Philosophical Psychology

ISSN: 0951-5089 (Print) 1465-394X (Online) Journal homepage: www.tandfonline.com/journals/cphp20

# Quantifying the Knobe effect: population-level patterns and heterogeneity analyses

Christoph Schmidt-Petri, Daniel Labarca-Pinto & Carsten Schröder

**To cite this article:** Christoph Schmidt-Petri, Daniel Labarca-Pinto & Carsten Schröder (03 Jun 2026): Quantifying the Knobe effect: population-level patterns and heterogeneity analyses, Philosophical Psychology, DOI: 10.1080/09515089.2026.2676274

**To link to this article:** https://doi.org/10.1080/09515089.2026.2676274

PHILOSOPHICAL PSYCHOLOGY
https://doi.org/10.1080/09515089.2026.2676274



# Quantifying the Knobe effect: population-level patterns and heterogeneity analyses

Christoph Schmidt-Petri [a], Daniel Labarca-Pinto [b] and Carsten Schröder [c]

[a]Department of Philosophy, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany; [b]German Institute for Economic Research, Berlin, Germany; [c]Department of Economics, Free University of Berlin & German Institute for Economic Research, Berlin, Germany

**ABSTRACT**

An influential survey by Joshua Knobe (2003) indicates that the cognitive process ascribing intentionality to an agent is mediated by the perceived moral worth of the action. To assess the external validity of the finding and to obtain more precise estimates of effect sizes, the present study replicates this canonical experiment with a large scale random sample of the German population ($N = 2{,}295$). The difference we observe between the harm and help scenarios is more pronounced than in nearly all previous studies (90.5% [88.754, 92.148] attribute intentionality in the harm scenario, 92.5% [90.947, 94.005] do not do so in the help scenario). In exploratory multivariate analyses, we also investigate whether the magnitude of the Knobe effect differs between different population groups, as characterized by a broad set of socio-demographics and personality traits as given by the Big Five. We find that the magnitude of the Knobe effect varies with gender, migration status, and being in a managerial position.

**ARTICLE HISTORY**
Received 23 July 2025
Accepted 30 March 2026

**KEYWORDS**
Knobe effect; side-effect effect; action; intentionality; Intentional action; replication

## Introduction

Joshua Knobe presented participants of a survey with two vignettes involving a company's chairman (see Knobe, 2003). In the first ("harm") scenario, the chairman approves a plan that boosts profits but also harms the environment, asserting that he only cares about the profit and not about the environment. After the plan is implemented, when asked whether the chairman had *intentionally* harmed the environment, 82% of participants said he had. In the analogous second ("help") scenario, the chairman also approves a plan that boosts profits but which instead benefits the environment, again asserting that he only cares about the profit and not about the environment. This time, when asked whether the chairman had intentionally helped the

**CONTACT** Christoph Schmidt-Petri christoph.schmidt-petri@kit.edu Department of Philosophy, Karlsruhe Institute of Technology (KIT), Kaiserstrasse 12, Karlsruhe 76131, Germany
Supplemental data for this article can be accessed online at https://doi.org/10.1080/09515089.2026.2676274.

environment, 77% of the participants said that he had *not*. This large discrepancy is surprising, as the chairman's mental state with respect to the environment, whatever exactly it is, seems identical in both scenarios. The only difference between the two scenarios seems to be whether the effect on the environment is beneficial or not.

This observation suggests that the cognitive process used to ascribe intentionality to the chairman is influenced by the perceived moral worth of his action: apparently, the moral assessment of an action's effect makes a difference to whether people view the action as having been performed intentionally or not. More precisely, it looks as if the moral assessment of the *side-effects* of the respective actions influences whether respondents believe this effect to have been brought about intentionally or not. The respondents' replies to a further question concerning the blameworthiness or praiseworthiness of the chairman's action coheres with this reading: it turns out that the participants were much more willing to blame the chairman for his action in the harm scenario than to praise him in the help scenario.

Knobe's initial experiment involved a rather small convenience sample of 78 respondents. In subsequent studies, Knobe and many other researchers found similar results with larger samples, from a number of countries, some of which also focused on specific population groups such as children, people with certain psychological conditions etc. (e.g., Cardinale et al., 2014; Cova & Naar, 2012; Dalbauer & Hergovich, 2013; Knobe & Burra, 2006; Leslie et al., 2006; Zalla & Leboyer, 2011). The existence of the effect was also confirmed using different contexts and variations of the original scenarios. These observations have generated extensive discussions in philosophy, psychology and related disciplines dealing with the normative assessment of human behavior, ranging from law (Kneer & Bourgeois-Gironde, 2017), political economy and behavioral economics (Bhatia et al., 2024; Utikal & Fischbacher, 2014) to social ontology (Michael & Szigeti, 2019). The results are generally considered to be quite robust (Cova et al., 2021).

A number of theories have been proposed to explain how this puzzling phenomenon comes about (see e.g., Adams & Steadman, 2004a, 2004b; Cushman & Mele, 2008; Nadelhoffer, 2006; Nichols & Ulatowski, 2007; Pettit & Knobe, 2009). These theories start from the empirical observations, but themselves fall into moral psychology or moral philosophy, theorizing about and assessing how individuals reason about moral phenomena. These discussions are of independent interest for the respective disciplines and many arguments in them may be pursued without direct reference to details of specific empirical observations. For some of these authors, the Knobe effect shows that the concept of intentionality is used normatively (most prominently Knobe himself (Knobe, 2010; Knobe & Burra, 2006)), for others, it shows that respondents have further beliefs about the chairman's mental and moral attitudes that could explain the results without this

assumption (see e.g., Adams & Steadman, 2004b). It is also controversial whether the respondents would be making some sort of mistake with the ascription of intentionality (Alicke, 2008; Nadelhoffer, 2004) or not (McCann, 2005).

While the debate about what causes the Knobe effect is far from settled, there is broad agreement that the effect itself is strong, as robustly observed across many studies. Yet, in all studies a non-negligible proportion of respondents do not follow the canonical pattern of ascribing intentionality to the chairman in the harm scenario while refraining from doing so in the help scenario. Surprisingly, relatively few attempts have been made to (statistically) explain this diversity in responses (with the notable exception of Nichols and Ulatowski (2007) and Feltz and Cokely (2024)) which is just as much in need of explanation as the effect itself. The current study tries to contribute to this explanation.

From an empiricist perspective, one of the key strengths of the current study is its use of a large-scale, population-wide random sample ($N = 2{,}295$), which allows for more robust generalizations compared to prior studies that relied on small-scale convenience samples (Woolfolk, 2013). Randomized sampling eliminates the potential selection biases associated with nonrandom sampling designs and ensures that the results can be generalized to the overall population (here: private households in Germany). Instead, previous research used samples from specific and limited subgroups, such as "people spending time in a Manhattan public park" (Knobe, 2003, p. 191) or "Hindi-speaking students in South Asian clubs at Princeton University and Yale University" (Knobe & Burra, 2006, p. 126). While these studies were valuable in their own right, their results cannot be easily generalized to the broader U.S. population or Hindi-speaking people around the world. By drawing from a diverse and representative population, we can confidently extend the findings to the general population, enhancing the external validity of the conclusions. In addition, our large sample size provides greater statistical power, providing more precise estimates[1] and richer insights into heterogeneity across different subgroups.

Furthermore, the design of the present study addresses important issues regarding internal validity. Earlier studies often employed between-subjects designs, assigning an interviewee to one of the two scenarios and then comparing responses of interviewees in the harm and help scenario. While these designs are valuable, they typically did not include a thorough assessment of randomization effectiveness. This left open the possibility that observed differences might stem from unintended systematic differences between subpopulations, rather than purely from the experimental conditions. By ensuring rigorous randomization and carefully considering sample composition, we improve the internal validity of the present study, making

it possible to more confidently attribute observed effects to the scenarios themselves, rather than to external confounds.[2]

The present study also addresses a broader and unresolved question in the literature: whether the Knobe effect is universal across different demographic and psychological factors. While many previous studies did not examine the role of age, gender, socio-economic status, or personality traits, this study takes a step forward by investigating these factors. For instance, research has shown that gender (Buckwalter & Stich, 2014) and socio-economic status (Weinberg et al., 2001) can influence intuitive judgments, and our study aims to explore whether these variables affect the Knobe effect. By incorporating these key covariates, we gain a more nuanced understanding of who exhibits the Knobe effect and under what conditions. Even though the existence of the effect has been established, it remains valuable to explore its variability across different groups.

In summary, this study makes several important contributions to the literature. It addresses methodological limitations from prior research by using a large-scale, population-wide random sample, improving both internal and external validity. Furthermore, it explores the role of socio-demographic and personality factors in shaping the Knobe effect, providing new insights into its universality. By embedding the study in the high-quality Innovation Sample of the German Socio-Economic Panel (SOEP-IS), which offers detailed background information on a wide range of variables (see Fischbacher et al., 2024; Goebel et al., 2019), this study opens up new avenues for investigating the heterogeneity of the effect across different populations. These exploratory analyses enrich our understanding of how the Knobe effect manifests in diverse contexts and provide a foundation for future research.

As far as the first objective is concerned, the replication, the main result of this paper should be reassuring to the discipline of experimental philosophy. The Knobe effect is *even stronger* than hitherto recognized, and it is also nearly universal (at least in Germany), but different population groups exhibit it to different degrees. As far as the second objective is concerned, the heterogeneity analyses, we find that while the intensity of the effect does not vary with personality traits, some socio-demographic characteristics do matter, with the effect varying systematically with gender, migration background, and being in a managerial position.

Our research design also contributes indirectly to the theorizing about *why* individuals reason about intentionality and side effects as they do. Even though none of the available theories is ruled out as an explanation of why *some* people display the side-effect effect, any *complete* explanation of the population-wide effect would in addition have to be able to show how it coheres with the findings of our heterogeneity analyses. In other words, our

large random sample does not help in finding the one true explanation of what is going on in people's minds, but it shows that one true explanation almost certainly does not exist. Some people do not display the Knobe effect, and its strength varies systematically with certain socio-economic variables. Further research should now try to explore in more detail how and why these variables matter. Both empirical and non-empirical approaches might be of help in this.

## Methods

### *Data*

Data collection was performed in the 2023 wave of SOEP-IS, the Innovation Sample of the German Socio-Economic Panel. Since its initiation in 1984, the SOEP has become one of the largest and longest-running panel surveys globally, comprising up to 32,050 participants each year. As a key component of Germany's research infrastructure, the data are freely accessible to the international scientific community, supported by extensive knowledge-sharing initiatives, user support, and detailed online documentation.

