side-effect outcome, as the cost of false negatives, facilitates intentionality attribution. Although the side-effect effect was replicated, the results show that the severity of the consequences of the effect did not affect the magnitude of intentionality attribution. Only the positive or negative outcome known to the actor affected the magnitude of intentionality attribution, which might be consistent with the idea that the Knobe effect arises as a result of responding to different mental states of the actor.

Keywords
experimental philosophy, Knobe effect, side-effect effect, intentionality, error management theory, false negative

Introduction

People tend to attribute intentionality to negative, but not positive, side effects, which is known in the literature as the Knobe effect (Knobe, 2003a). One of the scenarios used in the first experiment conducted by Knobe (2003a) represented a "help" version where the side effect was positive, while the other contained a "harm" version where the side effect was negative. The "harm" version scenario was as follows:

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 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.

The "help" version scenario was as follows:

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 help the environment.' The chairman of the board answered, 'I don't care 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.

Participants were asked whether the chairman intentionally harmed or helped the environment (depending on the version of the scenario). It was found that they were more likely to attribute intentionality when the side effect was negative (82%) than positive (23%).

Since the Knobe effect was discerned in the context of philosophy, numerous empirical studies have attempted to explain the observed effect, and the reason for the effect has been considered mainly in terms of concept analysis (e.g., Nado, 2008), semantics (e.g., Mizumoto, 2017) or psychological processes (e.g., Knobe, 2010). The purpose of this study is to investigate the effect in terms of function as a result of adaptation. The first point to be clarified is that the intention with which an actor performs a certain action and the intentionality that a bystander attributes to the action cannot be treated as the same thing. Why, then, do we attribute intentionality to the actions of others? Clark (2022) pointed to the relevance of error management theory (EMT; Haselton & Buss, 2000) to the Knobe effect. In her "Blame Efficiency Hypothesis," Clark (2022) argued that people would be more likely to assume that a person has the requisite characteristics for responsibility in cases of harmful rather than helpful or neutral behaviour because it would be more costly to mistakenly fail to blame a perpetrator who could have been deterred by blame (i.e., a false negative in EMT) than to mistakenly blame a perpetrator who could not have been deterred by blame (i.e., a false positive in EMT). Although the Blame Efficiency Hypothesis is a pioneering hypothesis that considers the Knobe effect in terms of adaptation, it has not been empirically tested. If the side-effect effect can be explained by EMT, then increasing the cost of false negatives would facilitate the degree of intentionality attribution because the optimal threshold for adopting a belief of responsibility decreases as the cost increases (Haselton & Nettle, 2006). The decrease in the magnitude of intentionality attribution could therefore be caused by the decrease in the cost of false negatives.

Although Clark (2022) emphasizes attributions of responsibility and blame, as these would be linked to attribution of intentionality, attribution of intentionality

doi: 10.5178/lebs.2023.107
Received 29 May 2023.
Accepted 02 July 2023.
Published online 23 July 2023.
© 2023 Oda.

This work is licensed under the Creative Commons Attribution 4.0 International License.
To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

does not always lead to blame. Indeed, Knobe (2003b) found a case where participants attributed a low degree of intentionality but assigned a high degree of blame. One possible function of attributing intentionality is to make it easier to predict an actor's future actions. When the cost of false negatives is high, the need to deal with the future actions of the perpetrator increases, which would lead to an increase in the degree of intentionality attribution. In other words, intentionality attribution makes it easier to control situations. The fact that the Knobe effect can be seen not only for intentions but also for desires or attitudes supports this hypothesis because attributions of desire and attitude also function to address others' behaviour as elements of theory of mind (e.g., Guglielmo & Malle, 2010; Pettit & Knobe, 2009). Thus, it could be that it is not blame that leads to the intentionality attribution; rather the cost of false negatives might directly increase the degree of intentionality attribution.

