Contextual dependence of the *Knobe effect*: reversals under counterfactual contrasts

Sean Dae Houlihan (dae.houlihan@dartmouth.edu)
Greyson Xiao (greyson.xiao.25@dartmouth.edu)
Jonathan Phillips (jonathan.s.phillips@dartmouth.edu)
Cognitive Science Program, Dartmouth College
5 Maynard Street, Hanover, NH, 03755 USA A Department, 1234 Example Street

Abstract

Lay concepts of intentionality and causality are sensitive to normative factors. Prior work on the "Knobe effect" has found that norm-violating actions are attributed greater causal importance than norm-conforming actions. Such norm-depended asymmetries in people's intuitive theory of psychology are highly replicable and represents striking deviations from prescriptively rational inference. To better understand the bases of norm-dependent asymmetries in intuitive psychology, we manipulate and measure counterfactual cognition in Knobe-like vignettes. In several preregistered studies, we replicate the Knobe effect under certain conditions. However, we also find strong reversals of the expected effect direction under other conditions. These data suggest that the classic Knobe effect depends critically on counterfactual cognition.

Keywords: Knobe effect; Side-effect effect; Counterfactual; Norm-dependence; Causal attribution; Causal judgment; Intentional action; Intuitive psychology; Intentional stance

Introduction

People's lay intuitions about intentional actions are sensitive to normative factors. Joshua Knobe famously conducted several survey studies in which participants were presented with a vignette about the chairman of a company. The scenario described the chairman implementing a profit maximizing policy knowing fully, but with no concern for, the impact on the environment. The studies found a striking asymmetry: when the policy damaged the environment, the majority of participants indicated that the chairman caused the harm intentionally. However, when the policy helped the environment, the majority of participants indicated that the chairman had not intentionally benefitted the environment (Knobe, 2003a, 2003b). This effect was generalized to showing that folk intuitions about intentional action were highly sensitive to whether the agent's action is judged to be norm-conforming or norm-violating (see, e.g., Uttich & Lombrozo, 2010). A closely related phenomenon is how norms influence attributions of causality to intentional actions (Hitchcock & Knobe, 2009): Actions that violate prescriptive norms tend to be attributed greater causal effect on downstream consequences than norm-conforming actions that result in the same outcome (Kominsky & Phillips, 2019).

Norm-dependent asymmetries such as these are prevalent in folk-psychology and highly replicable (Cova et al., 2021). Diverse proposals have been put forth to account for these empirical patterns (e.g. Cova et al., 2016; Feltz, 2007; Knobe, 2010), and while the proposals vary widely, they largely agree that norm-dependent asymmetries are surprising and warrant explanation.

One line of research argues that there is a unified explanation for these norm-dependent asymmetries which posits an underlying mechanism of differences in the relevant counterfactual contrasts (Phillips et al., 2015). This approach has been extended to argue that differences in counterfactual reasoning can explain observed asymmetries arising from both prescriptive and descriptive norms, and how these effects interact with the causal structure of the events and agents' mental states, among other things (see, e.g. Icard et al., 2017; Kirfel & Phillips, 2023; Kominsky et al., 2015).

An outstanding question is what counterfactual representations are involved. Some evidence indicates that norm violations enhance the salience of norm-conforming counterfactuals (Halpern & Hitchcock, 2015; Icard et al., 2017). This prior work has focused primarily on counterfactuals of actions and consequents. However, many other counterfactuals are possible. The goal of the ongoing work presented here is to dissect the cognitive structure of counterfactual reasoning about intentional actions.

In a series of preregistered studies (n=200 each), we employ a set of eight vignettes that follow the general pattern:

> In situation S, agent M can choose action A, which has externality E_A, or action B, which has externality E_B. Agent M does action A for the purpose of effecting outcome O, with full comprehension—and no concern—that the action will effect externality E_A and not externality E_B.

We implement a fully-crossed $2 \times 2 \times 2$ design. For each vignette we manipulate (i) whether the externality (i.e. the 'side-effect') is positive (beneficial) or negative (harmful), (ii) whether the action is norm-violating or norm-conforming, and (iii) whether the situation would default to the positive or negative externality if no action was executed by the agent (Halpern & Hitchcock, 2015). Note that the objective (i.e. the 'main-effect', outcome O) is always self-serving, just as in the original chairman paradigm.