SOEP-IS is SOEP's "experimental laboratory," running since 2012, with a sample of around 4,500 respondents randomly drawn from the German population and representative of it. As part of SOEP, it allows external users, after a scientific review, to introduce novel contents, ranging from survey modules to more intricate experimental designs, free of charge. The design and research plan for our module can be found in the SOEP-IS online companion, including the original questions in German (see https://companion-is.soep.de/)

Just as in Knobe's original setup, the SOEP-IS Knobe module comprises two questions, framed in a harm scenario and a help scenario. The assignment of respondents to scenarios is random. The exact formulation, retranslated from the original German, is detailed below.[3]

**Harm scenario**. An executive is making a suggestion to the chairman of a company. A new business idea would increase the company's profits but harm the environment. The chairman replies: "I don't care about harming the environment, I just want to make as much profit as possible." Subsequently, the idea is implemented and the environment is harmed.

**Question 1**: How much blame does the chairman deserve for his actions?
Response scale: 1 (no blame) to 6 (most severe blame) in steps of 1
**Question 2**: Did the chairman intentionally harm the environment?
Response scale: yes; no

The help scenario is analogous.

**Help scenario.** An executive is making a suggestion to the chairman of a company. A new business idea would increase the company's profits and help the environment. The chairman replies: "I don't care about helping the environment, I just want to make as much profit as possible." Subsequently, the idea is implemented and the environment is helped.

**Question 1**: How much praise does the chairman deserve for his actions?
Response scale: 1 (no praise) to 6 (highest praise) in steps of 1
**Question 2**: Did the chairman intentionally help the environment?
Response scale: yes; no

### *Sample composition and effectiveness of random assignment*

In order to reduce the cognitive burden of the SOEP-IS respondents, each respondent only received a subset of all modules. The Knobe module was provided to 2,420 respondents. Of these, 2,347 respondents answered the module questions. Because for another 52 respondents standard SOEP-IS information on socio-demographic characteristics and personality traits, which we want to use in the heterogeneity analyses, was missing, our estimation sample comprised 2,295 individuals, of which 1,152 were in the harm scenario and 1,143 in the help scenario.

To identify the treatment effect of the two scenarios, the two groups should differ only by chance with respect to both observed and unobserved characteristics. To ensure this, respondents must be randomly assigned to the scenarios. Otherwise, selection could bias the results, as explained above. Table 1 shows that our randomization was effective. The table provides, by scenario, from left to right, the number of valid observations and sample means. Furthermore, the far-right column provides the *p*-value of a student t-test for a difference in means tests. We observe that, first, the number of respondents in both scenarios is almost identical (bottom row). Second, of all observed characteristics, the scenario-specific means are very similar. With one exception (having no children in the household), *p*-values indicate that the scenario means are not significantly different from each other. Any observable difference between the harm and the help scenario hence cannot be due to observable differences of the respondents in the two scenarios.[4]

### *Analytic approach and estimation*

To address the two objectives of this study, the empirical strategy proceeds in two stages.

**Table 1.** Sample composition.

| | Harm | | Help | | |
|---|---|---|---|---|---|
| | Valid | Mean | Valid | Mean | *p*-value |
| Age (in years) | 1152 | 52.67 | 1143 | 52.82 | 0.840 |
| Female (%) | 1152 | 50.61 | 1143 | 50.74 | 0.948 |
| Married (%) | 1152 | 57.47 | 1143 | 55.47 | 0.335 |
| East Germany (%) | 1152 | 19.01 | 1143 | 19.51 | 0.762 |
| No children in the HH (%) | 1152 | 75.87 | 1143 | 79.35 | 0.045 |
| Children < 9 years (%) | 1152 | 11.72 | 1143 | 10.24 | 0.256 |
| Children 9–16 years (%) | 1152 | 12.41 | 1143 | 10.41 | 0.132 |
| Native (%) | 1152 | 82.20 | 1143 | 81.98 | 0.887 |
| First-generation immigrant (%) | 1152 | 9.03 | 1143 | 9.36 | 0.782 |
| Second-generation immigrant (%) | 1152 | 8.77 | 1143 | 8.66 | 0.928 |
| Lower than tertiary education[a] (%) | 1108 | 61.55 | 1102 | 62.20 | 0.745 |
| Tertiary education[a] (%) | 1108 | 34.64 | 1102 | 34.21 | 0.830 |
| Log net household income (euros) | 1092 | 8.12 | 1071 | 8.14 | 0.500 |
| Non-managerial position[a] (%) | 1139 | 80.64 | 1129 | 80.05 | 0.722 |
| Managerial position[a] (%) | 1139 | 18.23 | 1129 | 18.72 | 0.761 |
| Not affiliated with religious community[a] (%) | 928 | 34.81 | 934 | 34.38 | 0.830 |
| Affiliated with religious community[a] (%) | 928 | 45.75 | 934 | 47.33 | 0.447 |
| Extraversion | 1146 | 4.66 | 1137 | 4.76 | 0.052 |
| Conscientiousness | 1149 | 5.65 | 1136 | 5.69 | 0.279 |
| Openness to experience | 1141 | 4.81 | 1128 | 4.83 | 0.640 |
| Neuroticism | 1148 | 3.82 | 1136 | 3.72 | 0.093 |
| Agreeableness | 1148 | 5.38 | 1137 | 5.40 | 0.668 |
| Observations | 1152 | | 1143 | | |

Note: The first and third columns show the number of valid observations for selected variables in the harm and help scenarios, respectively. The second and fourth columns display the means for each scenario. The fifth column reports the *p*-values for the differences in means between the two scenarios. *Source*: SOEP-IS 2023.

[a]Indicates that these variables include an additional category for missing values. Missing values in the variables log gross household income and all Big Five personality traits were replaced with the respective variable's sample mean. Additionally, for these variables a dichotomous variable was created, set to 1 for missing observations.

### *Replication without controls*

The first stage involves a replication exercise aimed at testing the generalizability and external validity of prior experimental findings. In line with previous research, we compare responses between the two scenarios using standard statistical procedures: a test of proportions for the attribution of intentionality and a t-test for the comparison of blame/praise ratings. We examine whether, among all respondents, (1) a greater proportion ascribes intentionality in the harm scenario than in the help scenario; (2) whether the blame ratings are higher in the harm scenario than the praise ratings in the help scenario; and (3) how the fact of ascribing intentionality to the action matters for the respective level of blame and praise. An additional aim is to quantitatively assess whether our findings align with the effect sizes observed in earlier studies. To this end, we provide a descriptive comparison of our results with those from previous research, focusing on the shares ascribing intentional behavior (see Section Discussion).[5]

### *Heterogeneities in response patterns*

The second stage, which represents an entirely novel contribution to the literature, explores heterogeneities in response patterns using multivariate

regression models that statistically explain (1) the attribution of intentionality, and (2) levels of blame and praise. The results show whether the Knobe effect manifests similarly across different population groups: Are male and female respondents, respondents from different age groups, educational backgrounds, or occupational roles (e.g., managers versus non-managers) equally likely to attribute intentionality to the chairman, and do their responses regarding the level of blame and praise differ? Following an observation by Knobe (2003), we also look at whether levels of blame/praise vary with people's judgments about whether or not the side effect was brought about intentionally or not.

Because this is an exploratory analysis without strong prior hypotheses about the role of respondents' socio-demographics and personality, we did not apply multiple-testing corrections. Imposing such adjustments would reduce power and would risk obscuring potentially meaningful relationships. The reported associations should therefore be interpreted as hypothesis-generating rather than confirmatory, and require validation in follow-up large-scale studies from Germany and other countries.

For explaining the attribution of intentionality, a binary variable, we rely on a logit estimation. Our logit model of probability $P(I_i = 1|\mathbf{Z_i})$ (i.e., that respondents attribute intention) takes the following form:

$$P(I_i = 1|\mathbf{Z_i}) = \Lambda\left(\alpha^{\mathrm{I}} + \gamma^{\mathrm{I}} \times \mathrm{H_i} + \beta^{\mathbf{I}} \times \mathbf{X_i} + \delta^{\mathbf{I}} \times (\mathrm{H_i} \times \mathbf{X_i})\right), \qquad (1)$$

with $I_i$ denoting an indicator (equal to one if respondents attribute intention), and $\mathbf{Z_i}$ the full set of explanatory variables ($\mathbf{X}_i$, $H_i$ and the interaction between the two). Because logit coefficients are expressed in log-odds, their magnitude is not directly interpretable in terms of the probability of the outcome (intentionality); only their sign reveals the direction of the relationship. Therefore, we estimate average marginal effects (AMEs) to quantify the change in predicted probabilities. They are derived from the logistic cumulative distribution function, $\Lambda(\mathrm{k}) = \frac{1}{1+\mathrm{e}^{-\mathrm{k}}}$, and represent the average change in the predicted probability of the outcome for a change in a predictor, averaged across all observations in the dataset. For a continuous predictor like age, the AME refers to a marginal change of the predictor – ceteris paribus – and can be interpreted as an elasticity. For a discrete predictor like gender, the AME refers to a zero-to-one change (e.g., a comparison of male and female respondents).

The model includes three sets of main explanatory variables. The first set, $H_i$, is an indicator that equals 1 for respondents in the harm scenario and 0 for those in the help scenario. The second set, the vector $\mathbf{X}_i$ represents socio-demographic characteristics and personality traits of the respondents. This vector includes age, gender, marital status, region of residence, family composition, migration background, education, household income,

managerial position, religious community affiliation, and all Big Five traits.[6] To ease the interpretation of the estimation results, the variables age, log net household income, and all Big Five traits are centered by subtracting their respective sample means from each individual observation.[7] Thus, for each control variable in $X_i$, the AME indicates how response behavior changes relative to the reference characteristic (the sample mean for centered continuous variables and the base category for categorical variables).[8] The third set comprises the interaction of the vector $\mathbf{X}_i$ with the indicator $H_i$ to explore whether the role of sociodemographic characteristics and personality for intentionality attribution differs between the harm and help scenario.

We estimate scenario-specific AMEs for all covariates in $\mathbf{X}_i$ by evaluating them separately at $H_i = 1$ (harm) and $H_i = 0$ (help) and test differences between the AMEs using Wald tests. This approach allows us to quantify scenario-specific magnitudes and to test whether the influence of sociodemographics and personality traits on the probability of attributing intentionality differs by scenario. For example, if the difference between the AMEs in the harm and help scenarios is positive for a given sociodemographic characteristic, that characteristic is associated with a larger increase (or smaller decrease) in the probability of attributing intentionality in the harm scenario relative to the help scenario. To provide a baseline, we also estimate the predicted probability for a reference individual (labeled as the "constant") by setting all explanatory variables to zero. We apply an analogous Wald test to these baseline predicted probabilities to assess whether the reference group's propensity to attribute intentionality differs significantly between the two scenarios. For instance, a positive and statistically significant difference in these baseline probabilities would support the existence of the Knobe effect.

In a nutshell, the intentionality regression allows us to answer the following questions: (1) To what extent do attributions of intentionality differ between the harm and help scenarios? (2) How does the attribution of intentionality correlate with respondents' socio-demographic characteristics? (3) Is this correlation the same in the harm and help scenario?