This study examines whether increasing the cost of false negatives (i.e., the severity of side-effect outcome) facilitates intentionality attribution. I modified the scenario used in Knobe's (2003a) second experiment to control for the cost of false negatives. The original scenario with the "harm" version was as follows:

A lieutenant was talking with a sergeant. The lieutenant gave the order: 'Send your squad to the top of Thompson Hill.' The sergeant said: 'But if I send my squad to the top of Thompson Hill, we'll be moving the men directly into the enemy's line of fire. Some of them will surely be killed!' The lieutenant answered: 'Look, I know that they'll be in the line of fire, and I know that some of them will be killed. But I don't care at all about what happens to our soldiers. All I care about is taking control of Thompson Hill.' The squad was sent to the top of Thompson Hill. As expected, the soldiers were moved into the enemy's line of fire, and some of them were killed.

And the original scenario with the "help" version was as follows:

A lieutenant was talking with a sergeant. The lieutenant gave the order: 'Send your squad to the top of Thompson Hill.' The sergeant said: 'If I send my squad to the top of Thompson Hill, we'll be taking the men out of the enemy's line of fire. They'll be rescued!' The lieutenant answered: 'Look, I know that we'll be taking them out of the line of fire, and I know that some of them would have been killed otherwise. But I don't care at all about what happens to our soldiers. All I care about is taking control of Thompson Hill.' The squad was sent to the top of Thompson Hill. As expected, the soldiers were taken out of the enemy's line of fire, and they thereby escaped getting killed.

Although the "lieutenant scenario" has been used less frequently than the "chairman scenario" in previous studies of the Knobe effect, there is a replication study in Japan. In his Study 3, Nakamura (2018) presented the translated "harm" version to 64 participants and the "help" version to 59 participants and asked them to estimate the lieutenant's intentionality on an 8-point scale. Japanese participants estimated the lieutenant's intentionality significantly higher in the "harm" version than in the "help" version. The reason why I used the "lieutenant scenario" in the present study is that, in this scenario, the side-effect can be quantitatively manipulated. In the "chairman scenario", which has often been mentioned as an example of the side-effect effect, the side-effect is harm or help to the environment, which is difficult to quantitatively manipulate in terms of the severity of the outcomes. In the "lieutenant scenario", however, the side-effect is casualties, and the effect of the side-effect can be evaluated by manipulating the number of casualties in the squad. If the cost of false negatives affected the degree of intentionality attribution, then the cost of not attributing intentionality to the "harm" version in which no one was killed would not differ from the cost of not attributing intentionality to the "help" version, and the more casualties that occurred as a side-effect, the greater the degree of intentionality that would be attributed to the negative outcomes.

Methods

Questionnaire

On the website, participants read the "lieutenant scenario" modified so that the number of squad members in the scenario was set at 10. Four versions of the scenario were created: One used the "help" version, in which "they thereby escaped getting killed" was changed to "thereby no one was killed," and the other three utilized the "harm" version, in which "some of them were killed" was changed to: (1) no one was killed, (2) five of the 10 men were killed, and (3) all 10 men were killed. After reading a scenario corresponding to each of the four versions, participants were asked, as a comprehension check, to choose between the options regarding the side-effect (no one was killed/five of the 10 men were killed/all 10 men were killed) and the lieutenant's concern (he cared/he did not care what happened) in each scenario. Participants were then asked to rate the lieutenant's degree of intentionality on 9-point scales ranging from 0 (not intentional) to 8 (intentional), and to rate how they would praise (for the "help" version) or blame (for the "harm" version) the lieutenant on 9-point scales ranging from 0 (I do not think he should be praised/blamed at all) to 8 (I think he should be highly praised/blamed).

Participants

Three hundred and sixty-six Japanese adults (range: 18–82 years) were recruited through Cross Marketing, Inc. (Tokyo, Japan). Participants were randomly assigned to each of the four conditions. Excluding the participants who did not pass the comprehension check, data were analysed from 258 participants: 69 (36 females; median age = 53 years) for the "help" version, 62 (30 females; median age = 51.5 years) for the "harm" version with no casualties, 61 (30 females; median age = 54 years) for the "harm" version with five casualties, and 66 (34 females; median age = 51 years) for the "harm" version with 10 casualties.

