Deer, Lachlan; Krishna, Adithya; Zhang, Lyla

Working Paper

Replication Report: Corrupted by Algorithms? How AI-generated And Human-written Advice Shape (Dis)Honesty

I4R Discussion Paper Series, No. 212

Provided in Cooperation with:
The Institute for Replication (I4R)

Suggested Citation: Deer, Lachlan; Krishna, Adithya; Zhang, Lyla (2025) : Replication Report: Corrupted by Algorithms? How AI-generated And Human-written Advice Shape (Dis)Honesty, I4R Discussion Paper Series, No. 212, Institute for Replication (I4R), s.l.

This Version is available at:
https://hdl.handle.net/10419/313185

INSTITUTE for REPLICATION

No. 212
I4R DISCUSSION PAPER SERIES

Replication Report: Corrupted By Algorithms? How AI-generated And Human-written Advice Shape (Dis)Honesty

Lachlan Deer

Adithya Krishna

Lyla Zhang

March 2025

INSTITUTE for REPLICATION

I4R DISCUSSION PAPER SERIES

I4R DP No. 212

Replication Report: Corrupted by Algorithms? How AI-generated And Human-written Advice Shape (Dis)Honesty

Lachlan Deer[1], Adithya Krishna[2], Lyla Zhang[2]

[1]Tilburg University, Tilburg/The Netherlands
[2]Macquarie University, Macquarie Park/Australia

MARCH 2025

E-Mail: joerg.peters@rwi-essen.de
RWI – Leibniz Institute for Economic Research
Hohenzollernstraße 1-3
45128 Essen/Germany
www.i4replication.org

ISSN: 2752-1931

Replication Report: Corrupted By Algorithms? How AI-generated And Human-written Advice Shape (Dis)Honesty *

Lachlan Deer Adithya Krishna Lyla Zhang

February, 2025

Abstract

Leib et al. (2024) examine how artificial intelligence (AI) generated advice affects dishonesty compared to equivalent human advice in a laboratory experiment. In their preferred empirical specification, the authors report that dishonesty-promoting advice increases dishonest behavior by approximately 15% compared to a baseline without advice, while honesty-promoting advice has no significant effect. Additionally, they find that algorithmic transparency - disclosing whether advice comes from AI or humans - does not affect behavior. We computationally reproduce the main results of the paper using the same procedures and original data. Our results confirm the sign, magnitude, and statistical significance of the authors' reported estimates across each of their main findings. Additional robustness checks show that the significance of the results remains stable under alternative specifications and methodological choices.

KEYWORDS: artificial intelligence, dishonesty, laboratory experiment, computational reproducibility

JEL CODES: D01, D91, C91

*Author Contacts: Deer: Department of Marketing, Tilburg University. E-mail: lachlan.deer@gmail.com; Krishna: Department of Economics, Macquarie Business School, Macquarie University. Email: adithya.krishna@hdr.mq.edu.au Zhang: Department of Economics, Macquarie Business School, Macquarie University. Email: lyla.zhang@mq.edu.au.

The authors did not receive any financial support for this study and report no conflicts of interest. The authors thank the organizers of the Rotterdam Replication Games in June 2024 for providing the platform to initiate this replication.

1 Introduction

Leib et al. (2024), hereafter LKRHI, investigate how artificial intelligence (AI) generated advice shapes dishonesty compared to equivalent human advice. This question has become increasingly relevant as AI systems like large language models take on advisory roles across various domains. Although organizations commonly propose algorithmic transparency as a solution to potential AI risks, empirical evidence for its effectiveness in shaping ethical behavior remains limited. Through a series of laboratory experiments, the authors examine: (i) whether people's dishonest behavior is influenced by AI-generated advice, (ii) how such advice compares to human-written advice, and (iii) whether transparency about the advice source influences behaviour.

In their preferred empirical specification, LKRHI find that dishonesty-promoting advice increases dishonesty by 15% compared to a baseline with no advice, regardless of whether the advice comes from AI or humans. In contrast, honesty-promoting advice does not significantly increase honest behavior. Notably, algorithmic transparency - informing participants about whether advice comes from AI or humans - does not affect behavior. These findings suggest that AI advice can shape unethical behavior to the same extent as human advice, and that common policy proposals like algorithmic transparency may be insufficient to mitigate these risks.

In the present paper, we investigate both the computational reproducibility and robustness of LKRHI's empirical results. Using their data and code provided in their replication package, we successfully reproduce all five main findings from their analysis with identical point estimates and standard errors. Our robustness analysis examines the stability of these findings under alternative specifications and methodological choices including making regression standard errors heteroskedasticity robust, use Bonferroni adjusted standard errors for t-tests and alternative regression specifications.

The remainder of the paper proceeds as follows. Section 2 provides context on the

laboratory experiments to help contextualize the task and study's findings. Section 3 outlines the replication materials and describes our approach to both computational reproduction and robustness checks. Section 4 presents results from our re-analysis and Section 5 discusses the results of our robustness checks. Section 6 provides a brief conclusion and discusses main takeaways from the exercise.

2 Context: The Experiments

Leib et al. (2024) conduct a series of experiments to investigate how AI-generated and human-written advice shapes dishonesty. The core of the experimental design is and incentivized die-rolling task where participants privately roll a die and report the outcome, with higher reported numbers corresponding to higher payments (following Fischbacher and Föllmi-Heusi (2013)). In this setup, participants face a trade-off between honesty and monetary gain, as they could potentially increase their earnings by misreporting higher numbers.

The experiment consists of two main parts. In the first part, the authors collect human-written advice and generate AI advice. For the human-written advice, advisors (N = 367) were incentivized to write either honesty-promoting or dishonesty-promoting advice for future participants. The authors then use this human-written advice to fine-tune GPT-J, an open-source natural language processing algorithm, to generate comparable AI advice. In the second part of the experiment, the authors examine how this advice influences participants' dishonesty through a between-subjects design with nine treatment conditions, as described in Table 1.