We specify the following ordinary least squares (OLS) regression model for the level ($L$) of blame/praise (measured on a one-to-six scale):

$$L_i = \alpha^L + \gamma^L \times H_i + \beta^L \times \mathbf{X}_i + \delta^L \times (H_i \times \mathbf{X}_i) + \varepsilon_i^L, \tag{2}$$

where the set of explanatory variables is the same as in Equation 1. In contrast to the aforementioned logit model, where the coefficients represent the change in the log-odds of the outcome and require computation of marginal effects to interpret their direct impact on probabilities, OLS coefficients can be directly interpreted as the change in the outcome for

a one-unit change in the predictor. This is because OLS assumes a linear relationship between predictors and the outcome.

The blame/praise regression allows us to address three questions analogous to the preceding three: (1) How does the average level of blame differ from the average level of praise? (2) How do the respective levels of blame/praise correlate with respondents' socio-demographic characteristics? (3) Is this correlation the same in the harm and help scenario?

To further examine whether "the total amount of praise or blame that subjects offered was correlated with their judgments about whether or not the side effect was brought about intentionally" (Knobe, 2003, p. 193), we also estimate a more flexible specification of the blame/praise regression (2). This generalized model contains all explanatory variables in isolation and all possible interactions, and thus provides the most flexible framework to explain reported levels of blame and praise. In particular, it allows for the attribution of intentionality to not have the same implications for the respective levels of blame and praise reported by the respondents. For example, in the harm scenario, first-generation immigrants could assign less blame when they perceive the action as intentional than when they do not. We also examine whether the role of respondents' characteristics varies depending on the scenario presented. For instance, respondents in managerial roles could blame *less* in the harm scenario and praise *more* in the help scenario. To capture these joint effects, we introduce interaction terms among all three sets of variables. The notation in parentheses in the generalized model (see equation below) indicates these interactions, where $\eta^{L,g}$ represents the combined effect of the scenario and the attribution of intentionality on the levels of blame and praise; $\lambda^{L,g}$ captures the interaction between intentionality and respondent characteristics; $\delta^{L,g}$ reflects the combined effect of scenario and respondent characteristics; and $\theta^{L,g}$ denotes the three-way interaction among intentionality, scenario, and respondent characteristics.

The generalized ($g$) blame/praise OLS regression takes the form:

$$\begin{aligned} L_i = {} & \alpha^{L,g} \\ & + \gamma^{L,g} \times H_i + \psi^{L,g} \times I_i + \beta^{L,g} \times \mathbf{X}_i \\ & + \eta^{L,g} \times (I_i \times H_i) + \lambda^{L,g} \times (I_i \times \mathbf{X}_i) + \delta^{L,g} \times (H_i \times \mathbf{X}_i) \\ & + \theta^{L,g} \times (I_i \times H_i \times \mathbf{X}_i) + \varepsilon_i^{L,g}. \end{aligned} \tag{3}$$

The model includes the three sets of main explanatory variables from (2), an indicator of intentionality, and four sets of interaction terms. The coefficient $\psi^{L,g}$ indicates if blame/praise attributions systematically differ

with people's judgments about whether or not the side effect was brought about intentionally. The coefficient $\eta^{L,g}$ indicates if the difference between the blame and the praise attributions correlates with intentionality attribution. The vector $\lambda^{L,g}$ indicates if the role of covariates for blame/praise attributions correlates with respondents' judgments of intentionality. Finally, the vector $\theta^{L,g}$ reveals potential triple-interactions of intentionality attribution, socio-demographics and personality, and the harm/help treatment.

The subsequent regression tables report AMEs of each explanatory variable. For Equations (1) and (2), we compute the AMEs separately for the harm and help scenarios, respectively. For Equation (3), within each scenario, we estimate the AMEs for respondents who indicated that the chairman intentionally harmed or helped the environment and for those who indicated that he did not.[9] We test differences between AMEs using Wald tests, with robust standard errors being computed using the delta method.

Some control variables in $\mathbf{X}_i$ may be correlated (e.g., income and education – a higher level of education could be correlated with a higher level of income), which creates a problem known as multicollinearity. Multicollinearity can lead to inflated standard errors: the estimations might appear not to be statistically significant even if they actually are. Consequently, one may mistakenly conclude that certain variables do not affect the dependent variable when they actually do.

As a remedy, we additionally estimate stepwise regressions. Stepwise regressions identify the most important predictors (independent variables) by automatically selecting or removing variables in a data-driven manner, balancing simplicity with predictive power. Specifically, stepwise regressions select a subset of the available predictor variables that contribute the most to explaining the variance in the dependent variable. The stepwise selection is performed using backward elimination. At each step, the term with the highest $p$-value is removed from the model if its $p$-value exceeds the threshold of $p = 0.01$. This process continues iteratively until only terms statistically significant at the 1-percent level remain in the model.[10] We include a covariate along with all its corresponding interaction terms if the covariate and its interactions are jointly significant. For example, assume the population consists of natives, first-generation immigrants, and second-generation immigrants. Using the natives as the reference group, two dummy variables describe the two groups of immigrants. In addition, in the intentionality regression, these dummies are also interacted with a scenario dummy. If this set of covariates is jointly significant, all the covariates are included in the regression even if single covariates are insignificant.

## Results

### *Replication without control variables*

This section replicates the existing evidence on the Knobe effect. As mentioned above, existing research suggests that: (1) a greater proportion of respondents ascribe intentionality in the harm scenario than in the help scenario; (2) the blame ratings are higher in the harm scenario than the praise ratings in the help scenario; and (3) ascribing intentionality to the action matters for the respective levels of blame and praise.

Turning to the first point, Table 2 presents the point estimates for the population shares attributing intentionality in the two scenarios, together with the respective 95% confidence intervals. In the harm scenario, the vast majority of respondents – 90.5% – attribute intentionality to the chairman's actions, whereas in the help scenario, only 7.5% do so. A two-sample test of proportions confirms that the proportion of respondents attributing intentionality is significantly higher in the harm scenario (with a z-score of 39.733 and a $p$-value below 0.0001). These findings qualitatively align with the results of previous studies, including Knobe (2003, 2004) and subsequent research which consistently shows the attribution of intentionality to be much higher in the harm scenario. By contrast, the difference between the two scenarios is much larger than in previous studies, more precisely estimated because of the large sample size, and highly significant. The confidence interval for the difference is tightly estimated at $[80.643, 85.212]$. As noted, these findings generalize to the full population and are not confounded by unwanted selection into the two treatments (see Table 1).

Let's turn to the next two points: (2) How does the evaluation of the chairman's action between the harm and the help scenario differ, and (3) what role does the assumption that he acted intentionally play? Figure 1 illustrates the results in histogram form, while Table 3 provides the corresponding summary statistics and statistical test results.

The top two panels of Figure 1 provide the distributions of reported levels of blame for the harm and praise for the help scenario. Consistently with previous evidence, with an average of 5.180 (95% confidence interval

**Table 2.** Evaluation of intentionality.

| Intentionality | Harm | Help | Δ frac. | z score |
|---|---|---|---|---|
| Yes (%) | 90.451 | 7.524 | 82.927 | 39.733*** |
| | [88.754, 92.148] | [5.995, 9.053] | [80.643, 85.212] | |
| No (%) | 9.549 | 92.476 | | |
| | [7.852, 11.246] | [90.947, 94.005] | | |

Note: This table presents results from two-sample tests of proportions. Δ indicates the difference of intentionality shares between the harm and help scenario. Number of observations: Harm: 1,152; Help: 1,143. *Source*: SOEP-IS 2023.

*$p$ <.1,**$p$ <.05,***$p$ <.01.

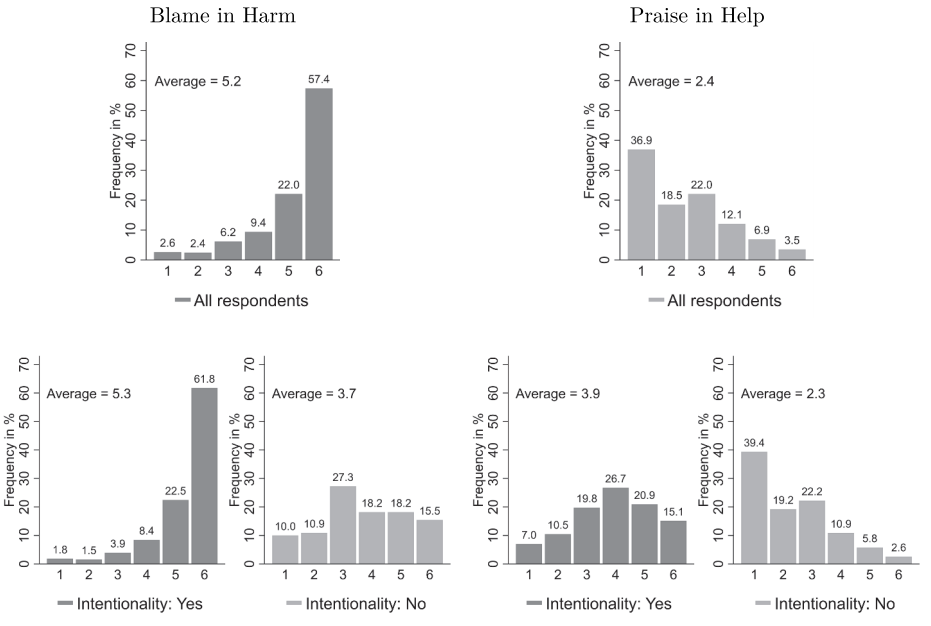


**Figure 1.** Distribution of blame and praise. Note: This figure presents the distribution of responses to the question about how much blame/praise the chairman deserves for his action. The top row shows the levels of blame and praise that respondents attribute to the chairman in the harm and help scenarios, respectively. The bottom row disaggregates responses within each scenario by whether individuals indicated that the chairman intentionally harmed/helped the environment or not. Number of observations: Harm – Intentionality (Yes: 1,042; No: 110); Help – Intentionality (Yes: 86; No: 1,057). *Source*: SOEP-IS 2023.

**Table 3.** Evaluations of blame and praise.

| | Harm | Help | Δ of means | t stat. |
|---|---|---|---|---|
| All respondents | 5.180 | 2.440 | 2.740 | 49.054*** |
| | [5.108, 5.251] | [2.357, 2.523] | [2.630, 2.849] | |
| Intentionality: Yes | 5.336 | 3.895 | 1.441 | 11.505*** |
| | [5.270, 5.402] | [3.587, 4.204] | [1.195, 1.686] | |
| Intentionality: No | 3.700 | 2.322 | 1.378 | 9.927*** |
| | [3.411, 3.989] | [2.239, 2.404] | [1.106, 1.651] | |
| Δ of means | 1.636 | 1.574 | | |
| | [1.413, 1.859] | [1.271, 1.876] | | |
| t stat. | 14.376*** | 10.203*** | | |

Note: This table presents results from two-tailed independent samples. Δ indicates the difference of means. 95 % confidence bands in parentheses. Number of observations: Harm - Intentionality (Yes: 1,042; No: 110); Help - Intentionality (Yes: 86; No: 1,057). *$p < .1$,**$p < .05$,***$p < .01$. Source: SOEP-IS 2023.

