Negative Perceptions of Outsourcing to Artificial Intelligence

Scott Claessens, Pierce Veitch, and Jim Everett

School of Psychology, University of Kent

Author Note

Scott Claessens https://orcid.org/0000-0002-3562-6981

Pierce Veitch https://orcid.org/0009-0005-3364-7470

Jim Everett https://orcid.org/0000-0003-2801-5426

Correspondence concerning this article should be addressed to Jim Everett, School of Psychology, University of Kent, Keynes College, Canterbury CT2 7NP, UK, Email: j.a.c.everett@kent.ac.uk

Abstract

xxxx

Keywords: keyword1, keyword2, keyword3

Negative Perceptions of Outsourcing to Artificial Intelligence

Study 1

Study 2

Study 3

Study 4

Study 5

Study 6

General Discussion

Supplementary Materials

Negative Perceptions of Outsourcing to Artificial Intelligence

Scott Claessens, Pierce Veitch, and Jim Everett

School of Psychology, University of Kent

Table of contents

Pilot Study 1 **3**

Methods . . . 3

Participants . . . 3

Procedure . . . 3

Statistical analysis . . . 3

Results . . . 4

Pilot Study 2 **5**

Methods . . . 5

Participants . . . 5

Design . . . 5

Procedure . . . 5

Pre-registration . . . 6

Statistical analysis . . . 6

Results . . . 6

Methods for Text Analysis in Study 5 **8**

Supplementary Figures **10**

Supplementary Tables **23**

Supplementary References **25**

Pilot Study 1

Methods

Participants

We recruited a convenience sample of 200 participants from the United Kingdom through Prolific. After excluding participants who failed our pre-treatment attention check, we were left with a final sample of 186 participants (118 female; 67 male; 1 non-binary / third gender; 0 undisclosed gender; mean age = 38.99 years).

Procedure

We presented participants with six different tasks "that people might perform in their daily lives". The six tasks were randomly drawn from a larger set of 20 tasks (see Supplementary Table 1 for the full list of tasks). For each task, we asked participants the following questions on 7-point Likert scales:

- Is this a social task?
- Does this task require social skills?
- Does this task impact other people?
- How important are the consequences of this task?
- How important is it that effort goes into this task?
- How important is it that others see the effort that goes into this task?

Statistical analysis

We fitted a Bayesian multivariate multilevel cumulative-link ordinal model to the data using the *brms* R package. We modelled each task evaluation as a separate response variable and included correlated varying intercepts for participants and tasks. We used regularising priors for all parameters to impose conservatism on parameter estimates (see Supplementary Materials for full model specification). The model converged normally ($\hat{R} \leq 1.01$).

Results

We found that participants' responses to all six questions tended to be positively correlated. For example, tasks rated as more social were also rated as requiring more social skills (see Supplementary Figure 1). Estimated averages and rankings for the 20 tasks across each of the questions can be found in Supplementary Figures 2 – 7.

Pilot Study 2

Methods

Participants

We conducted a power simulation to determine our target sample size. The simulation suggested that a sample size of 150 participants per condition (overall $n = 450$ for three conditions) would be required to detect a small difference between conditions (Cohen's $d \approx 0.20$) with above 80% power.

We recruited a convenience sample of 500 participants from the United Kingdom through Prolific. After excluding participants who failed our pre-treatment attention check, we were left with a final sample of 466 participants (292 female; 169 male; 4 non-binary / third gender; 1 undisclosed gender; mean age = 42.32 years). 73% of these participants reported having used ChatGPT before (see Supplementary Figure 8).

Design

We randomly allocated participants into one of three conditions in a between-subjects design: (*i*) the control condition, (*ii*) the AI outsourcing condition, or (*iii*) the human outsourcing condition. These conditions determined how scenarios were presented to participants.

Procedure

We presented participants with six scenarios (see Supplementary Figure 9 for examples). Each scenario described a person completing a task, such as writing computer code or writing a love letter. The six tasks were randomly drawn from a larger set of 20 tasks (see Supplementary Table 1 for the full list of tasks). For each scenario, we told participants:

- *Control condition*: "In order to complete this task, [the person] works on it by themselves from start to finish."
- *AI outsourcing condition*: "In order to complete this task, [the person] gets the AI tool ChatGPT to do it for them."

- *Human outsourcing condition*: "In order to complete this task, [the person] gets someone else to do it for them."

We then asked participants how well each of the following words described the person in the scenario: competent, warm, moral, lazy, and trustworthy. Participants answered these questions on 7-point Likert scales, ranging from "does not describe [the person] well" to "describes [the person] extremely well".