The main dependent variable is the reported die roll outcome (ranging from 1 to 6), with an expected average of 3.5 if the participants report truthfully. While individual-level dishonesty cannot be detected because the result of the actual die role is not reported, systematic deviations from this expected average across treatments reveal the influence of different types of advice on dishonest behavior. Following the die-rolling task, participants completed a post-experimental survey

Table 1: Overview of Experimental Treatments – Part 2 of LKRHI

Treatment	Source	Type	Information	No. Subjects
Control Condition				
0	No advice	-	-	201
Treatment Conditions				
1	AI	Honesty	Transparent	201
2	AI	Honesty	Opaque	201
3	AI	Dishonesty	Transparent	200
4	AI	Dishonesty	Opaque	200
5	Human	Honesty	Transparent	203
6	Human	Honesty	Opaque	201
7	Human	Dishonesty	Transparent	205
8	Human	Dishonesty	Opaque	205
			Total # Subjects	1,817

Notes: The Table reports the treatments implemented in Part 2 of the experiment by Leib et al. (2024). '*Source*' refers to whether the advice was given by humans (Human) or generative AI (AI). '*Type*' refers to kind of advice given to a subject about how to report their die roll and is set to be either honesty promoting (Honesty) or dishonesty promoting (Dishonest). '*Information*' refers to whether subjects are informed about the source of information being human or AI (Transparent) or not informed (Opaque). In the Control Condition (No Advice) subjects were not provided with advice from an adviser.

measuring several potential mechanisms, including perceptions of social norms, justifiability of dishonesty, and attribution of responsibility between themselves and the advisor. In treatments where the advice source was not disclosed (i.e. the opacity conditions), participants also completed an incentivized static Turing test to assess whether they could distinguish between AI and human-written advice.

3 Data, Replication Materials & Approach

We access the data and code provided by LKRHI on OSF.[1] The authors of the original study include their R code and two data files along with their preregistration documents. Preregistration documents provided pertain to the design of the experiment and do not discuss planned analyses of the collected data. Web Appendix Table A.1 summarizes the features of the replication package. The codes and data used in our reproduction, including a workflow to execute the analysis are made available in a separate OSF repository.[2]

[1]OSF URL: https://osf.io/g3sw2/
[2]OSF URL: https://osf.io/g6249/.

Our approach. We independently re-coded the analysis without reference to the original scripts. Whilst mostly successful, due to the absence of a README file describing the variables in the data set we did need to verify which variable in the provided dataset indicated a subject completed all parts of the experiment and should be included in the analysis by looking at some lines of the author's code. We reproduced all results (tables and figures) in the main text of LKHRI except their Bayesian analysis reported in the appendix. We opted not to reproduce the Bayesian analysis, as these results primarily serve to support the main findings.

After re-coding the results, we ran the scripts of LKHRI and verified that all the main results in the published manuscript are produced by their code.

4 Computational Reproducibility Results

4.1 Main Results

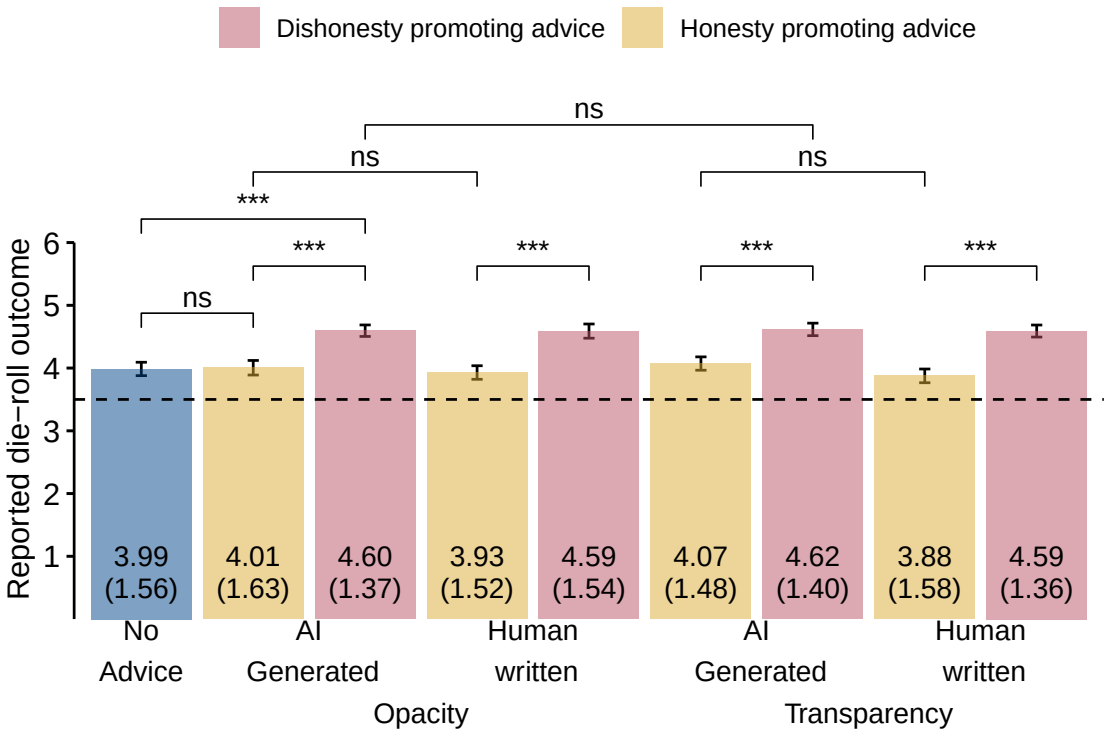
This section reproduces the main results of LKHRI as presented in Section 3 of their paper.

Over-reporting of die-roll outcomes. LKHRI's first result is that subjects over-report the die-roll outcomes across all treatments. Our results confirm this finding. Figure 1 reports the average reported die-roll for each treatment along with the standard error, partially reproducing Figure 3 of LKHRI. The figure also reports the standard deviation of the die-roll outcome in parentheses under the mean for each treatment. Mean reported outcomes in each treatment are higher than the expected value of 3.5 ($p < 0.00$ for all treatments. See Web Appendix Table B.1 for one-sample t-test results).

Is people's behaviour influenced by AI-generated advice? Column (1) of Web Appendix Table B.3 reveals average die-roll reports higher in the AI-generated dishonesty treatment compared to the No Advice treatment ($b = 0.61$, $p < 0.00$).[3]

[3]Estimates in We Appendix Table B.3 reproduce the results in Table 1 of LKHRI. Panel A reports their estimates on the treatment coefficients, and Panel B reports our estimates. The

Figure 1: Average Die-Roll Reports Across Treatments



Notes: The Figure shows mean reported die-roll outcomes (bars) by treatment and their standard errors. The dashed black line represents the expected mean if participants were honest. The means (standard deviations) of die-roll reports are given at the bottom of each bar. Statistical test results reported are from pairwise t-tests with standard errors that are not corrected for multiple testing. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Figure reproduces the mean die-roll related aspects of Figure 3 in Leib et al. (2024).