$[5.108, 5.251]$) the level of blame is much larger than the level of praise with an average of 2.440 $[2.357, 2.523]$. The difference of the two valuations is highly significant (see Table 3).

The bottom panels differentiate whether respondents attributed intentionality to the chairman's behavior or not, distinguished by scenario. Among those who did attribute intentionality in the harm scenario, shown on the left, nearly 62% assigned the highest level of blame (level 6

on the 6-point scale), resulting in an average blame score of 5.336 [5.270, 5.402]. By contrast, respondents who did not attribute intentionality reported a significantly lower average blame score of 3.700 [3.411, 3.989]. The difference of 1.636 is also statistically significant (see Table 3). The right panels show the respective responses for the help scenario. The average level of praise among those who did attribute intentionality is 3.895 [3.587, 4.204], while it is 2.322 [2.239, 2.404] among those who did not. The difference of 1.574, too, is statistically significant. These results align with the patterns previously observed.

### *Heterogeneities in response patterns*

Our study confirms the existence of a strong Knobe effect. At the same time, not *all* respondents perceive the chairman as intentionally harming (or not helping) the environment, nor do they *all* assign blame or praise in the same way. So, what accounts for these differences in response patterns? One way to explain this would be by claiming that people use different theories about the link between intentions (or other mental states) and side effects, or about the difference in moral responsibility for bad as opposed to good side effects. These are certainly relevant considerations, but given our research design, the workings of the individuals' minds remain unobservable to us. It might, however, also be possible to explain, at least in the statistical sense, the variations we have observed by socio-demographic factors and personality traits which *are* observable to us. This is what this section investigates.

#### *Heterogeneities in the attribution of intentionality*

We begin by examining the attribution of intentionality made by the respondents. Table 4 presents the estimation results from the intentionality regression, Equation (1), in terms of average marginal effects. The column "help" provides the results for the help scenario. In this scenario, the indicator $H_i$ and the vector of coefficients of the interaction terms $H_i \times \mathbf{X_i}$ are zero. Therefore, the scenario-specific AMEs are derived from the regression intercept, $\hat{\alpha}^I$, and the estimated coefficient vector for the explanatory variables, $\hat{\beta}^I$. Conversely, in the harm scenario, where $H_i = 1$, the computation of the AMEs also includes the coefficients $\hat{\gamma}^I$ and $\hat{\delta}^I$, in addition to those used in the help scenario.

In both columns, a positive marginal effects means that the variable makes the attribution of intentionality more likely. Hence, the Knobe effect – the asymmetric attribution of intentionality between the harm and help scenarios – would be more pronounced than on average when the marginal effect is positive in the harm column and negative in the help column.

**Table 4.** Probability of attributing intentionality.

| | (1) Harm | (2) Help | (3) Δ |
|---|---|---|---|
| Age | 0.002 | −0.005** | 0.007* |
| | (0.003) | (0.003) | |
| Age squared | −0.000 | 0.000** | −0.000* |
| | (0.000) | (0.000) | |
| Female | 0.025 | −0.022 | 0.047* |
| | (0.018) | (0.016) | |
| Married | 0.011 | 0.006 | 0.005 |
| | (0.023) | (0.018) | |
| East Germany | −0.017 | 0.019 | −0.035 |
| | (0.026) | (0.023) | |
| Children < 9 years | −0.014 | 0.116*** | −0.130** |
| | (0.034) | (0.044) | |
| Children 9–16 years | 0.004 | −0.048*** | 0.052 |
| | (0.028) | (0.017) | |
| First-generation immigrant | −0.096** | 0.107*** | −0.203*** |
| | (0.039) | (0.037) | |
| Second-generation immigrant | 0.013 | −0.017 | 0.030 |
| | (0.028) | (0.023) | |
| Tertiary education | −0.014 | −0.019 | 0.004 |
| | (0.020) | (0.017) | |
| Log net household income | 0.011 | −0.016 | 0.027 |
| | (0.017) | (0.014) | |
| Managerial position | 0.046** | −0.060*** | 0.107*** |
| | (0.020) | (0.014) | |
| Affiliated with religious community | −0.053*** | 0.018 | −0.071*** |
| | (0.021) | (0.017) | |
| Extraversion | −0.010 | 0.002 | −0.012 |
| | (0.008) | (0.006) | |
| Conscientiousness | −0.007 | −0.003 | −0.004 |
| | (0.010) | (0.008) | |
| Openness to experience | 0.006 | −0.008 | 0.013 |
| | (0.009) | (0.008) | |
| Neuroticism | 0.009 | −0.003 | 0.013 |
| | (0.008) | (0.006) | |
| Agreeableness | 0.013 | 0.000 | 0.013 |
| | (0.010) | (0.007) | |
| Constant | 0.926*** | 0.058*** | 0.867*** |
| | (0.021) | (0.018) | |
| Adj. pseudo $R^2$ | 0.548 | | |
| Log pseudo-likelihood | −616.275 | | |
| $\chi^2$ (degrees of freedom) | 879.659 (47) | | |
| Observations | 2295 | | |

Note: This table presents the estimated AMEs from the logit model based on Equation (1). The first and second columns show the probability that individuals consider the chairman to have intentionally harmed/helped the environment in the harm and help scenarios, respectively. Δ indicates the difference between the AMEs in the harm and help scenarios, and stars denote the significance level of a Wald test. AMEs for missing dummies are not shown. Robust standard errors are reported in parentheses. *$p < .1$, **$p < .05$,***$p < .01$. Source: SOEP-IS 2023.

The bottom row "constant" indicates the conditional probability of ascribing intentionality predicted by the model when all independent variables are set to zero. As shown in the third column, respondents with baseline characteristics are 86.7% points more likely to ascribe intentionality in the harm scenario than in the help scenario. According to a Wald test, this difference is highly significant. Interestingly, the conditional predicted probabilities for individuals with baseline characteristics in the harm

scenario (92.6%) and the help scenario (5.8%) closely resemble the unconditional descriptive probabilities reported in Table 2 for the harm (90.451%) and help (7.524%) scenarios. This similarity can be attributed to two factors: the mean-standardization of the explanatory variables and the fact that all average marginal effects of the explanatory variables are quantitatively small.[11]

The explanatory power of the model, as measured by the pseudo adjusted coefficient of determination of 0.548 is rather high for a cross-sectional regression. A large portion of this explanatory power comes from the regression constant and the scenario indicator, $H_i$. Removing all the socio-demographics and personality traits, $\mathbf{X_i}$, and the interaction terms from the regression gives a marginally larger adjusted coefficient of determination of 0.578.[12] Thus, the control vector $\mathbf{X_i}$ does not add much information in explaining the variation in the dependent variable, and its removal has little impact on the model's ability to fit the data. In other words, only a limited number of respondent characteristics appear to explain the respective attributions of intentionality. Whether respondents perceive the effect as intentional or not is almost entirely due to the scenario they find themselves in. Nonetheless, certain respondent characteristics warrant particular attention.

Turning to the socio-demographics, first, while the marginal effects within each scenario are not statistically significant on their own, the probability of *women* attributing intentionality to the chairman's actions differs significantly between the two scenarios. After controlling for other observed characteristics, women are more likely than men to attribute intentionality in the harm scenario and less likely to do so in the help scenario, with the difference between the scenarios amounting to about 4.7% points, being statistically significant at the ten percent level (col. (3)). In other words, women show a slightly *stronger* Knobe effect.[13]

Second, respondents with *young children* are more likely to attribute intentionality in the help scenario, with a difference of about minus 13% points after controlling for other factors, significant at the five percent level. As a result, they show a somewhat *weaker* Knobe effect.

Third, *first-generation immigrants* exhibit a systematically different response pattern compared to native Germans: in the harm scenario, they are less likely to attribute intentionality (by approximately 9.6% points), whereas in the help scenario, they are more likely to do so (by about 10.7% points). This difference of 20.3% points is negative and significant at the one percent level. Thus, while the Knobe effect is still very much present, it is *much weaker* with first-generation immigrants. Interestingly, no statistically significant difference is observed between second-generation immigrants and natives.

Fourth, also respondents in *managerial positions* display a distinct pattern. For this group, the side-effect effect is even *stronger* than in the general

population. Ceteris paribus, holding a managerial position is associated with a 4.6% point increase in the probability of attributing intentionality in the harm scenario, and a 6% point decrease in the help scenario. The difference between the two, 10.7% points, is highly significant.

Finally, affiliation with a *religious community* significantly *reduces* the strength of the Knobe effect. The probability of attributing intentionality in the harm scenario is reduced by 5.3% points compared to those without religious affiliation, controlling for other factors. Although no significant effect is found in the help scenario, the difference between the AMEs of the two scenarios is 7.1% points, which is highly significant.

The personality of the respondents, as captured by the Big Five, cannot statistically explain whether they attribute intentionality to the chairman or not.

To illustrate the heterogeneity of the Knobe effect in absolute terms and to highlight its variation across different subgroups in our estimation, we present a few selected average predicted probabilities for specific stylized groups. The differences between these groups are quite substantial. At one extreme, the group with the weakest Knobe effect are men who are first-generation immigrants, affiliated with a religious community and have children younger than 9 years old (with all other explanatory variables held at their observed values). They show a predicted probability of attributing intentionality of 71.2% in the harm scenario and 42.9% in the help scenario. The other extreme, the group with the strongest Knobe effect, are women native to Germany in managerial positions who are not affiliated with a religious community: they have a probability of ascribing intentionality to the chairman's action of 96.6% in the harm scenario and 1.3% in the help scenario.

Table 5 shows the corresponding results from the stepwise regression model.[14] Measured by the pseudo adjusted $R^2$ of 0.581, this more parsimonious model has slightly higher explanatory power as the comprehensive model just discussed (see footnote 12 for an explanation of why the value can increase with a small number of explanatory variables.). Further, the model confirms, quantitatively and qualitatively, the results regarding the role of children in the help scenario, as well as the importance of a first-generation migration background[15] and managerial position in both scenarios.

### *Heterogeneities in evaluations of blame and praise*

Next, we turn to the results from the blame/praise regressions. The first set of results corresponds to regression specification (2), as reported in Table 6. It shows how sociodemographic characteristics and personality traits are correlated with the reported levels of blame and praise in the respective scenarios. As in the intentionality regression, the first column presents the AMEs in the harm, and the second column presents those in the help scenario.