Rowena caused the group to [catch/miss] the bus home. Jalen caused his team to [win first place / not even place] in

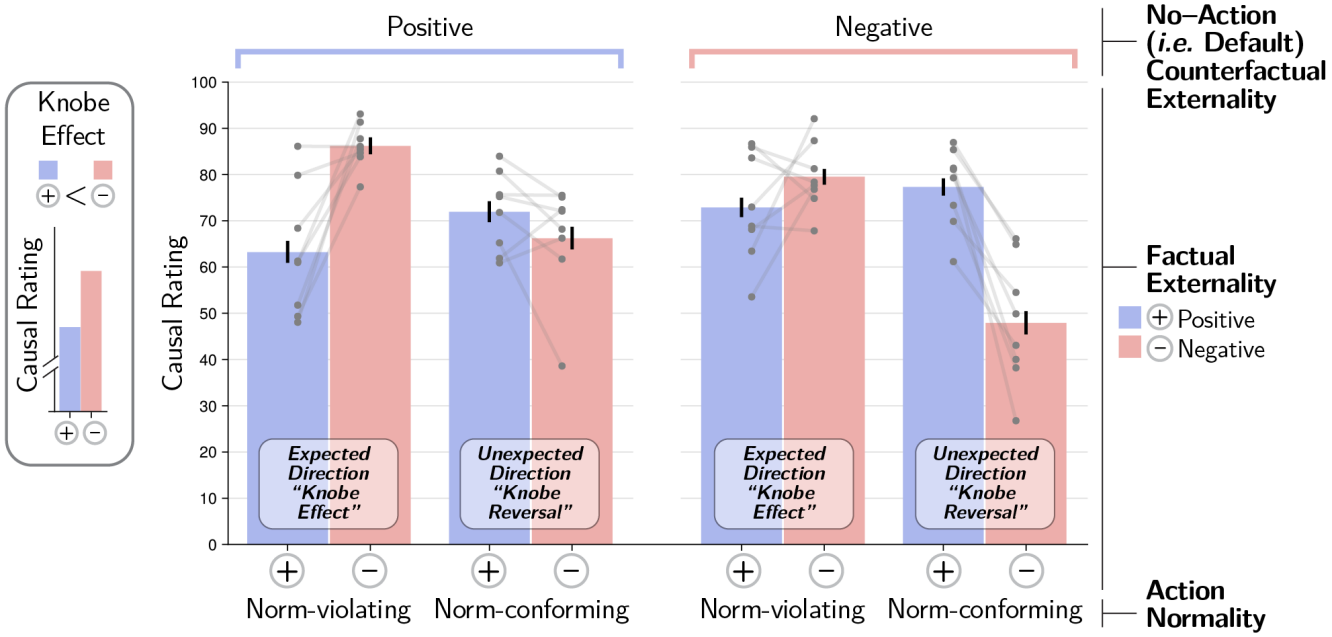

Figure 1: Mean ratings of the agent's causal role in effecting the externality. Bars give the mean across all eight vignettes and error-bars give the 95% bootstrap CI (participants resampled with replacement, $n = 10,000$ iterations). The ratings responses are normalized by the number of responses such that each vignette is equally weighted in the summary statistics. The mean rating of each vignette are shown as grey points. **(*inlay*)** Prior work on the "Knobe Effect" has established reliably asymmetry in people's intuitive theory of the causal influence of intentional actions. The established directional effect, depicted qualitatively in the inlay, is that intentional actions are rated as more causally important when associated with negative externalities, compared with positive externalities.

the competition.

We collected judgments about these 64 stimulus variations (8 conditions of 8 vignettes) from online participants. Participants made causal and counterfactual judgments about the agents, their actions, and their mental contents. For the present analysis, the data of interest are participants' causal ratings of the agents, which were prompted with statements of the form "agent M caused externality E_A". Participants rated their endorsements on a scale from 0 to 100. Consistent with prior work, we find that causal attributions are highly sensitive to counterfactuals. In one condition, we robustly replicate the "Knobe effect". When the situation would, in the absence of any action by the agent, default to a favorable externality, norm-violating actions that produce harmful externalities were judged to be more causally important to the externality (Figure 1, norm-violating action with a positive counterfactual externality).