Results

Figure 1 shows the distribution of the magnitude of intentionality attribution in each condition. In the "harm" version with no casualties, the median was seven and 48.4% of participants chose eight. In the "harm" version

with five casualties, the median was eight and 59.0% of participants chose eight, and in the "harm" version with 10 casualties, the median was eight and 59.1% of participants chose eight. The distributions were highly skewed towards the maximum value. The distribution of intentionality attribution in the "help" version with no casualties (median = 5), however, was not as skewed as in the "harm" versions (Figure 1).

As the distributions were far from normal, a Kruskal-Wallis test was conducted to compare the magnitude of intentionality attribution across the four conditions. The result shows that the effect of the condition was significant (χ^2 (3) = 36.13, p < .001, η^2 = .141). The magnitude of intentionality attribution in the "help" version with no casualties was significantly lower than in the "harm" versions (vs. no casualties: z = 3.70, p < .001; vs. five casualties: z = 4.90, p < .001; vs. 10 casualties: z = 4.78, p < .001). There were no significant differences, however, between each of the "harm" versions (z = 0.22–1.19).

Figure 2 shows the distribution of the magnitude of praise/blame in each condition. The distributions of blame were also highly skewed towards the maximum value, albeit not as much as in the intentionality attribution. A Kruskal-Wallis test was conducted to compare the magnitude of blame across the three conditions of the "harm" version. The result shows that the effect of the

