Jan García Olier
Universität Zürich
Ethik Zentrum, Zollikerstrasse 117, 8008 Zürich, Schweiz

The 'Knobe effect' as an instance of a 'severity effect'

In the last two decades, several empirical studies have documented an asymmetry in people's assessments of intentional action, so-called 'Knobe effect' (cf. Knobe, 2003; Cushman & Mele, 2008; Knobe & Mendlow, 2004; Knobe & Burra, 2006; Nadelhoffer, 2004, 2005; for detailed reviews see Feltz, 2007; and Nichols & Ulatowski, 2007). Accordingly, foreseen (yet undesired) outcomes that are harmful are judged intentional, whereas foreseen (yet undesired) outcomes that are helpful are judged nonintentional. In a more recent study, however, Kneer & Bourgeois-Gironde (2017) found that people's ascriptions of intentionality are susceptible not only to the Knobe effect but to a 'severity effect': the more harmful a foreseen (yet undesired) outcome, the more inclined people are to say that it was intentionally brought about.

The severity effect can be perceived as a challenge to some explanations of the Knobe effect. The latter has standardly focused but on two data points: intentionality ascriptions for harmful v. helpful (or neutral) outcomes. Not surprisingly, many scholars have conceived of the Knobe effect as a *binary*, absolute effect: foreseen harmful outcomes are deemed intentional, whereas foreseen helpful —or at least not harmful— outcomes are deemed nonintentional. However, the severity effect findings suggest that things might be somewhat more complex. Rather than being of binary nature, the relation between an action's outcome and intentionality ascriptions appears to be a matter of *degrees*. It is not just that people judge harmful outcomes as intentional; it is that people's propensity to say that an outcome is intentional is commensurate with the outcome's degree of harm.

The research conducted by Kneer & Bourgeois-Gironde (2017) on the severity effect, however, suffers from a shortcoming. Focus was placed only on intentionality ascriptions for graded *harmful* outcomes (somewhat bad v. very bad); beneficial outcomes were not tested. While there is empirical evidence concerning the relation between graded harmful outcomes and intentionality ascriptions (cf. Kneer & Bourgeois-Gironde, 2017; Prochownik, 2020), the relation between differently graded positive outcomes and intentionality attributions is completely unexplored.

To address the former lacuna and provide a clearer understanding of the relation between outcomes and intentionality ascriptions, we conducted a study exploring attributions of intentionality (and knowledge) across a range of different outcomes: *very bad, somewhat bad, neutral, somewhat good and very good*. In contrast to classic side-effect effect research, we were less interested in the difference between good and bad outcomes, but in the correlation between perceived goodness and badness of outcome (which we also measured) on intentionality (and knowledge). Given that we contend that perceived blameworthiness (or praiseworthiness), rather than outcomes per se, might drive intentionality ascriptions, we were also interested in the correlation between perceived blame and praise on intentionality. Importantly, we decided to explore the positive and the negative part of the outcome spectrum, as well as the blame and praise parts, separately. In the next paragraph we will briefly summarize the results of Experiment 1 of our study.

On the negative part of the outcome spectrum, consistent with the severity-effect findings by Kneer & Bourgeois-Gironde (2017), we found a positive correlation between intentionality ascriptions and perceived outcome badness (r=.416, p <.001): the worse participants perceived the side effect to be, the more likely they were to agree with the claim that the side effect was

Jan García Olier
Universität Zürich
Ethik Zentrum, Zollikerstrasse 117, 8008 Zürich, Schweiz

intentionally brought about. On the positive side of the outcome-spectrum, by contrast, we found a weak negative correlation between intentionality ascriptions and perceived outcome goodness (r=-.192, p=.013): the more desirable participants perceived the side effect to be, the less likely they were to agree with the claim that it was intentionally brought about. Interestingly, we also found a strong positive correlation between blame and intentionality ascriptions (r = .741, p < .001), and a weaker positive correlation between praise and intentionality ascriptions (r = .385, p >.001).

The former results suggest that the Knobe effect data points are but two data points of a broader, more fine-grained phenomenon, and that Knobe effect explanations that have conceived of it in binary terms are at best incomplete. Explanations that, on the contrary, turn to gradable features such as blame or praise, seem to be on the right track. The results further suggest that the relation between intentionality ascriptions and graded outcomes is different when the valence of the latter is negative than when it is positive. Whereas an increase in the degree of harm of a foreseen outcome warrants a considerable increase in the propensity to ascribe intentionality to it, an increase in the degree of desirability of a foreseen outcome only warrants a minor (if not irrelevant) increase in people's propensity to judge it nonintentional. The results of Experiment 2, which will also be presented, demonstrate that these findings are robust.

References:

Cushman, F., & Mele, A. (2008). Intentional action: Two and a half folk concepts? In J. Knobe, & S. Nichols (Eds.). Experimental philosophy (pp. 171–188). Oxford: Oxford University Press.

Feltz, A. (2007). The Knobe effect: A brief overview. Journal of Mind and Behavior, 3(4), 265–277.

Kneer, & Bourgeois-Gironde. (2017). Mens rea ascription, expertise and outcome effects: Professional judges surveyed. *Cognition, 169*, 139-146

Knobe, J. (2003). Intentional Action and Side Effects in Ordinary Language. *Analysis, 63*(3), 190- 194.

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

Knobe, J., & Mendlow, G. (2004). The good, the bad, and the blameworthy: Understanding the role of evaluative reasoning in folk psychology. *Journal of Theoretical and Philosophical Psychology*, 24, 252–258.

Nadelhoffer, T. (2004). On praise, side effects, and folk ascriptions of intentional action. *Journal of Theoretical and Philosophical Psychology*, 24, 196–213.

Nadelhoffer, T. (2005). Skill, luck, control, and intentional action. *Philosophical Psychology*, 18, 341–352.

Nichols, S., & Ulatowski, J. (2007). Intuitions and individual differences: The Knobe effect revisited. *Mind & Language*, 22(4), 346–365.

Prochownik, K., Krebs, M., Wiegmann, A., & Horvath, J. (2020). CogSci 2020 Paper "Not as Bad as Painted? Legal Expertise, Intentionality Ascription, and Outcome Effects Revisited." Retrieved from osf.io/n9h2b