The Explanation Explanation of Knobe Effects

Numerous studies, starting with Knobe (2003), have revealed asymmetries in subjects’ judgments concerning agents’ relations to side effects that the agents do not care about. When the chairman of a board knowingly allows a profitable program that will harm the environment because he cares only about profit, subjects tend to say that the agent intentionally brought about the harm and that he is blameworthy for doing so. By contrast, when a chairman who only cares about profit knowingly allows a profitable program that will help the environment, subjects are reluctant to say that he intentionally helped the environment, or that he is praiseworthy for doing so. Similarly, subjects are more inclined to say that the agent in the harm condition was in favor of harming or that he decided to bring about the harm than they are to say that the agent in the benefit condition was in favor of helping or decided to help, and more willing to say that the agent achieved his goals by bringing about the harm, or that he brought about the harm in order to achieve his goals, than to say, in the benefit condition, that he achieved his goal by helping, or that he helped in order to achieve his goal (Pettit and Knobe 2009, e.g.).

A number of explanations have been suggested for these asymmetries. It is clear that the asymmetry is related, somehow, to the different valuation of the effects, or to the norms involved, but exactly how has been unclear. For example, the norms in question need not be endorsed by subjects attributing intentionality (Nichols & Ulatowski 2007; Knobe 2007), and the agent need not be seen as blameworthy in order for people to see him as intentionally bringing about the effect (Knobe & Mendlow 2004). Knobe suggests that these asymmetries depend on whether agents’ attitudes fall short of (or exceed) a default set by normative expectations (Knobe forthcoming). But without an account of how deviations from defaults affect judgments, this falls short of a convincing explanation. Cole Wright & Bengson (2009) argue that asymmetries in intentionally descriptions depend on attributions of responsibility, and more specifically on attributions of praiseworthiness and blameworthiness, but they admit that this cannot be the whole story, as we sometimes attribute intentionality without blameworthiness.

In this paper, I propose a way of turning hunches into a detailed account that covers all the relevant cases. According to this account, the Explanation Explanation, normative expectations affect explanatory judgments, which in turn affect the judgments subject to the Knobe effect, in particular judgments of intentionality, blameworthiness and praiseworthiness. In the harm scenario, there is a simple, intuitively straightforward explanation of the outcome in terms of the agent’s motivational states: the environment was harmed because the chair did not care enough about the environment. In the benefit scenario, there is nothing nearly as straightforward: we might say the environment was helped because the chair cared about profit and because the most profitable action happened to be have beneficial environmental effects, but that makes reference to happenstance outside the agent’s motivation.

The paper proceeds as follows. In section 1, I set up the problem and explain why extant
explanations are either mistaken or in need of supplementation In section 2, I introduce the Explanation Explanation and detail how explanatory judgments in Knobe cases are asymmetrical because they are sensitive to ways in which various factors stand out from an explanatory background by violating normative expectations. In section 3, I argue that descriptions of actions in “intentionally” contexts are strongly influenced by explanatory interests: the act needs to be described in terms appropriately connecting it to motivational states of the agent. Because of the explanation asymmetry, this has the effect of allowing “intentionally caused the harm” in the harm condition but ruling out “intentionally caused the benefit” in the benefit condition.

In section 4, I outline how the Explanation Explanation accounts for data unaccounted for by Cole Wright & Bengson as well as for correlations between responsibility judgments and intentionality judgments. The Explanation Explanation provides a straightforward account of why agents in harm scenarios are seen as intentionally bringing about the outcome even when they are not seen as blameworthy: in those cases, the outcome is still intuitively explained by the agents’ motivational states. Moreover, the correlation is well accounted for given the hypothesis that responsibility judgments are, partly, explanatory judgments. Elsewhere, I have argued that this hypothesis accounts for a wide range of phenomena involving our judgments of moral responsibility (Björnsson & Persson Forthcoming, Björnsson Forthcoming). In section 5, I supplement these arguments by reviewing new empirical work indicating a strong correlation between responsibility judgments and corresponding explanatory judgments. In section 6, finally, I look at other locutions that show milder Knobe effects than “intentionally”, “blameworthy” and “praiseworthy”, briefly suggesting that we should expect the considerations at play with “intentionally” to carry over, though less strongly. If the overall argument is correct, we now have a detailed and unified account of Knobe effects.

Knobe, J. Forthcoming: ‘Person as Scientist, Person as Moralist.’ Behavioral and Brain Sciences

1 Around 30 pairs of judgments about (i) degrees moral responsibility and (ii) acceptability of an explanation of outcome in terms of the agent’s motivation were collected for each of 15 scenarios. The correlation for all answers combined was 0.71 and highly significant.