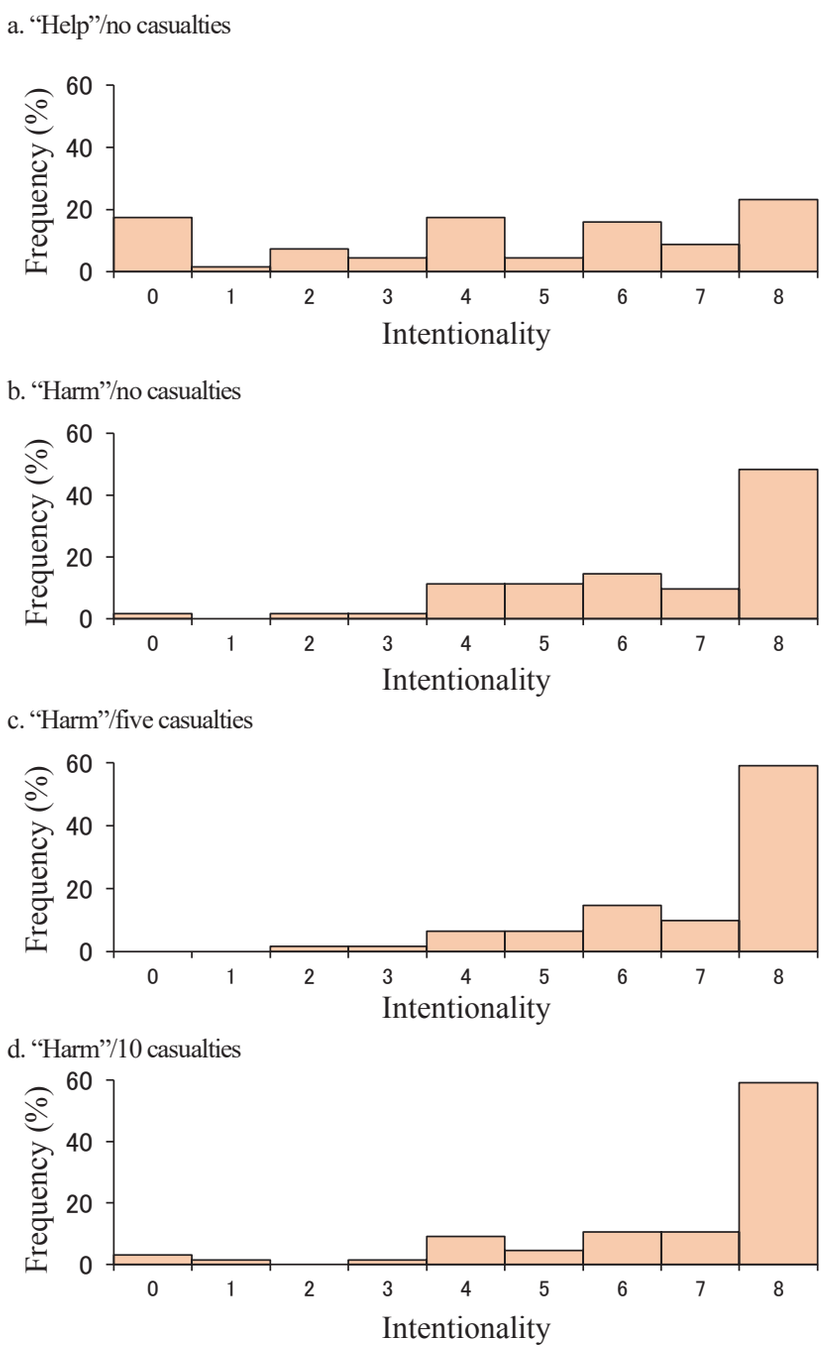

Figure 1. Distributions of the magnitude of intentionality attribution: (a) "help" version with no casualties, (b) "harm" version with no casualties, (c) "harm" version with five casualties, and (d) "harm" version with 10 casualties.

condition was significant (χ^2 (2) = 7.44, p = .024, η^2 = .040). The magnitude of blame in the "harm" version with no casualties (median = 6) was significantly lower than that in the "harm" version with five casualties (median = 7; z = 2.08, p = .038) and that in the 10 casualties condition (median = 7; z = 2.46, p = .014). There was no significant difference between the magnitude of blame in the "harm" version with five casualties and that in the 10 casualties condition (z = 0.34, p = .738). When there were casualties as a side effect, the blame was stronger than when there were none. In contrast to the intentionality attribution, the outcomes of the side-effect affected the blame ascribed to the lieutenant. The magnitude of blame did not increase, however, even if the number of casualties doubled.

Correlations were calculated between the magnitude of intentionality attribution and that of praise/blame in each condition. Spearman rank correlations were .70 (t (67) = 8.07, p < .001) in the "help" version with no casualties, .40 (t (60) = 3.47, p = .001) in the "harm" version with no casualties, .36 (t (59) = 2.97, p = .004) in the "harm" version with five casualties, and .60 (t (64) = 6.04, p < .001) in the "harm" version with 10 casualties.