After the six scenarios, we asked participants several questions about the AI tool ChatGPT, including their familiarity with ChatGPT, whether they had used ChatGPT before, how frequently they used ChatGPT, and how trustworthy they thought ChatGPT was (see Supplementary Figure 8).

Pre-registration

We pre-registered the study on the Open Science Framework (https://osf.io/khr42).

Statistical analysis

We fitted Bayesian multivariate multilevel cumulative-link ordinal models to the data using the *brms* R package. We modelled each character evaluation – competence, warmth, morality, laziness, and trustworthiness – as a separate response variable and included fixed effects for conditions, varying intercepts for participants, and varying intercepts and slopes for tasks. We used regularising priors for all parameters to impose conservatism on parameter estimates (see Supplementary Materials for full model specifications). All models converged normally ($\hat{R} \leq 1.01$).

Results

We found that people who outsourced tasks to AI or other humans were perceived more negatively than people who completed the tasks themselves (Supplementary Figure 10). In particular, people who outsourced were perceived as lazier and less competent, with smaller yet detectable differences for perceptions of warmth, morality, and trustworthiness (Supplementary Table 2). Across all measures, outsourcing to other humans was perceived more negatively than

outsourcing to AI.

We found that the effects of outsourcing varied across the different tasks, especially for perceptions of warmth and morality (Supplementary Figure 11). For example, people were perceived as less warm if they outsourced writing a love letter, but not if they outsourced writing computer code. Similarly, people were perceived as less moral if they outsourced writing an apology letter to a friend, but not if they outsourced writing a dinner recipe. By contrast, the effects of outsourcing on competence, laziness, and trustworthiness were more consistent across tasks.

To determine the factors that predict variation across tasks, we incorporated ratings of tasks from the first pilot study. Participants were asked to rate the 20 tasks on several features: whether the task is social, requires social skills, impacts others, has important consequences, and requires effort. All of these features predicted stronger causal effects of outsourcing compared to control (Supplementary Figures 12 and 13). In other words, outsourcing to AI or other humans is perceived more negatively for tasks that have these features, compared to tasks without these features.

Methods for Text Analysis in Study 5

To generate frequency lists for each experimental condition in Study 5, we created three documents containing the raw text submissions to the open-ended question "In your own words, describe how you feel about Adam and why". Each raw text submission was paired with a numbered text ID column. The number of submissions was roughly equivalent across conditions: the control condition (N = 196), the tool outsourcing condition (N = 215), and the full outsourcing condition (N = 202).

All text processing was conducted using the Basic Unit-Transposable Text Experimentation Resource (BUTTER; Version 0.9.4.1; Boyd, 2019). To prepare the data, each CSV file was converted into a folder containing individual text files – one per submission – using two plugins: *Read Text from CSV* (Version 1.0.2) and *Save .txt Files to Folder* (Version 1.0.6). The settings for *Read Text from CSV* were as follows: file encoding = UTF-8, row identifier = ID, text column = Text, CSV delimiter = , and CSV quote = ".

To generate frequency lists, we first loaded the .txt files using the *Load .txt Files from Folder* plugin (Version 1.0.4). Tokenization was performed using the Twitter-Aware Tokenizer (Version 1.0.2), with the options *convert text to lowercase* and *reduce elongation* enabled to minimize superficial variation in tokens. We removed filler and function words using the *Remove Stop Words* plugin (Version 1.0.31), applying the default English stop word list.