Average die-roll reports are not statistically different from the No Advice treatment when the AI generated advice is honesty promoting ($b = 0.019$, $p = 0.898$). Die rolls in AI generated dishonesty promoting are higher the an in AI generated honesty promoting treatments ($b = -0.590$, $p < 0.00$).

Is people's behaviour influenced by AI-generated advice? AI generated advice performs similarly to human generated advice. Column (2) of Web Appendix Table B.3 reports the results. For the opacity treatment, AI generated advice does lead to differences in average die-roll reporting compared to human generated advice when dishonesty promoting ($b = 0.070$, $p = 0.744$) of honesty promoting ($b = -0.076$, $p = 0.631$).

Can individual's distinguish AI and human-written advice? In the opacity treatments, participants cannot distinguish AI advice from human advice (Binomial test results using a frequency threshold of 50 percent, p = 0.999).

Does transparency about the advice source matter? Column (3) of Web Appendix Table B.3 include interactions with an indicator for treatments that reveal advice sources to subjects. The results are statistically insignificant. There is no evidence that transparency on the advice source influences reported die-roll outcomes.

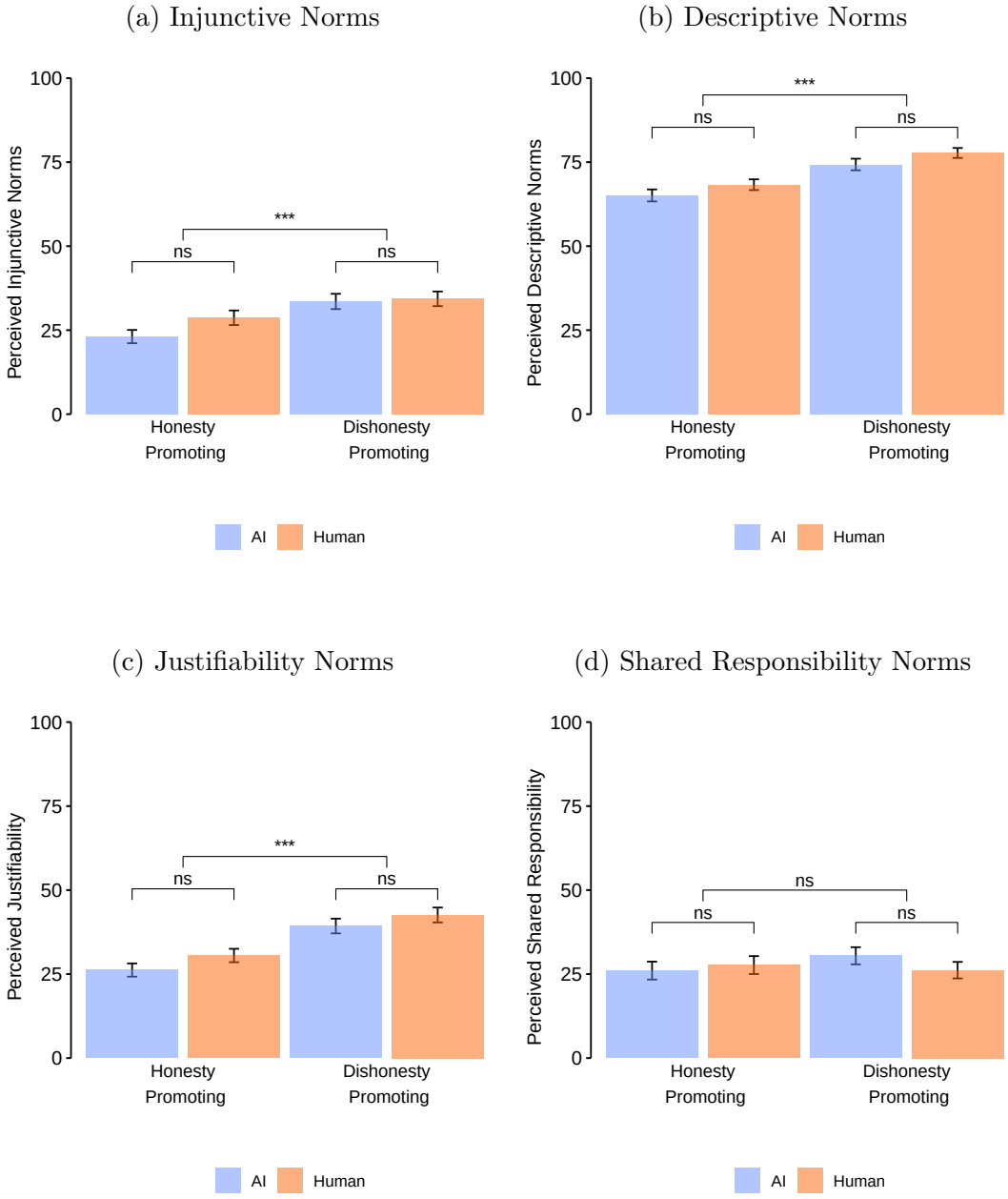
The estimates discussed above are robust to the inclusion of controls including perceived norms, gender, age, advice readability and whether a subject correctly guesses the advice source (Columns (4) to (7) of Web Appendix Table B.3).

4.2 Mechanisms

LKHRI examine how the advice source (AI or human) and type (honesty or dishonesty promoting) affect participants' perceptions across four key aspects: whether

estimates and standard errors coincide in all columns. Web Appendix Table B.2 reports analogous results using pairwise t-tests with a Bonferroni adjustment for multiple hypothesis testing.

Figure 2: Reports of Perceived Norms by Treatment Type



Notes: The Figure shows mean reports of perceived norms (bars) by treatment and their standard errors. Statistical test results reported are from pairwise t-tests with standard errors that are not corrected for multiple testing. Figure reproduces Figure 4 in Leib et al. (2024). See Section 2.2.3 of LKHRI for norm definitions.

misreporting is appropriate (injunctive social norms), how common they think it is (descriptive social norms), how justifiable they consider it (justifiability norms), and how they share responsibility with their advisor (shared responsibility norms). Subjects completed a post-experimental survey, indicating on a scale from 0 to 100 their perception of each norm. To understand how known advice source and advice type shaped perceptions, the authors focus on subjects who where in treatments where they informed about the advice source (i.e. transparency treatments).