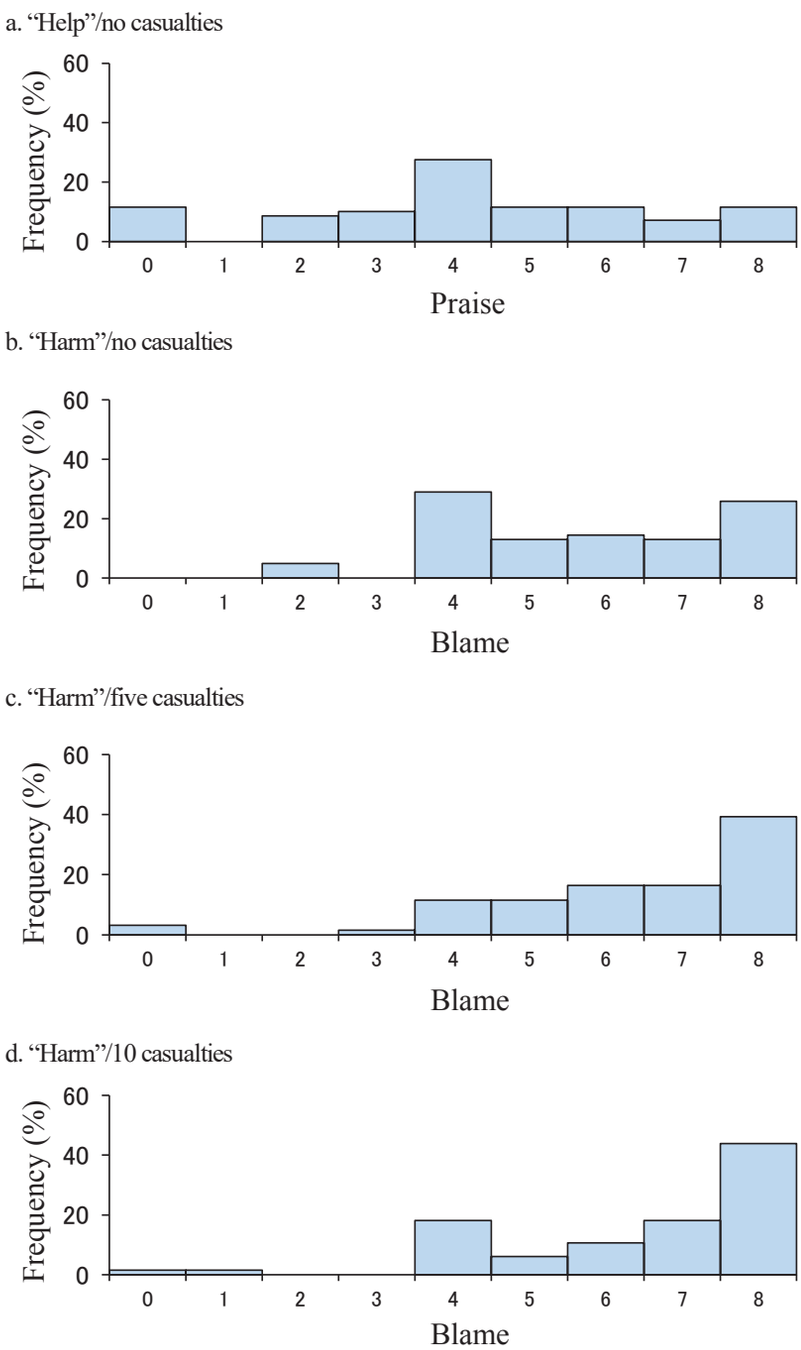

Figure 2. Distributions of the magnitude of praise/blame to the lieutenant: (a) "help" version with no casualties, (b) "harm" version with no casualties, (c) "harm" version with five casualties, and (d) "harm" version with 10 casualties.

Discussion

The distributions of intentionality attributions in the "harm" versions were highly skewed towards the maximum, suggesting that people's intentionality attributions were dichotomous. While there have been many previous studies that have measured intentionality attribution as a continuous value, it might be better to measure attribution in a dichotomous way, as in the original paper by Knobe (2003a).

A surprising result was that the magnitude of intentionality attribution increased when the content of the scenarios was changed from positive to negative, even when there were no casualties in either scenario. Furthermore, the magnitude of intentionality attribution did not increase as the number of casualties increased. Given that their comprehension was checked, it is not possible that participants did not recognize the lieutenant's concern and the number of the casualties. In fact, the "harm" version with five casualties corresponded to the original "harm" version in which "some of them were killed," and the magnitude of intentionality attribution increased significantly from the "help" version, which replicated the previous study with Japanese participants (Nakamura, 2018). The results do not support the hypothesis that increasing the cost of false negatives facilitates intentionality attribution. Only the positive or negative outcome known by the lieutenant affected the magnitude of intentionality attribution, and the attribution might be done in a dichotomous way. These results might be consistent with the idea that the Knobe effect arises as a result of responding to the different mental states of the actor, such as reasoning and norm consideration (e.g., Hindriks, 2019; Scaife & Webber, 2013).

The magnitude of praise for the lieutenant correlated strongly with that of intentionality attribution in the "help" version, whereas the magnitude of blame correlated strongly with that of intentionality attribution only in the "harm" version with 10 casualties. Although the correlations were statistically significant, those in the "harm" versions with no casualties and with five casualties were weak. The results suggest that harm is more likely to be judged as intentional regardless of the degree of blame. Although Clark (2022) emphasizes attributions of blame, as these would be linked to intentionality attributions, the relationship between blame and intentionality attributions in this study was not straightforward.