Manipulating the "default" externality in a situation, or the norm-congruence of an action, induced dramatically different judgements of how causally important the agents were to the externalities. Most notably, the causal attribution patterns include pronounced reversals of the expected direction of the Knobe effect. For instance, in situations that would default to an unfavorable externality if the agents took no action, agents were attributed more causal influence when they performed norm-conforming actions that resulted in a positive externality (Figure 1, norm-conforming action with a negative counterfactual externality). Note that the pattern cannot be explained by whether the action produced an externality of the same valance as the default externality.

We collected several types of counterfactual judgments. In addition to judgments about counterfactual actions, participants furnished judgments of counterfactual mental states and "agent replaceability", i.e. how likely would have the externality consequence been if the agent had been replaced with (i) a similarly competent agent, (ii) a random person, (iii) no one. We find complex interactions between these counterfactual judgments and the causal attribution patterns.

In sum, our series of tightly-controlled experiments explicitly articulate necessary conditions for the "Knobe effect" to occur. The data suggest that the classic Knobe effect, and the broader class of norm-dependent asymmetries, depend critically on counterfactual cognition. At present, these data point to specific counterfactual representations likely involved, but the precise mechanisms of norm-dependence are not yet understood. Nonetheless, these results suggest ways that decades-old questions about intentionality and skill (Knobe, 2003b, 2006) can be integrated with recent work on the roles of counterfactuals in moral cognition (Wu & Gerstenberg, 2024), and efforts to uncover the core counterfactual system for social and non-social causal reasoning (Kominsky & Phillips, 2019).

References

Cova, F., Lantian, A., & Boudesseul, J. (2016). Can the Knobe Effect Be Explained Away? Methodological Controversies in the Study of the Relationship Between Intentionality and Morality. *Personality and Social Psychology Bulletin*, *42*(10), 1295–1308.

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., Beebe, J., Berniūnas, R., Boudesseul, J., Colombo, M., Cushman, F., Diaz, R., N'Djaye Nikolai van Dongen, N., Dranseika, V., Earp, B. D., Torres, A. G., Hannikainen, I., Hernández-Conde, J. V., Hu, W., … Zhou, X. (2021). Estimating the Reproducibility of Experimental Philosophy. *Review of Philosophy and Psychology*, *12*(1), 9–44.

Feltz, A. (2007). The Knobe effect: A brief overview. *Journal of Mind and Behavior*, *28*(3-4), 265–278.

Halpern, J. Y., & Hitchcock, C. (2015). Graded Causation and Defaults. *The British Journal for the Philosophy of Science*, *66*(2), 413–457.

Hitchcock, C., & Knobe, J. (2009). Cause and Norm. *Journal of Philosophy*, *106*(11), 587–612.

Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, *161*, 80–93.

Kirfel, L., & Phillips, J. (2023). The pervasive impact of ignorance. *Cognition*, *231*, 105316.

Knobe, J. (2003a). Intentional action and side effects in ordinary language. *Analysis*, *63*(3), 190–194.

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

Knobe, J. (2006). The Concept of Intentional Action: A Case Study in the Uses of Folk Psychology. *Philosophical Studies*, *130*(2), 203–231.

Knobe, J. (2010). Person as scientist, person as moralist (2010/10/22). *Behavioral and Brain Sciences*, *33*(4), 315–329.

Kominsky, J. F., & Phillips, J. (2019). Immoral Professors and Malfunctioning Tools: Counterfactual Relevance Accounts Explain the Effect of Norm Violations on Causal Selection. *Cognitive Science*, *43*(11), e12792.

Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D., & Knobe, J. (2015). Causal superseding. *Cognition*, *137*, 196–209.

Phillips, J., Luguri, J. B., & Knobe, J. (2015). Unifying morality's influence on non-moral judgments: The relevance of alternative possibilities. *Cognition*, *145*, 30–42.

Uttich, K., & Lombrozo, T. (2010). Norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition*, *116*(1), 87–100.

Wu, S. A., & Gerstenberg, T. (2024). If not me, then who? Responsibility and replacement. *Cognition*, *242*, 105646.