Figure 2 reproduces Figure 4 of LKHRI's manuscript and reports the mean of each perceived by treatment. The figure exactly mirror those of the authors and shows that that dishonesty-promoting advice increases participants' perceptions of the appropriateness, prevalence, and justifiability of their behavior compared to honesty-promoting advice, regardless of whether the advice comes from AI or humans. Perceptions of shared responsibility are similar across treatments. These results are supported via linear regression of each of the perceived norms on treatment indicators. The results are presented in Web Appendix Table B.4 and mirror those of the in text discussion of the authors on pages 778-780.[4]

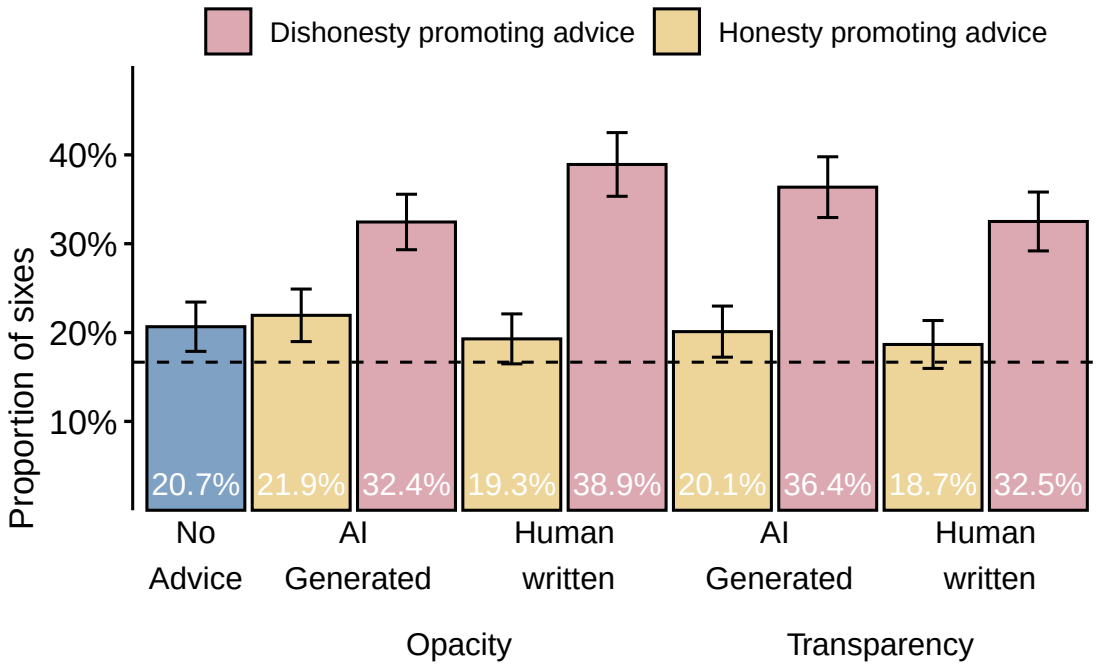
4.3 Author Reported Robustness

After establishing the main results, LKHRI show that their results are robust across two dimensions. First, the authors show that the results are robust to a change of outcome variable, using the proportion of reported sixes (which is the report that yields subjects the maximum payment). Second, they show their results are robust to the incentive alignment schemes of the advisors. In what follows, we reproduce their results for these robustness exercises.

Reported Sixes. LKHRI show that their results are robust to an alternative outcome, whether a subject reports a die-roll outcome of six. Figure 3 partially reproduces Figure 3 of LKHRI reporting the proportion of sixed reported within each

[4]The regression coefficients and statistical test results are identical to those produced when running the scripts of LKHRI.

Figure 3: Proportion of Die Rolls Reported as Six by Treatment



Notes: The Figure shows mean proportions of sizes (bars) by treatment and their standard errors. The dashed black line represents the proportion of sixes if participants were honest. The means of die-roll reports are given at the bottom of each bar. Figure reproduces the proportion of sixes related aspects of Figure 3 in Leib et al. (2024).

treatment. The results show a similar pattern the average die roll outcomes shown in Figure 1. The proportion of sixes are higher in dishonesty promoting advice treatments. Treatments promoting honesty have a similar proportion of sixes to the no advice condition. These results are confirmed via regression. Columns 1 to 3 of Table B.5 reproduce the (unreported) probit regression results of LKHRI that show the main findings in terms of treatment differences are robust to the change in the outcome variable. Our estimates align with those reported in the manuscript.

Additional Treatments. LKHRI run four additional treatments to explore the robustness of their results to the advisor's incentive scheme. In these treatments, participants in Part Two of the experiment read advice written by advisors whose incentives were aligned with those of the advisees. Otherwise the treatments were identical to the original experiment: they differed by advice source (human written

Table 2: Replication of Pairwise t-statistics: Aligned Treatments

	Opacity	
	AI Generated	Human Written
No Advice	-1.73	-4.08 ***
AI × Dishonest × Opaque	-	-2.14
	Transparency	
	AI Generated	Human Written
No Advice	-4.00 ***	-3.27 **
AI × Dishonest × Opaque	-2.07	-1.41
Human × Dishonest × Opaque	0.08	0.73
AI × Dishonest × Transparent	-	0.65

Notes: Table reports (two-sided) t-test results of whether the mean reported die-roll outcome were equal between treatments. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$ using Bonferroni-Holm adjusted p-values. Reproduces reported statistics of Leib et al. (2024) reported in Section 3.1.2. See Table 1 and notes therein for treatment definitions.

versus AI generated) and information (transparency versus opacity). We reproduce the pairwise t-test results from the paper in which the authors compare the average reported die roll of these treatments to the "no advice" condition and to each other. We report the results in Table 2. Again, our results mirror those of the authors.

5 Robustness Reproducibility

5.1 Econometric Specifications.

Heteroskedasticity Robust Standard Errors. LKHRI's regressions in their Table 1, which we have reproduced in Web Appendix Table B.3 assume that the regression residual is homoskedastic. Web Appendix Tables C.1 and C.2 show that the results are robust to using HC2 and HC3 heteroskedasticity robust standard errors.