**Table 5.** Stepwise selected model: probability of attributing intentionality.

| | (1) Harm | (2) Help | (3) Δ |
|---|---|---|---|
| Children < 9 years | −0.021 | 0.096*** | −0.117** |
| | (0.030) | (0.036) | |
| Children 9–16 years | 0.000 | −0.053*** | 0.053* |
| | (0.026) | (0.016) | |
| First-generation immigrant | −0.090** | 0.098*** | −0.188*** |
| | (0.038) | (0.035) | |
| Second-generation immigrant | 0.010 | −0.017 | 0.027 |
| | (0.028) | (0.022) | |
| Managerial position | 0.042** | −0.070*** | 0.112*** |
| | (0.019) | (0.012) | |
| Constant | 0.907*** | 0.073*** | 0.835*** |
| | (0.011) | (0.010) | |
| Adj. pseudo $R^2$ | 0.581 | | |
| Log pseudo-likelihood | −636.501 | | |
| $\chi^2$ (degrees of freedom) | 910.836 (13) | | |
| Observations | 2295 | | |

Note: This table presents the estimated AMEs of a model that includes only the variables selected in the stepwise selection process. The backward stepwise selection was conducted using a critical significance level of $p = 0.01$ for term removal. The first and second columns show the probability that individuals consider the chairman to have intentionally harmed/helped the environment in the harm and help scenarios, respectively. Δ indicates the difference between the AMEs in the harm and help scenarios, and stars denote the significance level of a Wald test. AMEs for missing dummies are not shown. Robust standard errors are reported in parentheses. *$p < .1$,**$p < .05$, ***$p < .01$. *Source*: SOEP-IS 2023.

Starting with the row "constant," for respondents with baseline characteristics, the degree of blame assigned in the harm scenario is 2.88 points higher than the degree of praise assigned in the help scenario, the difference being highly significant.

With regard to the socio-demographics, the level of blame in the harm scenario is significantly higher among older respondents (at a decreasing rate) and females, and lower among first-generation migrants. In the help scenario, we find that the level of praise is significantly higher for married respondents and lower for second-generation migrants.

In terms of personality traits, openness to experience and agreeableness are positively associated with the reported level of blame. Interestingly, more agreeable respondents report *lower* level of praise in the help scenario, and this opposing correlation between the two scenarios is highly significant. In addition, in the help scenario, we find negative association of praise with neuroticism, but it is only weakly significant.

The stepwise model of Equation (2), as reported in Table 7, confirms the relevance of the socio-demographic variables age, gender, marital and migration status, while none of the personality traits is selected.

The second set of results on the determinants of blame and praise refers to the generalized specification (3), using intentionality and all the related interactions added as additional controls. Specifically, this is determined by whether the indicator $I_i$ equals 1 (indicating the attribution of intentionality) or 0 (indicating the opposite), generating four distinct sets of

**Table 6.** Chairman blame and praise evaluation.

| | (1) Blame in harm | (2) Praise in help | (3) Δ |
|---|---|---|---|
| Age | 0.03** | 0.00 | 0.03 |
| | (0.01) | (0.01) | |
| Age squared | −0.00* | −0.00 | −0.00 |
| | (0.00) | (0.00) | |
| Female | 0.27*** | −0.12 | 0.39*** |
| | (0.08) | (0.10) | |
| Married | 0.04 | 0.20* | −0.16 |
| | (0.09) | (0.11) | |
| East Germany | −0.05 | −0.01 | −0.03 |
| | (0.09) | (0.11) | |
| Children < 9 years | −0.13 | 0.22 | −0.35* |
| | (0.13) | (0.16) | |
| Children 9–16 years | −0.11 | −0.03 | −0.08 |
| | (0.13) | (0.14) | |
| First-generation immigrant | −0.55*** | −0.02 | −0.53** |
| | (0.16) | (0.16) | |
| Second-generation immigrant | 0.09 | −0.22* | 0.31* |
| | (0.13) | (0.13) | |
| Tertiary education | 0.06 | 0.08 | −0.03 |
| | (0.08) | (0.10) | |
| Log net household income | 0.10 | −0.01 | 0.11 |
| | (0.08) | (0.09) | |
| Managerial position | 0.04 | −0.03 | 0.08 |
| | (0.10) | (0.13) | |
| Affiliated with religious community | −0.13 | 0.10 | −0.23* |
| | (0.08) | (0.10) | |
| Extraversion | −0.03 | −0.00 | −0.03 |
| | (0.03) | (0.04) | |
| Conscientiousness | −0.01 | 0.01 | −0.02 |
| | (0.04) | (0.05) | |
| Openness to experience | 0.07* | −0.02 | 0.09 |
| | (0.04) | (0.04) | |
| Neuroticism | 0.03 | −0.06* | 0.09* |
| | (0.03) | (0.04) | |
| Agreeableness | 0.06* | −0.12*** | 0.19*** |
| | (0.04) | (0.05) | |
| Constant | 5.18*** | 2.30*** | 2.88*** |
| | (0.10) | (0.11) | |
| Adj. $R^2$ | 0.528 | | |
| Observations | 2295 | | |

Note: This table presents the estimated AMEs based on Equation (2). The first and second columns show the estimation results for the attribution of blame/praise for the chairman's actions in the harm and help scenarios, respectively. Δ indicates the difference between the AMEs in the harm and help scenarios, and stars denote the significance level of a Wald test. AMEs for missing dummies are not shown. Robust standard errors are reported in parentheses. *$p < .1$,**$p < .05$,***$p < .01$. *Source*: SOEP-IS 2023.

estimates. In the help scenario "without intentionality," the results are determined by the regression intercept, $\hat{\alpha}^{L,g}$, and the coefficient vector $\hat{\beta}^{L,g}$. If, in the same scenario, respondents do attribute intentionality, the constant is $\hat{\alpha}^{L,g} + \hat{\psi}^{L,g}$ and the role of the explanatory variables is captured by $\hat{\beta}^{L,g} + \hat{\lambda}^{L,g}$. The results for the harm scenario must consider all the regression coefficients, as $H_i$ equals 1. Table 8 illustrates how the results can be derived for each of the four combinations. It also shows how, within a scenario, the difference in the level of blame/praise is determined depending on whether respondents attribute intentionality to the chairman.

**Table 7.** Stepwise selected model: Chairman blame and praise evaluation.

| | (1) Harm | (2) Help | (3) Δ |
|---|---|---|---|
| Age | 0.01*** | −0.01** | 0.02*** |
| | (0.00) | (0.00) | |
| Female | 0.27*** | −0.19** | 0.46*** |
| | (0.07) | (0.08) | |
| Married | 0.09 | 0.26*** | −0.17 |
| | (0.07) | (0.09) | |
| First-generation immigrant | −0.55*** | −0.00 | −0.54** |
| | (0.16) | (0.16) | |
| Second-generation immigrant | 0.07 | −0.25* | 0.31* |
| | (0.13) | (0.13) | |
| Constant | 5.04*** | 2.41*** | 2.62*** |
| | (0.08) | (0.08) | |
| Adj. $R^2$ | 0.525 | | |
| Observations | 2295 | | |

Note: This table presents the estimated AMEs of a model that includes only the variables selected in the stepwise selection process. The backward stepwise selection was conducted using a critical significance level of $p = 0.01$ for term removal. The first and second columns show the estimation results for the attribution of blame/praise for the chairman's actions in the harm and help scenarios, respectively. Δ indicates the difference between the AMEs in the harm and help scenarios, and stars denote the significance level of a Wald test. AMEs for missing dummies are not shown. Robust standard errors are reported in parentheses. *$p < .1$, **$p < .05$,***$p < .01$. Source: SOEP-IS 2023.

**Table 8.** Definition of displayed estimates from blame/praise regression.

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | Harm | | | Help | | |
| *Intentionality:* | Yes | No | Δ | Yes | No | Δ |
| Constant | $\alpha + \psi + \gamma + \eta$ | $\alpha + \gamma$ | $\psi + \eta$ | $\alpha + \psi$ | $\alpha$ | $\psi$ |
| Explanatory variables | $\beta + \lambda + \delta + \theta$ | $\beta + \delta$ | $\lambda + \theta$ | $\beta + \lambda$ | $\beta$ | $\lambda$ |

Table 9 presents the regression results. It is a more detailed version of Table 6. The first two columns show the results for the harm scenario, now differentiating between respondents that attribute intentionality and those who do not, while the third column reports the AME differences along with the results of a Wald test. This test indicates whether the statistical relationship between a particular control variable and the reported level of blame/praise varies between respondents who attribute intentionality and those who do not. The next three columns provide the analogous information for the help scenario.

The bottom row ("constant") displays the average level of blame and praise when all explanatory variables are equal to zero: As shown in the descriptive replication and consistent with previous evidence, the average level of blame is higher than the average level of praise, and both levels are higher when respondents attribute intentionality to the chairman. The difference between respondents who attribute intentionality and those who do not is 1.4 in the harm scenario (column 3, row "constant") and