This study examined the effect of the severity of the side effects caused by the actor's previous decision on the attribution of intentionality. As I described, however, one possible function of attributing intentionality is to make it easier to predict an actor's future actions. Therefore, the severity of future negative outcomes is expected to be more relevant than the severity of the past negative side-effects. Future research should consider whether the likelihood that the actor will continue to make decisions in the future affects the magnitude of intentionality attribution. Thus, the results of the present study are not sufficient to deny that the Knobe effect is due to error management. However, the adaptive perspective has successfully explained human cognitive biases (e.g., Gigerenzer et al., 1999). If the Knobe effect is a type of cognitive bias, then further studies from an adaptive perspective might help us to better understand the reasons and mechanisms behind it.

Acknowledgments

I would like to thank an anonymous reviewer for the helpful comments. This work was supported by JSPS KAKENHI Grant Number 20H01755.

Ethical statement

This study was approved by the Bioethics Review Committee of Nagoya Institute of Technology (No. 2022-4).

Data accessibility & program code

All the data is accessible as a supplemental file.

Supplementary material

Electronic supplementary materials (original materials in Japanase) are available as a supplemental file.

References

Clark, C. J. (2022). The blame efficiency hypothesis: An evolutionary framework to resolve rationalist and intuitionist theories of moral condemnation. In T. Nadelhoffer & A. Monroe (Eds.), *Advances in Experimental Philosophy of Free Will and Responsibility* (pp. 27–44). Bloomsbury Publishing.

Gigerenzer, G., Todd, P. M., & ABC Research Group (1999). *Simple Heuristics That Make Us Smart*. Oxford University Press.

Guglielmo, S., & Malle, B. F. (2010). Can unintended side effects be intentional? Resolving a controversy over intentionality and morality. *Personality and Social Psychology Bulletin*, *36*(12), 1635–1647. https://doi.org/10.1177/0146167210386733

Haselton, M. G., & Buss, D. M. (2000). Error management theory: A new perspective on biases in cross-sex mind reading. *Journal of Personality and Social Psychology*, *78*(1), 81–91. https://doi.org/10.1037/0022-3514.78.1.81

Haselton, M. G., & Nettle, D. (2006). The paranoid optimist: An integrative evolutionary model of cognitive biases. *Personality and Social Psychology Review*, *10*(1), 47–66. https://doi.org/10.1207/s15327957pspr1001_3

Hindriks, F. (2019). Explanatory unification in experimental philosophy: Let's keep it real. *Review of Philosophy and Psychology*, *10*, 219–242. https://doi.org/10.1007/s13164-018-0397-0

Knobe, J. (2003a). Intentional action and side effects in ordinary language. *Analysis*, *63*(3), 190–194. https://doi.org/10.1093/analys/63.3.190

Knobe, J. (2003b). Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology*, *16*(2), 309–324. https://doi.org/10.1080/09515080307771

Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences*, *33*(4), 315–329. https://doi.org/10.1017/S0140525X10000907

Mizumoto, M. (2018). A simple linguistic approach to the Knobe effect, or the Knobe effect without any vignette. *Philosophical Studies*, *175*, 1613–1630. https://doi.org/10.1007/s11098-017-0926-1

Nado, J. (2008). Effects of moral cognition on judgments of intentionality. *British Journal for the Philosophy of*

Science, *59*(4), 709–731. https://doi.org/10.1093/bjps/axn035

Nakamura, K. (2018). Harming is more intentional than helping because it is more probable: The underlying influence of probability on the Knobe effect. *Journal of Cognitive Psychology*, *30*(2), 129–137. https://doi.org/10.1080/20445911.2017.1415345

Pettit, D., & Knobe, J. (2009). The pervasive impact of moral judgment. *Mind and Language*, *24*(5), 586–604. https://doi.org/10.1111/j.1468-0017.2009.01375.x

Scaife, R., & Webber, J. (2013). Intentional side-effects of action. *Journal of Moral Philosophy*, *10*, 179–203. https://doi.org/10.1163/17455243-4681004