Proportion of Sixes via LPM. Economists when estimating casual effects of binary treatments advocate for the use of Linear Probability Models (LPMs) rather than the generalised linear model counterpart (see, for example Section 3.4.1 of Angrist and Pischke (2009)). Columns (4) to (6) of Web Appendix Table B.5 report the results from estimating LKRI's probit regression that models whether a six was reported as a function of treatments via an LPM with HC2 heteroskedastic standard errors. Their results are robust to the alternative specification.

Table 3: Distribution of High vs Low Reported Die-Roll Outcomes between Treatments

Condition	Number of 1-3	Number of 4-6
Honesty Promoting	291 (36.56%)	505 (63.44%)
Dishonesty Promoting	176 (21.78%)	632 (78.22%)
No Advice	77 (36.15%)	136 (63.85%)

Notes: The Table reports the number (percentage) of subjects who report a die roll outcome between 1 and 3 (Number of 1 -3) and 4 to 6 (Number 4 - 6). Treatment conditions aggregated to the Control Condition (No Advice) and whether the advice was honesty promoting ot dishonest promoting. See Table 1 for definitions.

5.2 Proportions of die roll values

Our second robustness exercise further explores the distribution of die roll outcomes, expanding beyond using reported sixes an alternate outcome variable by LKHRI. In the case of a fair die, the proportions of each value of the die roll, i.e., 1 through 6, are equal be 1/6. This would imply that the cumulative proportions of 1-3 (lower values) appearing on the dice should be equal to the cumulative proportions of 4-6 (higher values) appearing on the dice. Since the main results of LKHRI document a difference in behaviour along the dishonesty-honesty dimension of advice giving, we focus on whether the proportion of higher vs lower outcomes differ along this dimension.

Table 3 reports the cumulative number of reported values for the three conditions. Chi-square tests comparing the Honesty Promoting and No Advice treatments indicate indicating that the two proportions are statistically not different ($\chi^2_1 = 0.000877$, $p = 0.9764$). In contrast, chi-square test comparing the Dishonesty Promoting and No Advice treatments provide strong evidence that the two proportions are different ($\chi^2_1 = 17.907$, $p < 0.001$). Similar results hold for comparing Dishonesty Promoting and Honest promoting treatments ($\chi^2_1 = 41.704$, $p < 0.001$).

We further explore the distribution of all reported die-roll outcomes by treatment. To do this we estimate six separate linear probability models where each potential outcome is used a dependent variable. Web Appendix Table C.3 reports results comparing opacity treatments to the baseline, equivalent to the specification of Col-

umn (1) in Web Appendix Table B.5. The results show an increase in reports of fives and sixes coming from a decrease in reports of ones and threes in the dishonesty promoting treatments. Web Appendix Table C.4 reports using all treatment conditions, mirroring the specification of Column (3) in Web Appendix Table B.5. These results corroborate those above and provide suggestive insights into subject behaviour differences across treatments.[5] Subjects in the dishonesty promoting advice treatment that is sourced from humans decrease their reports of rolling a three, and when the advice source is opaque, dishonesty promoting and sourced from humans increase their reports of a four.

6 Concluding Remarks

We conducted a computational reproducibility analysis of Leib et al. (2024) by recoding their analysis from the raw data. Our point estimates and standard errors align with the original article, suggesting that their empirical results are reproducible. We conduct further robustness checks by robustifying regression standard errors to be heteroskedasticity robust, use Bonferroni adjusted standard errors for t-tests and alternative regression specifications. LKHRI's results are robust to these changes.

We suggest several improvements for future replication repositories to adhere to best practice (Gentzkow and Shapiro 2014, Koren et al. 2022, Vilhuber et al. 2022). First, repositories should include a README file in the root directory that clearly explains the structure and organization of all materials. Second, detailed documentation for datasets should be provided, including thorough descriptions of all variables and their measurements. Third, repositories should maintain a clear separation between code, data, inputs, and outputs to enhance navigability and reproducibility. Finally, we encourage authors to adopt best practice by modularizing their analysis (rather than using single scripts) and thoroughly documenting their code with comments that explain not just what the code does, but why specific analytical choices were

[5] Differences discussed are significant at the 10% significance level.

made. We believe these practices would significantly enhance the transparency and reproducibility of empirical research.

References

Angrist, J. D. and Pischke, J.-S.: 2009, *Mostly harmless econometrics: An empiricist's companion*, Princeton university press.

Fischbacher, U. and Föllmi-Heusi, F.: 2013, Lies in disguise—an experimental study on cheating, *Journal of the European Economic Association* **11**(3), 525–547.

Gentzkow, M. and Shapiro, J. M.: 2014, Code and data for the social sciences: A practitioner's guide.

Koren, M., Connolly, M., Lull, J. and Vilhuber, L.: 2022, *Data and Code Availability Standard*.
URL: *https://doi.org/10.5281/zenodo.7436134*

Leib, M., Köbis, N., Rilke, R. M., Hagens, M. and Irlenbusch, B.: 2024, Corrupted by algorithms? how ai-generated and human-written advice shape (dis) honesty, *The Economic Journal* **134**(658), 766–784.

Vilhuber, L., Connolly, M., Koren, M., Llull, J. and Morrow, P.: 2022, *A template README for social science replication packages*.
URL: *https://doi.org/10.5281/zenodo.7293838*

Web Appendices

A Contents of the Replication Package

Table A.1: Replication Package Contents and Reproducibility

Replication Package Item	Fully	Partial	No
Raw data provided	✓		
Cleaning code provided	✓		
Analysis code provided	✓		
Provided code generates reported results	✓		
Reproducible from analysis data via recoding	✓		
Preregistration of experiment design	✓		
Preregistration of analysis plan			✓

B Tables Accompanying Computational Reproducibility Analysis

Table B.1: One Sample t-tests for over-reporting of die-roll outcomes

Treatment	Test Statistic
AI × Honest × Opaque	4.35 ***
AI × Honest × Transparent	5.38 ***
Human × Dishonest × Opaque	9.64 ***
AI × Dishonest × Opaque	11.97 ***
AI × Dishonest × Transparent	11.21 ***
Human × Honest × Transparent	3.43 ***
Human × Dishonest × Transparent	11.36 ***
Human × Honest × Opaque	3.97 ***
No Advice	4.55 ***

Notes: Table reports (two-sided) t-test results of whether the mean reported die-roll outcome was equal to that of a fair die (EV = 3.5). * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$ using Bonferroni-Holm adjusted p-values. Reproduces reported statistics of Leib et al. (2024) reported in the first sentence of Section 3 on page 774.