**Table 9.** Chairman blame and praise evaluation by attribution of intentionality.

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | Blame in harm | | | Praise in help | | |
| *Intentionality:* | Yes | No | Δ | Yes | No | Δ |
| Age | 0.04*** | −0.01 | 0.05 | −0.01 | 0.01 | −0.02 |
| | (0.01) | (0.05) | | (0.05) | (0.01) | |
| Age squared | −0.00** | 0.00 | −0.00 | 0.00 | −0.00 | 0.00 |
| | (0.00) | (0.00) | | (0.00) | (0.00) | |
| Female | 0.18** | 0.73** | −0.55* | 0.10 | −0.10 | 0.20 |
| | (0.07) | (0.31) | | (0.33) | (0.10) | |
| Married | 0.13 | −0.56* | 0.69** | 0.68* | 0.17* | 0.51 |
| | (0.08) | (0.31) | | (0.40) | (0.10) | |
| East Germany | −0.07 | 0.22 | −0.29 | 0.35 | −0.06 | 0.41 |
| | (0.09) | (0.37) | | (0.27) | (0.11) | |
| Children < 9 years | −0.14 | −0.27 | 0.13 | 0.36 | 0.12 | 0.24 |
| | (0.12) | (0.47) | | (0.39) | (0.17) | |
| Children 9–16 years | −0.02 | −1.05** | 1.03** | −0.74 | 0.03 | −0.77 |
| | (0.12) | (0.46) | | (0.69) | (0.14) | |
| First-generation immigrant | −0.37** | −0.19 | −0.18 | −0.72 | −0.12 | −0.60 |
| | (0.16) | (0.33) | | (0.49) | (0.16) | |
| Second-generation immigrant | 0.09 | −0.06 | 0.16 | −0.94 | −0.17 | −0.77 |
| | (0.12) | (0.59) | | (0.63) | (0.13) | |
| Tertiary education | 0.05 | 0.24 | −0.18 | −0.23 | 0.13 | −0.36 |
| | (0.08) | (0.33) | | (0.34) | (0.10) | |
| Log net household income | 0.00 | 0.64*** | −0.64*** | 0.33 | −0.02 | 0.35 |
| | (0.08) | (0.21) | | (0.32) | (0.09) | |
| Managerial position | −0.04 | 0.39 | −0.44 | −0.65 | 0.05 | −0.70* |
| | (0.10) | (0.48) | | (0.40) | (0.13) | |
| Affiliated with religious community | −0.03 | 0.05 | −0.09 | 0.03 | 0.05 | −0.02 |
| | (0.08) | (0.32) | | (0.41) | (0.10) | |
| Extraversion | −0.01 | −0.05 | 0.04 | −0.01 | −0.00 | −0.00 |
| | (0.03) | (0.15) | | (0.15) | (0.04) | |
| Conscientiousness | 0.00 | −0.20 | 0.21 | 0.25 | −0.00 | 0.25 |
| | (0.04) | (0.17) | | (0.18) | (0.05) | |
| Openness to experience | 0.07** | 0.04 | 0.03 | −0.01 | −0.01 | 0.00 |
| | (0.03) | (0.14) | | (0.13) | (0.04) | |
| Neuroticism | −0.00 | 0.15 | −0.15 | −0.20 | −0.05 | −0.15 |
| | (0.03) | (0.11) | | (0.15) | (0.04) | |
| Agreeableness | 0.05 | 0.02 | 0.03 | 0.15 | −0.14*** | 0.29 |
| | (0.04) | (0.15) | | (0.18) | (0.05) | |
| Constant | 5.26*** | 3.86*** | 1.40*** | 3.82*** | 2.20*** | 1.62*** |
| | (0.10) | (0.42) | | (0.54) | (0.11) | |
| Adj. $R^2$ | 0.584 | | | | | |
| Observations | 2295 | | | | | |

Note: This table presents the estimated AMEs based on Equation (3). The first and second columns show the estimation results for the attribution of blame for the chairman's actions in the harm scenario, conditionally on whether individuals indicated that the chairman acted intentionally or did not act intentionally, respectively. The fourth and fifth columns show the corresponding estimation results for attributing praise in the help scenario. Δ indicates the difference between the AMEs by intentionality, and stars denote the significance level of a Wald test. AMEs for missing dummies are not shown. Robust standard errors are reported in parentheses. *$p < .1$,**$p < .05$,***$p < .01$. *Source*: SOEP-IS 2023.

1.62 in the help scenario (column 6, row "constant"). In both scenarios, the difference is statistically significant at the 1-percent level.

In terms of goodness-of-fit, the generalization of the basic OLS model (2) by intentionality implies a rise of the adjusted coefficient of determination ($\bar{R}^2$) from 0.528 to 0.584. There are two reasons for this. First, the indicator for intentionality captures differences in the reported levels of blame and praise between respondents who believe the chairman acted

intentionally and those who do not. i.e., respondents who attribute intentionality to the chairman assign significantly more blame or praise than those who do not perceive the chairman's action as intentional. The generalized model further demonstrates that the effect of certain covariates is conditional on whether respondents perceive the action as intentional. By contrast, the more restrictive model (2) captures only the average effect across both subgroups.

We begin by commenting on the role of socio-demographic variables. First, in the harm scenario, the reported level of blame rises in age at a deceasing rate but only among respondents who do attribute intentionality. In the help scenario, age is not among the significant covariates. Second, women tend to assign more blame than men, and this tendency is even more pronounced when they do *not* attribute intentionality: on average, women assign 0.18 points more blame when they attribute intentionality, and 0.73 points more when they do not. Interestingly, the corresponding difference between men and women does not emerge in the help scenario, indicating that women are stricter with the chairman in the harm scenario but not more critical than men in the help scenario. Third, married respondents blame the chairman less when they attribute intentionality than their unmarried counterparts. In the help scenario, married respondents report higher levels of praise than unmarried respondents, and this difference is larger when they attribute intentionality. Note however, that the two groups' AMEs are not statistically different.

Other significant findings include having older children, being a first generation immigrant and net household income. However, these effects do not present a coherent or systematic pattern across scenarios. All other socio-demographic factors, including managerial status, are found to be insignificant and do not differ between respondents attributing intentionality or not. The same holds for personality traits play. One exception is openness to experience among respondents who attribute intentionality in the harm scenario.

Table 10 shows the corresponding results of the stepwise regressions. Measured by the adjusted $R^2$ of 0.580, this more parsimonious stepwise regression model has almost exactly the same explanatory power as the comprehensive model just discussed. Further, the model confirms, quantitatively and qualitatively, the results regarding the role of age, gender, marital status, and income.

### *Synthesis of the heterogeneity analyses*

Overall, the multivariate heterogeneity analyses confirm the results of the descriptive replication: (1) more people ascribe intentionality in the harm scenario than in the help scenario; (2) the blame ratings are higher in the harm scenario than the praise ratings in the help scenario; and (3) ascribing

**Table 10.** Stepwise selected model: Chairman blame and praise evaluation by attribution of intentionality.

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | Blame in harm | | | Praise in help | | |
| *Intentionality:* | Yes | No | Δ | Yes | No | Δ |
| Age | 0.01*** | 0.02*** | −0.01* | 0.00 | −0.01* | 0.01 |
| | (0.00) | (0.01) | | (0.01) | (0.00) | |
| Female | 0.19*** | 0.74*** | −0.55** | −0.02 | −0.17** | 0.15 |
| | (0.07) | (0.26) | | (0.30) | (0.08) | |
| Married | 0.13* | −0.66** | 0.79*** | 0.68* | 0.21** | 0.47 |
| | (0.08) | (0.28) | | (0.36) | (0.10) | |
| Log net household income | 0.01 | 0.48** | −0.48** | 0.38 | 0.06 | 0.32 |
| | (0.07) | (0.20) | | (0.23) | (0.08) | |
| Constant | 5.16*** | 3.84*** | 1.32*** | 3.67*** | 2.29*** | 1.37*** |
| | (0.07) | (0.24) | | (0.33) | (0.08) | |
| Adj. $R^2$ | | | 0.580 | | | |
| Observations | | | 2295 | | | |

Note: This table presents the estimated AMEs of a model that includes only the variables selected in the stepwise regression. The backward stepwise selection was conducted using a critical significance level of $p = 0.01$ for term removal. The first and second columns show the estimation results for the attribution of blame for the chairman's actions in the harm scenario, conditionally on whether individuals indicated that the chairman acted intentionally or did not act intentionally, respectively. The fourth and fifth columns show the corresponding estimation results for attributing praise in the help scenario. Δ indicates the difference between the AMEs by intentionality, and stars denote the significance level of a Wald test. AMEs for missing dummies are not shown. Robust standard errors are reported in parentheses. *$p < .1$,**$p < .05$,***$p < .01$. *Source*: SOEP-IS 2023.

intentionality to the action matters for the respective level of blame and praise. However, we do find some heterogeneous associations not only in the probability of attributing intentionality but also in the blame/praise ratings. Heterogeneity also arises when comparing blame/praise ratings between those who believe the chairman brought about the side effects intentionally and those who do not.

Now, do people who are more likely to attribute intentionality also assign more blame or praise in the two scenarios? We would expect the direction of the associations in the intentionality regression to be identical with that in the blame/praise regression. If that is the case, we would have reason to believe that the associations observed between some socio-demographic characteristics and the attribution of intentionality are also at the root of the level of blame and praise attributed to the chairman's actions.

Let's look at the harm scenario first. There, all significant estimates for the socio- demographic variables discussed above indeed exhibit the same sign as the estimated AMEs of these variables on the level of blame (though not all reach statistical significance there too). For example, being a woman is associated with an *increased* probability of ascribing intentionality to the chairman's actions (Equation (1), Table 4). As one might expect, being a woman is also correlated with a sizable and statistically significant difference regarding the blame ratings: women assign substantially *more* blame than men in the harm scenario (Equation (2), Table 6). Interestingly, this gender difference exists both when respondents attribute intentionality and when they do not: women perceived the chairman's actions as bad,

regardless of whether they believed he brought about the side effects intentionally or not (Equation (3), Table 9). First-generation immigrants exhibit a significantly *lower* probability of attributing intentionality in the harm scenario than natives (see Table 4). At the same time, they show a sizable and statistically significant effect on blame ratings (see Table 6): they assign 0.55 points *less* blame compared to natives (which represents the largest average marginal effect among all AMEs on the blame rating) – again, as one might expect. By contrast with the case of women, this association is mostly driven by those respondents who indicated that the chairman acted intentionally (see Table 9).

Yet not all significant associations observed in blame ratings are mirrored by significant relationships in the attribution of intentionality. An illustrative case is agreeableness, which exhibits a weakly significant association on blame levels but shows no significant relationship with the probability of attributing intentionality (its AME is also substantively negligible).

Now let's look at the help scenario. By contrast to the harm scenario, it is not the case that all significant estimates for the socio-demographic variables in the probability of attributing intentionality exhibit the same sign as the AMEs of these variables on praise ratings. Furthermore, none of the AMEs on intentionality attribution are mirrored by statistically significant associations in the level of praise.

As an example, consider again first-generation immigrants. Even though they are 10.1% points *more* likely than natives to believe that the chairman *intentionally* brought about the environmental benefit (see Table 4) – a highly statistically significant difference and one of the largest AMEs – they show no corresponding significant association in the level of praise, whereas one might expect them to praise more. In fact, their AME on praise is close to zero (see Table 6). Generally speaking, the significant associations observed in praise ratings are not reflected in significant ones in the attribution of intentionality. Agreeableness is again a case in point: even though a one-point increase in agreeableness significantly *decreases* the praise rating by 0.12 points (see Table 6), the AME on the probability of attribution in the help scenario is close to zero and not statistically significant (see Table 4).

## Discussion

This study pursued two objectives: to replicate previous experimental studies, all relying on convenience samples, with a large-scale population-wide random sample, representative for the population of Germany, and to determine whether the response behavior this replication would observe varies with socio-demographic characteristics and personality traits.

Regarding the first objective, we observe that *almost all* of the respondents attribute intentionality in the harm scenario and *hardly any* attributed intentionality in the help scenario. This is strong empirical confirmation of what the literature had previously assumed on much weaker evidence: not just the existence, but also the universality of the Knobe effect – at least in Germany.