Table B.2: Replication of Pairwise t-statistics

	Opacity			
	AI Generated		Human Written	
	Honest	Dishonest	Honest	Dishonest
No Advice	-0.12	-4.33 ***	0.38	-3.88 ***
AI × Honest × Opaque	-	-3.99 ***	0.48	-3.61 **
AI × Dishonest × Opaque	-	-	4.71 ***	0.04
Human × Honest × Opaque	-	-	-	-4.22 ***
Human × Dishonest × Opaque	-	-	-	-
AI × Honest × Transparent	-	-	-	-
AI × Dishonest × Transparent	-	-	-	-
Human × Honest × Transparent	-	-	-	-
	Transparency			
	AI Generated		Human Written	
	Honest	Dishonest	Honest	Dishonest
No Advice	-0.57	-4.32 ***	0.72	-4.21 ***
AI × Honest × Opaque	-0.43	-4.00 ***	0.81	-3.88 ***
AI × Dishonest × Opaque	3.73 ***	-0.15	5.05 ***	0.04
Human × Honest × Opaque	-0.94	-4.68 ***	0.35	-4.57 ***
Human × Dishonest × Opaque	3.33 **	-0.18	4.54 ***	-0.01
AI × Honest × Transparent	-	-3.73 ***	1.29	-3.62 **
AI × Dishonest × Transparent	-	-	5.00 ***	0.19
Human × Honest × Transparent	-	-	-	-4.91 ***

Notes: Table reports (two-sided) t-test results of whether the mean reported die-roll outcome were equal between treatments. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$ using Bonferroni adjusted p-values. Reproduces reported statistics in Figure 1 in this manuscript. See Table 1 and notes therein for treatment definitions.

Table B.3: Replication of Table 1

	Dependent variable: reported die-roll outcome						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Estimates of LKRHI							
No advice	0.019						
	(0.150)						
Dishonesty-promoting	0.610***	0.590***	0.590***	0.396**	0.439**	0.436**	0.370*
	(0.145)	(0.148)	(0.145)	(0.143)	(0.144)	(0.145)	(0.152)
Human-written		−0.076	−0.076	−0.166	−0.127	−0.024	0.086
		(0.152)	(0.150)	(0.147)	(0.147)	(0.157)	(0.171)
Transparent			0.067	0.071	0.112	0.114	
			(0.150)	(0.147)	(0.148)	(0.148)	
Interactions							
Dishonesty-promoting × Human		0.070	0.070	0.104	0.033	−0.046	0.041
		(0.214)	(0.210)	(0.205)	(0.206)	(0.210)	(0.219)
Dishonesty-promoting × Transparent			−0.046	−0.031	−0.088	−0.067	
			(0.209)	(0.203)	(0.204)	(0.204)	
Human × Transparent			−0.120	−0.094	−0.146	−0.162	
			(0.211)	(0.205)	(0.206)	(0.206)	
Dishonesty-promoting × Human × Transparent			0.101	0.083	0.170	0.161	
			(0.297)	(0.290)	(0.290)	(0.290)	
Intercept	3.986***	4.005***	4.005***	3.350***	3.571***	3.362***	1.838*
	(0.104)	(0.108)	(0.106)	(0.145)	(0.195)	(0.511)	(0.756)
Additional Controls							
Norms	No	No	No	Yes	Yes	Yes	Yes
Demographics	No	No	No	No	Yes	Yes	Yes
Advice Readability	No	No	No	No	No	Yes	Yes
Guess Correctly	No	No	No	No	No	No	Yes
R2	0.035	0.042	0.044	0.096	0.101	0.105	0.105
Num.Obs.	634	803	1604	1604	1589	1589	794
Panel B: Our Estimates							
No advice	0.019						
	(0.150)						
Dishonesty-promoting	0.610***	0.590***	0.590***	0.396**	0.439**	0.436**	0.370*
	(0.145)	(0.148)	(0.145)	(0.143)	(0.144)	(0.145)	(0.152)
Human-written		−0.076	−0.076	−0.166	−0.127	−0.024	0.086
		(0.152)	(0.150)	(0.147)	(0.147)	(0.157)	(0.171)
Transparent			0.067	0.071	0.112	0.114	
			(0.150)	(0.147)	(0.148)	(0.148)	
Interactions							
Dishonesty-promoting × Human		0.070	0.070	0.104	0.033	−0.046	0.041
		(0.214)	(0.210)	(0.205)	(0.206)	(0.210)	(0.219)
Dishonesty-promoting × Transparent			−0.046	−0.031	−0.088	−0.067	
			(0.209)	(0.203)	(0.204)	(0.204)	
Human × Transparent			−0.120	−0.094	−0.146	−0.162	
			(0.211)	(0.205)	(0.206)	(0.206)	
Dishonesty-promoting × Human × Transparent			0.101	0.083	0.170	0.161	
			(0.297)	(0.290)	(0.290)	(0.290)	
Intercept	3.986***	4.005***	4.005***	3.350***	3.571***	3.362***	1.838*
	(0.104)	(0.108)	(0.106)	(0.145)	(0.195)	(0.511)	(0.756)
Additional Controls							
Norms	No	No	No	Yes	Yes	Yes	Yes
Demographics	No	No	No	No	Yes	Yes	Yes
Advice Readability	No	No	No	No	No	Yes	Yes
Guess Correctly	No	No	No	No	No	No	Yes
R2	0.035	0.042	0.044	0.096	0.101	0.105	0.105
Num.Obs.	634	803	1604	1604	1589	1589	794

Notes: Table reports coefficient estimates of a linear regression of reported die-roll outcome on treatment indicators and control variables. Panel A reports the results of Leib et al. (2024). Panel B reports our reproduction results. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$ using homoskedastic standard errors. Column (1) uses data from opacity, AI advice and No Advice treatments. Columns (2) uses data from opacity treatments. Columns (3) - (6) use data from all treatments except No Advice. Column (7) uses data from opacity treatments and no advice. See Table 1 and notes therein for treatment definitions.

Table B.4: Mechanism Regressions

	Dependent Variable:							
	Injunctive Norms		Descriptive Norms		Justifiability		Shared Responsibility	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: LKHRI Estimates.								
Honest	−7.942***	−10.453***	−9.278***	−9.200***	−12.508***	−13.102***		−4.408
	(2.160)	(3.085)	(1.654)	(2.362)	(2.101)	(2.998)		(3.700)
Human		0.785		3.442		3.282	−1.327	−4.280
		(3.062)		(2.344)		(2.976)	(2.588)	(3.672)
Interaction								
Honest × Human		4.817		−0.257		1.043		5.917
		(4.318)		(3.305)		(4.197)		(5.179)
Intercept	33.935***	33.540***	76.023***	74.293***	40.962***	39.313***	28.273***	30.455***
	(1.532)	(2.170)	(1.140)	(1.661)	(1.559)	(2.109)	(1.845)	(2.603)
R2	0.017	0.021	0.038	0.043	0.043	0.047	0.000	0.002
Num.Obs.	801	801	801	801	801	801	801	801
Panel B: Our Estimates.								
Honest	−7.942***	−10.453***	−9.278***	−9.200***	−12.508***	−13.102***		−4.408
	(2.160)	(3.085)	(1.654)	(2.362)	(2.101)	(2.998)		(3.700)
Human		0.785		3.442		3.282	−1.327	−4.280
		(3.062)		(2.344)		(2.976)	(2.588)	(3.672)
Interaction								
Honest × Human		4.817		−0.257		1.043		5.917
		(4.318)		(3.305)		(4.197)		(5.179)
Intercept	33.935***	33.540***	76.023***	74.293***	40.962***	39.313***	28.273***	30.455***
	(1.532)	(2.170)	(1.140)	(1.661)	(1.559)	(2.109)	(1.845)	(2.603)
R2	0.017	0.021	0.038	0.043	0.043	0.047	0.000	0.002
Num.Obs.	801	801	801	801	801	801	801	801

Notes: Table reports coefficient estimates of a linear regression of perceived norms on treatment indicators and control variables. Reproduces the unreported regression results of Leib et al. (2024) used in Section 3.2. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$ using homoskedastic standard errors. All columns use data from the opacity treatments. See Table 1 and notes therein for treatment definitions.

Table B.5: Replication of Reported Sixes

	Dependent variable: reported die-roll outcome of six					
	Probit Coefficients			OLS Coefficients		
	(1)	(2)	(3)	(4)	(5)	(6)
No advice	0.076			0.013		
	(0.242)			(0.041)		
Dishonesty-promoting	0.612***	0.536**	0.536**	0.118***	0.105**	0.105**
	(0.221)	(0.224)	(0.224)	(0.042)	(0.043)	(0.043)
Human-written		−0.162	−0.162		−0.026	−0.026
		(0.250)	(0.250)		(0.041)	(0.041)
Transparent			−0.111			−0.018
			(0.249)			(0.041)
Interactions						
Dishonesty-promoting × Human		0.445	0.445		0.091	0.091
		(0.325)	(0.325)		(0.063)	(0.063)
Dishonesty-promoting × Transparent			0.284			0.058
			(0.322)			(0.062)
Human × Transparent			0.070			0.012
			(0.355)			(0.057)
Dishonesty-promoting × Human × Transparent			−0.524			−0.115
			(0.462)			(0.088)
Intercept	−1.346***	−1.269***	−1.269***	0.207***	0.219***	0.219***
	(0.169)	(0.173)	(0.173)	(0.028)	(0.030)	(0.030)
R2				0.015	0.030	0.030
Num.Obs.	634	803	1604	634	803	1604

Notes: Table reports coefficient estimates of regressions using an binary indicator that takes the value 1 when reported die-roll outcome is a six on treatment indicators and control variables. Columns (1) - (3) report our reproduction of the un-reported probit estimation results used by Leib et al. (2024) discussed in Section 3.1.1. Columns (4) - (6) use Linear Probability Model equivalents of the regressions run in Columns (1) - (3). * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Columns (4) - (6) report HC2 heteroskedasticity robust standard errors. Columns (1) and (4) uses data from opacity, AI advice and No Advice treatments. Columns (2) and (5) uses data from opacity treatments. Columns (3) and (6) use data from all treatments except No Advice. See Table 1 and notes therein for treatment definitions.

C Tables Accompanying Computational Reproducibility Analysis

Table C.1: Replication of Table 1 with HC2 Standard Errors

	Dependent variable: reported die-roll outcome						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
No advice	0.019						
	(0.158)						
Dishonesty-promoting	0.610***	0.590***	0.590***	0.396**	0.439**	0.436**	0.370*
	(0.141)	(0.148)	(0.148)	(0.147)	(0.147)	(0.150)	(0.155)
Human-written		−0.076	−0.076	−0.166	−0.127	−0.024	0.086
		(0.159)	(0.159)	(0.160)	(0.161)	(0.170)	(0.180)
Transparent			0.067	0.071	0.112	0.114	
			(0.157)	(0.156)	(0.156)	(0.156)	
Interactions							
Dishonesty-promoting × Human		0.070	0.070	0.104	0.033	−0.046	0.041
		(0.215)	(0.215)	(0.210)	(0.211)	(0.217)	(0.224)
Dishonesty-promoting × Transparent			−0.046	−0.031	−0.088	−0.067	
			(0.208)	(0.202)	(0.202)	(0.202)	
Human × Transparent			−0.120	−0.094	−0.146	−0.162	
			(0.220)	(0.220)	(0.221)	(0.221)	
Dishonesty-promoting × Human × Transparent			0.101	0.083	0.170	0.161	
			(0.298)	(0.290)	(0.291)	(0.290)	
Intercept	3.986***	4.005***	4.005***	3.350***	3.571***	3.362***	1.838*
	(0.107)	(0.116)	(0.116)	(0.157)	(0.207)	(0.518)	(0.785)
Additional Controls							
Norms	No	No	No	Yes	Yes	Yes	Yes
Demographics	No	No	No	No	Yes	Yes	Yes
Advice Readability	No	No	No	No	No	Yes	Yes
Guess Correctly	No	No	No	No	No	No	Yes
R2	0.035	0.042	0.044	0.096	0.101	0.105	0.105
Num.Obs.	634	803	1604	1604	1589	1589	794

Notes: Table reports coefficient estimates of a linear regression of perceived norms on treatment indicators and control variables. Reproduces Table 1 of Leib et al. (2024) using HC2 standard errors. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Column (1) uses data from opacity, AI advice and No Advice treatments. Columns (2) uses data from opacity treatments. Columns (3) - (6) use data from all treatments except No Advice. Column (7) uses data from opacity treatments and no advice. See Table 1 and notes therein for treatment definitions.