Figure 2 provides a summary of the main results of all previous studies on Knobe's (2003) original chairman experiment, ordered by sample size.[16] It yields two insights: First our results are more pronounced than generally observed so far. Only one study Nakamura (2018) observes even fewer respondents attributing intentionality to the chairman in the help scenario but no study has observed more people attributing intentionality in the harm scenario. Nevertheless the second-largest study Young et al. (2006) also suggests similarly pronounced differences in intentionality attributions between the

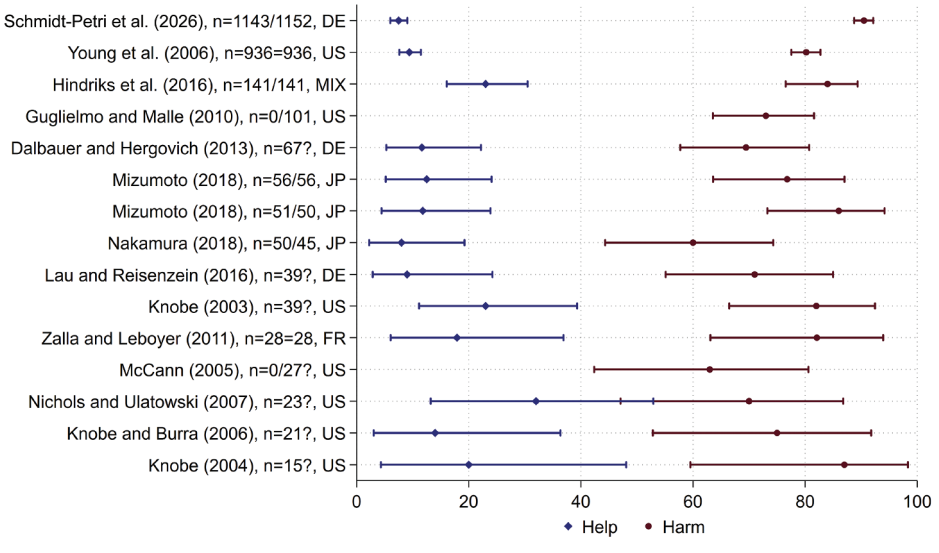


**Figure 2.** Results from previous studies. Note: This figure compares our results with the results from other studies examining the level of intentionality in the original 'chairman' scenarios taken from Knobe (2003). It displays the point estimates of proportions as reported in the studies (circles) along with 95% confidence intervals, which we derived from tests of equality of proportions. The studies are from a number of countries, as displayed; MIX stands for a mixture of many countries; the study by Mizumoto uses two different Japanese translations for 'intentionally,' 'Itoketi ni' and 'Wazato.' Not all studies have collected data for both the harm and help scenarios. The letter *n* denotes the scenario-specific sample sizes. 1143/1152 means that there were 1143 respondents in the help and 1152 in the harm scenario. 936 is the number of respondents participating in both scenarios. 67? means that it is unclear how respondents were divided among the different scenarios or random groups, and to calculate the confidence intervals, we used 67 – this being the result of dividing the overall number of respondents equally among the scenarios or random groups (sometimes rounding up or down). Supplementary Material **C** provides further details.

harm and help scenarios. Indeed their point estimate for the help scenario is not statistically different from ours and for the harm scenario their estimate is not that much smaller. Second our large sample yields a tight estimation of the shares of respondents attributing intentionality: the confidence bands from our study are significantly narrower than in any previous studies.

Given our sample size and the effective randomization, we consider our results to be the statistically most accurate results yet.[17] Such inter-study comparisons need to be taken with a grain of salt, however. In addition to the obvious facts that the questionnaires have been administered at different times, in different countries and in different languages (and even within any language might differ in many minor details), three more technical aspects must be considered. First, the confidence intervals of many previous studies are very wide and overlap with the confidence intervals of our study in about half of cases. Hence, despite the apparent large differences in the point estimates, one cannot reject the hypothesis that at least these estimates are statistically not different from ours. However, secondly, all previous studies have been based on convenience samples, which, due to the potential selection biases mentioned above, raises questions about the generalizability of their findings. Such selection biases are quite likely to be present even when the results do not differ from ours on the surface; any coherence would simply be a matter of chance. A third explanation of the large differences is also possible: that there *are* culture-specific differences in the cognitive process of ascribing intentionality and the mediation by the perceived moral worth of the action. Interestingly, it seems that the few studies from Germany and Japan, respectively, are more coherent among themselves than the more numerous studies performed in the United States.[18]

Regarding the second objective, we thought that we were unlikely to find heterogeneities among the different subpopulations, given the above results. Still, three socio-demographic characteristics are worth highlighting overall: gender, immigration status, and professional status. We observe a clear difference between men and women, both concerning the attribution of intentionality as well as concerning the level of praise and blame, especially in the harm scenario. First-generation immigrants, by contrast, generally do not display as strong a Knobe effect as native Germans, while second-generation immigrants tend to cohere more with the overall results. This might be due, for instance, to differing moral norms or language skills that change over time as people get accustomed to local circumstances. Respondents in managerial positions do not differ significantly from the general population in how much blame or praise they heap on their imaginary colleague – their final moral assessment is not generally divergent –, but they are much more likely to attribute intentionality in the harm

scenario and much less likely to do so in the help scenario. It is plausible to believe that these respondents are in a better position to look at the scenarios "from the inside" of the chairman's head, as it were. Contrary to what might have been expected, the personality traits of the respondents as measured by the Big Five barely influence the results. In particular, extraversion does not have any significant effect in any of our analyses, contrary to the observations described in Feltz and Cokely (2024).

Our study design leaves a few issues unexplored, and, together with the result that the heterogeneities in the overall results are minor, this has several implications for future research. Let's start with the empirical side. First, as might be expected, any future studies should use random samples if at all possible and control for (at least) the three important variables mentioned. Secondly, it deserves emphasis that even heterogeneities with minor effects might have a major impact in small convenience samples. For instance, if 3 out of 4 undergraduate students in psychology are female (Odic & Wojcik, 2020) and the sample is taken from an average psychology class, it will almost certainly give biased results. If the population of Germany is anything to go by, this bias will lead to a stronger Knobe effect than the overall population would display. Hence, it would help to balance the sample in this dimension, if for practical reasons convenience samples are inevitable. Third, – even though it almost goes without saying – it would be desirable to perform even larger studies so as to be able to zoom in on potential *combinations* of characteristics. A limitation of our sample is that it is still too small to make meaningful claims about the group of native German women in managerial positions without a religions affiliation, for instance, and as discussed, this is one group that differs substantially from the rest of the population in the attribution of intentionality.[19] With even larger samples, several further groups could also be included that were missing from our sample altogether, such as children and adolescents. Our sample also does not include residents in nursing homes or other institutional accommodations. As the importance of the immigration status suggests that language skills and language acquisition might be relevant factors, a larger sample should ideally be more fine-grained in these dimensions (for instance, control for different original cultural or linguistic backgrounds) and ideally be supplemented with longitudinal observations. Our findings may help to set up a testable theory for such studies.

Considering the more philosophical side, we need to accept that our study design does not allow us to draw inferences about individual's reasoning. It highlights the importance of designing any future empirical studies carefully, however. Again, if three competing hypotheses about individual-level thinking in the harm and help scenarios were to be tested against each other, at the very least all six groups should contain the same number of

women, for instance, again supposing that our results would not be specific to Germany.

One might also raise the more fundamental objection that precise estimates of the strength of the Knobe effect or its statistical explanation through socio-economic variables are anyway of little relevance to the philosophical or psychological issues raised by the Knobe effect. While we have some sympathy with this approach, it effectively begs the question against experimental philosophy altogether. If experiments in philosophy are performed to shed light on intuitions of the general population, and thus how the intuitions of non-philosophers compare to those of philosophers, it would simply be misguided to not collect the best data possible or to ignore an important aspect of it – namely that some people do not display the Knobe effect, or that women apparently think about it differently than men. We suggest that future research should take these results into consideration.

## Conclusion

We replicate Knobe's (2003) canonical experiment with a large random sample to assess the external validity of the findings and obtain more precise estimates of effect sizes. We find the effect to be even stronger than previously estimated. We also investigate whether the magnitude of the Knobe effect differs between different population groups, as characterized by a broad set of socio-demographics and personality traits, finding that gender, migration status, and being in a managerial position matter for the magnitude of the Knobe effect.

## Notes

1. The standard error is inversely proportional to the square root of the sample size.
2. For example, randomization minimizes the risk of biased sample composition, such as over-representation of certain demographic groups, which may occur when using convenience samples. This strengthens the validity of our comparisons across conditions.
3. To comply with the word limit allocated to us in the SOEP-IS questionnaire module, we had to rephrase the original wording minimally. The translation was ours and it was pretested. The German for the original "intentionally" is "absichtlich," there is no reason to believe it functions differently in German than in English.
4. We are performing multiple statistical tests on different variables within the same dataset (multiple hypothesis testing) and given that each test has a certain probability of producing a false positive result (a type I error), with 21 tests, it is likely that at least one will show a significant result just by random chance, even if there is no true effect.
5. As the scales of blame/praise vary across studies, we do not provide a comparison for average levels of blame/praise.
6. Not all respondents reported all socio-demographics. This item non-response is a standard issue in voluntary surveys. For the variables marital status, education,

and affiliation with a religious community, missing values were logically imputed by replacing them with valid values from previous survey waves, when available. For missing values in categorical variables, we added an additional "missing" category. For metric variables, we replaced missing values with the mean and included in the regressions a missing dummy. We removed the remaining 52 observations with missing values from our sample.

7. Supplementary Material **A** provides further details on the definition of the variables in the vector $\mathbf{X}_i$.
8. For example, a positive AME for the gender dummy would indicate that female respondents are more likely than male respondents to attribute intentionality.
9. AMEs are computed using the actual observed values of variables that are not held fixed. In the OLS models, they have the same interpretation as the coefficients.
10. The estimation results obtained from the stepwise selection with thresholds of $p = 0.1$ and $p = 0.05$ are not shown but are available upon request.
11. If all the average marginal effects of all control variables were zero, the estimated AME of the baseline probability would correspond to the descriptive sample mean.
12. The pseudo adjusted coefficient of determination accounts for the goodness of fit, but also penalizes the inclusion of additional variables. Hence, it can increase after removing variables (with little explanatory power) from the model.
13. All regression tables in the paper report standard errors in parentheses. Supplementary Material **D** presents the same tables, but with *p*-values in parentheses instead.
14. Following the suggestion of an anonymous reviewer, we also estimate a linear probability model using the same specification as the stepwise logit model. Table E1 in Supplementary Material **E** reports the results. AMEs from the logit model and the AMEs from the linear model are qualitatively consistent (same signs) and quantitatively very similar. For instance, in the harm scenario, the AME for managerial position is 0.042, while the corresponding AME in the linear probability model is 0.043.
15. We have included the second-generation immigrant variable for reasons provided in Section Analytic approach and estimation.
16. Beyond the studies presented in Figure 2, a number of additional studies employ variations of the original design. These were excluded from the overview to ensure the highest possible comparability of results. It should be noted that not all previous studies report confidence intervals for their point estimates or use divergent methods for their calculation. Therefore, we (re-)calculate the confidence intervals from the point estimates provided and numbers of scenario-specific observations. As the attribution of intentionality to the chairman implies a discrete distribution with two outcomes (yes = 1 or no = 0), we determine the confidence intervals using the binomial distribution with probability mass function, $f(k,n,p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$, with $X$ denoting the random variable, $k$ the successful of all $n$ trials, $p \in [0,1]$ the success probability, and $q = 1 - p$. The confidence interval from a binomial distribution is non-symmetric, because the binomial distribution itself is discrete and skewed when the probability of success, $p$, is not 0.5. For example, when $p$ is close to zero or one, the probability mass is concentrated near the endpoints of the interval; consequently the upper and lower bounds of the confidence interval will differ more widely. To compute the confidence bands, we use the Stata command *proportion*; for details see Stata's "Methods and formulas" section (p. 2220f.).