Table C.2: Replication of Table 1 with HC3 Standard Errors

	Dependent variable: reported die-roll outcome						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
No advice	0.019						
	(0.158)						
Dishonesty-promoting	0.610***	0.590***	0.590***	0.396**	0.439**	0.436**	0.370*
	(0.141)	(0.148)	(0.148)	(0.148)	(0.148)	(0.150)	(0.157)
Human-written		−0.076	−0.076	−0.166	−0.127	−0.024	0.086
		(0.159)	(0.159)	(0.161)	(0.162)	(0.171)	(0.182)
Transparent			0.067	0.071	0.112	0.114	
			(0.158)	(0.156)	(0.157)	(0.157)	
Interactions							
Dishonesty-promoting × Human		0.070	0.070	0.104	0.033	−0.046	0.041
		(0.216)	(0.216)	(0.211)	(0.212)	(0.218)	(0.226)
Dishonesty-promoting × Transparent			−0.046	−0.031	−0.088	−0.067	
			(0.208)	(0.203)	(0.203)	(0.203)	
Human × Transparent			−0.120	−0.094	−0.146	−0.162	
			(0.221)	(0.221)	(0.222)	(0.222)	
Dishonesty-promoting × Human × Transparent			0.101	0.083	0.170	0.161	
			(0.299)	(0.291)	(0.292)	(0.292)	
Intercept	3.986***	4.005***	4.005***	3.350***	3.571***	3.362***	1.838*
	(0.107)	(0.116)	(0.116)	(0.158)	(0.208)	(0.521)	(0.793)
Additional Controls							
Norms	No	No	No	Yes	Yes	Yes	Yes
Demographics	No	No	No	No	Yes	Yes	Yes
Advice Readability	No	No	No	No	No	Yes	Yes
Guess Correctly	No	No	No	No	No	No	Yes
R2	0.035	0.042	0.044	0.096	0.101	0.105	0.105
Num.Obs.	634	803	1604	1604	1589	1589	794

Notes: Table reports coefficient estimates of a linear regression of perceived norms on treatment indicators and control variables. Reproduces Table 1 of Leib et al. (2024) using HC3 standard errors. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Column (1) uses data from opacity, AI advice and No Advice treatments. Columns (2) uses data from opacity treatments. Columns (3) - (6) use data from all treatments except No Advice. Column (7) uses data from opacity treatments and no advice. See Table 1 and notes therein for treatment definitions.

Table C.3: Indicator for each Die Roll Outcome by Treatment, Simple

	Dependent variable: reported die-roll outcome					
	Outcome = 1	Outcome = 2	Outcome = 3	Outcome = 4	Outcome = 5	Outcome = 6
No advice	−0.002	0.049	−0.051	−0.042	0.034	0.013
	(0.029)	(0.032)	(0.036)	(0.040)	(0.042)	(0.041)
Dishonesty-promoting	−0.067***	−0.014	−0.072**	−0.038	0.073*	0.118***
	(0.023)	(0.026)	(0.033)	(0.038)	(0.041)	(0.042)
Intercept	0.094***	0.089***	0.178***	0.221***	0.211***	0.207***
	(0.020)	(0.020)	(0.026)	(0.028)	(0.028)	(0.028)
R2	0.016	0.008	0.008	0.002	0.005	0.015
Num.Obs.	634	634	634	634	634	634

Notes: Table reports coefficient estimates of linear regressions using an binary indicator that takes the value 1 when reported die-roll outcome is takes a given value on treatment indicators and control variables. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$ using HC2 heteroskedasticity robust standard errors. All columns use data data from opacity, AI advice and No Advice treatments. See Table 1 and notes therein for treatment definitions.

Table C.4: Indicator for each Die Roll Outcome by Treatment, Treatment Interactions

	Dependent variable: reported die-roll outcome					
	Outcome = 1	Outcome = 2	Outcome = 3	Outcome = 4	Outcome = 5	Outcome = 6
Dishonesty-promoting	−0.065***	−0.062**	−0.021	0.004	0.040	0.105**
	(0.023)	(0.030)	(0.032)	(0.038)	(0.043)	(0.043)
Human-written	−0.021	−0.011	0.055	0.055	−0.052	−0.026
	(0.028)	(0.034)	(0.037)	(0.041)	(0.042)	(0.041)
Transparent	−0.040	−0.004	0.017	0.053	−0.008	−0.018
	(0.026)	(0.035)	(0.035)	(0.041)	(0.043)	(0.041)
Dishonesty-promoting × Human	0.043	0.033	−0.081*	−0.086	0.000	0.091
	(0.034)	(0.044)	(0.047)	(0.055)	(0.060)	(0.063)
Interactions						
Dishonesty-promoting × Transparent	0.039	−0.001	0.008	−0.064	−0.039	0.058
	(0.030)	(0.043)	(0.047)	(0.055)	(0.061)	(0.062)
Human × Transparent	0.065*	0.006	−0.032	−0.081	0.030	0.012
	(0.038)	(0.048)	(0.051)	(0.058)	(0.059)	(0.057)
Dishonesty-promoting × Human × Transparent	−0.072	−0.068	0.066	0.145*	0.044	−0.115
	(0.046)	(0.060)	(0.068)	(0.079)	(0.085)	(0.088)
Intercept	0.092***	0.138***	0.128***	0.179***	0.245***	0.219***
	(0.021)	(0.025)	(0.024)	(0.027)	(0.031)	(0.030)
R2	0.013	0.015	0.008	0.005	0.004	0.030
Num.Obs.	1604	1604	1604	1604	1604	1604

Notes: Table reports coefficient estimates of linear regressions using an binary indicator that takes the value 1 when reported die-roll outcome is takes a given value on treatment indicators and control variables. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$ using HC2 heteroskedasticity robust standard errors. Each column uses data from all treatment conditions except No Advice used in the analysis. See Table 1 and notes therein for treatment definitions.