17. There are clear formal arguments for the claim that our study provides the most accurate results: The central limit theorem states that, provided an estimator (like the population mean) fulfills certain weak conditions, then, for reasonable sample sizes, the sampling distribution of the estimator converges to normality. Most importantly for our purposes, the theorem states that the distribution of the sample value of the estimator will approach the actual population-wide parameter as the sample size increases and sampling error decreases. How confident can we be that these sample-based probabilities comply with those for the overall German population? Our answer in short is: statistically speaking, we can be very confident. We illustrate this by Monte Carlo simulations detailed in the Supplementary Material **B**.
18. The relevance of the immigration status we observe also suggests that cultural differences might be at play.
19. Of course, it could also be worth exploring in more detail smaller samples of these specific groups.

## Acknowledgements

Our analyses rely on data from the Innovation Sample of the Socio-Economic Panel (SOEP-IS). All contents of the panel are approved by a Research Ethics Board. SOEP data are individual-level data; therefore, legal conditions for using the data apply; users are not permitted to disseminate the data. However, scientists can easily access the data at no cost by signing a data distribution contract (contact email: soepmail@diw.de). A replication package is available, stored in the *Reanalysis Archive of the SOEP*. We thank Ines Weyland for outstanding research assistance and Martin Gerike for his organizational support in implementing the Knobe module in the SOEP-IS. We thank Denis Gerstorf, Theresa Entringer, and Edmond Purssell as well as the audience at X-Phi 2025 at UEA Norwich for helpful comments.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

This is independent research, which was not funded by external parties.

## ORCID

Christoph Schmidt-Petri http://orcid.org/0000-0001-7990-1669
Daniel Labarca-Pinto http://orcid.org/0009-0006-2075-9622
Carsten Schröder http://orcid.org/0000-0002-6406-595X

## Artificial intelligence

No artificial-intelligence-assisted technologies were used in this research or the creation of this article.

## Analysis scripts

All analysis scripts and a replication file are available in the Reanalysis Archive of the SOEP https://www.diw.de/en/diw_01.c.604159.en/reanalysis_archive_of_the_soep.htmlReanalysis.

## Data

Our data come from the German Socio-Economic Panel (SOEP). The SOEP data is available, free of cost, to the worldwide scientific community. A data contract is required (see https://www.diw.de/en/diw_01.c.601584.en/data_access.htmlSOEP user contract).

## Ethics

This research and data collection received approval from the Survey Committee of the German Socio-Economic Panel.

## Materials

Our study materials are publicly available via https://www.diw.de/en/diw_01.c.604159.en/reanalysis_archive_of_the_soep.htmlReanalysis.

## Preregistration

The survey module and research questions were pre-registered before data collection (see https://companion-is.soep.de/)

## References

Adams, F., & Steadman, A. (2004a). Intentional action and moral considerations: Still pragmatic. *Analysis*, *64*(3), 268–276. https://doi.org/10.1093/analys/64.3.268

Adams, F., & Steadman, A. (2004b). Intentional action in ordinary language: Core concept or pragmatic understanding? *Analysis*, *64*(2), 173–181. https://doi.org/10.1093/analys/64.2.173

Alicke, M. (2008). Blaming badly. *Journal of Cognition and Culture*, *8*(1–2), 179–186. https://doi.org/10.1163/156770908X289279

Bhatia, D., Fischbacher, U., Hausfeld, J., & Stumpf, R. (2024). Blame and praise: Responsibility attribution patterns in decision chains. *Experimental Economics*, *27*(3), 637–663. https://doi.org/10.1007/s10683-024-09833-1

Buckwalter, W., & Stich, S. (2014). Gender and philosophical intuition. In J. Knobe & S. Nichols (Eds.), *Experimental philosophy: Volume 2* (pp. 307–346). Oxford University Press. https://doi.org/10.1093/acprof:osobl/9780199927418.003.0013

Cardinale, E. M., Finger, E. C., Schechter, J. C., Jurkowitz, I. T., Blair, R., & Marsh, A. A. (2014). The moral status of an action influences its perceived intentional status in adolescents with psychopathic traits. In T. Lombrozo, J. Knobe, & S. Nichols (Eds.), *Oxford Studies in Experimental Philosophy* (Vol. 1, pp. 131–151). Oxford University Press UK. https://doi.org/10.1093/acprof

Cova, F., & Naar, H. (2012). Side-effect effect without side effects: The pervasive impact of moral considerations on judgments of intentionality. *Philosophical Psychology*, *25*(6), 837–854. https://doi.org/10.1080/09515089.2011.622363

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., Beebe, J., Berniu¯nas, 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. . . . Wilkenfeld, D. (2021). Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology*, *12*(1), 9–44. https://doi.org/10.1007/s13164-018-0400-9

Cushman, F., & Mele, A. (2008). Intentional action: Two and half folk concepts? In J. Knobe & S. Nichols (Eds.), *Experimental philosophy* (pp. 171–188). Oxford University Press. https://doi.org/10.1093/oso/9780195323252.003.0009

Dalbauer, N., & Hergovich, A. (2013). Is what is worse more likely?-the probabilistic explanation of the epistemic side-effect effect. *Review of Philosophy and Psychology*, *4* (4), 639–657. https://doi.org/10.1007/s13164-013-0156-1

Feltz, A., & Cokely, E. T. (2024). *Diversity and Disagreement: From Fundamental Biases to Ethical Interactions*. Palgrave Macmillan. https://doi.org/10.1007/978-3-031-61935-9

Fischbacher, U., Neyse, L., Richter, D., & Schröder, C. (2024). Adding household surveys to the behavioral economics toolbox: Insights from the SOEP Innovation sample. *Journal of the Economic Science Association*, *10*(1), 136–151. https://doi.org/10.1007/s40881-023-00150-6

Goebel, J., Grabka, M. M., Liebig, S., Kroh, M., Richter, D., Schröder, C., & Schupp, J. (2019). The German Socio-Economic Panel (SOEP). *Jahrbücher für Nationalökonomie und Statistik*, *239*(2), 345–360. https://doi.org/10.1515/jbnst-2018-0022

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

Hindriks, F., Douven, I., & Singmann, H. (2016). A new angle on the Knobe effect: Intentionality correlates with blame, not with praise. *Mind & Language*, *31*(2), 204–220. https://doi.org/10.1111/mila.12101

Kneer, M., & Bourgeois-Gironde, S. (2017). Mens rea ascription, expertise and outcome effects: Professional judges surveyed. *Cognition*, *169*, 139–146. https://doi.org/10.1016/j.cognition.2017.08.008

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

Knobe, J. (2004). Intention, intentional action and moral considerations. *Analysis*, *64*(2), 181–187. https://doi.org/10.1093/analys/64.2.181

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

Knobe, J., & Burra, A. (2006). The folk concepts of intention and intentional action: A cross-cultural study. *Journal of Cognition and Culture*, *6*(1–2), 113–132. https://doi.org/10.1163/156853706776931222

Lau, S., & Reisenzein, R. (2016). Evidence for the context dependence of the side-effect effect. *Journal of Cognition and Culture*, *16*(3–4), 267–293. https://doi.org/10.1163/15685373-12342180

Leslie, A. M., Knobe, J., & Cohen, A. (2006). Acting intentionally and the side-effect effect theory of mind and moral judgment. *Psychological Science*, *17*(5), 421–427. https://doi.org/10.1111/j.1467-9280.2006.01722.x

McCann, H. J. (2005). Intentional action and intending: Recent empirical studies. *Philosophical Psychology*, *18*(6), 737–748. https://doi.org/10.1080/09515080500355236

Michael, J. A., & Szigeti, A. (2019). "The group "Knobe effect": Evidence that people intuitively attribute agency and responsibility to groups. *Philosophical Explorations*, *22* (1), 44–61. https://doi.org/10.1080/13869795.2018.1492007

Mizumoto, M. (2018). A simple linguistic approach to the Knobe effect, or the Knobe effect without any vignette. *Philosophical Studies*, *175*(7), 1613–1630. https://doi.org/10.1007/s11098-017-0926-1

Nadelhoffer, T. (2004). Blame, badness, and intentional action: A reply to Knobe and Mendlow. *Journal of Theoretical and Philosophical Psychology*, *24*(2), 259–269. https://doi.org/10.1037/h0091247

Nadelhoffer, T. (2006). Desire, foresight, intentions, and intentional actions: Probing folk intuitions. *Journal of Cognition and Culture*, *6*(1–2), 133–157. https://doi.org/10.1163/156853706776931259

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

Nichols, S., & Ulatowski, J. (2007). Intuitions and individual differences: The Knobe effect revisited. *Mind & Language*, *22*(4), 346–365. https://doi.org/10.1111/j.1468-0017.2007.00312.x

Odic, D., & Wojcik, E. H. (2020). The publication gender gap in psychology. *The American Psychologist*, *75*(1), 92–103. https://doi.org/10.1037/amp0000480

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

Utikal, V., & Fischbacher, U. (2014). Attribution of externalities: An economic approach to the Knobe effect. *Economics & Philosophy*, *30*(2), 215–240. https://doi.org/10.1017/s0266267114000170

Weinberg, J. M., Nichols, S., & Stich, S. (2001). Normativity and epistemic intuitions. *Philosophical Topics*, *29*(1 & 2), 429–460. https://doi.org/10.5840/philtopics2001291/217

Woolfolk, R. L. (2013). Experimental philosophy: A methodological critique. *Metaphilosophy*, *44*(1–2), 79–87. https://doi.org/10.1111/meta.12016

Young, L., Cushman, F., Adolphs, R., Tranel, D., & Hauser, M. (2006). Does emotion mediate the relationship between an action's moral status and its intentional status? Neuropsychological evidence. *Journal of Cognition and Culture*, *6*(1–2), 291–304. https://doi.org/10.1163/156853706776931312

Zalla, T., & Leboyer, M. (2011). Judgment of intentionality and moral evaluation in individuals with high functioning autism. *Review of Philosophy and Psychology*, *2*(4), 681–698. https://doi.org/10.1007/s13164-011-0048-1