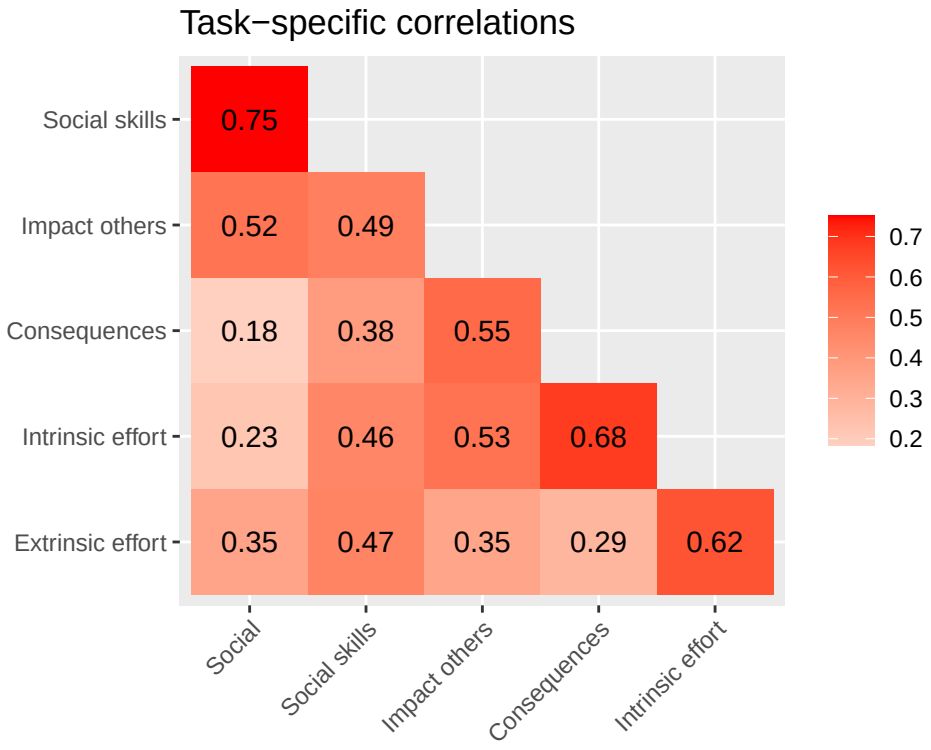
Frequency lists were created with the *Frequency List* plugin (Version 1.0.11). Settings included: unigram analysis (N = 1), omission of n-grams with frequency < 5, exclusion of n-grams appearing in fewer than 0.1% of documents, filtering collocates by Normalized Pointwise Mutual Information (NPMI), and removal of collocates with metric values < 0.5. Outputs were saved using the *Save Output to CSV* plugin (Version 1.0.5). This process was repeated separately for each condition folder.

For cross-condition comparison, we used the *Compare Frequencies* plugin (Version 1.1.02), retaining most default settings. The only modification was disabling the *Skip comparisons with 0 frequency values* option. This plugin calculates a range of comparative

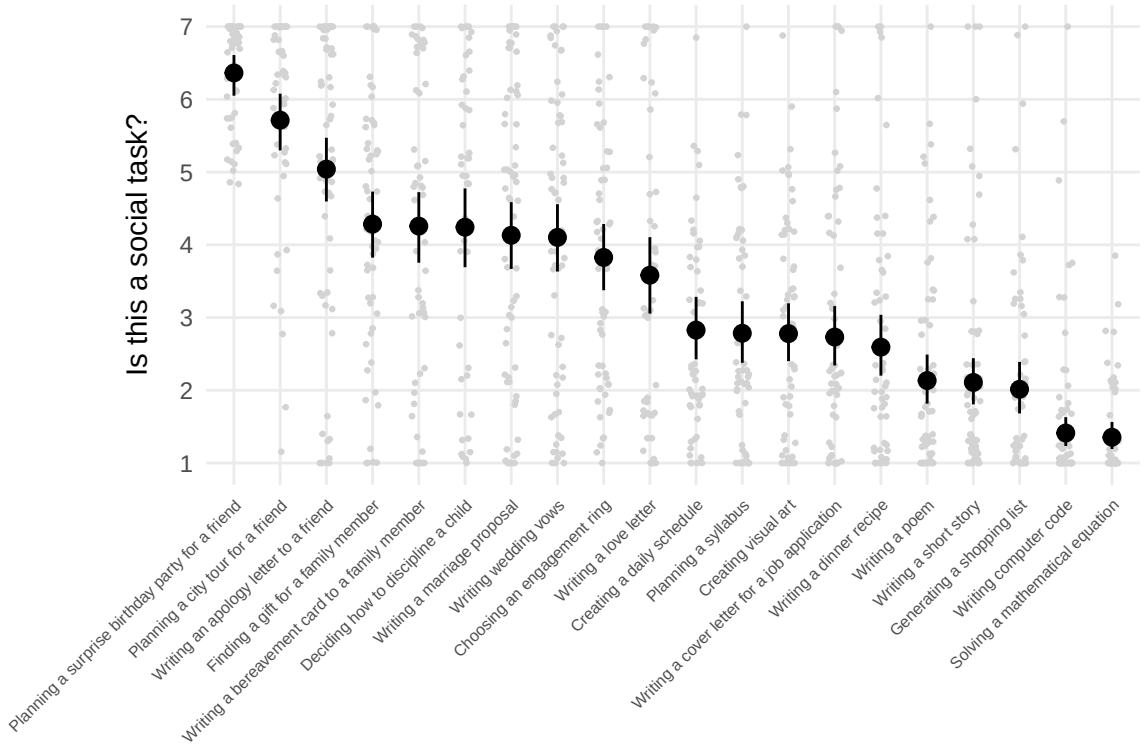
metrics, including log likelihood (LL), %DIFF, Bayes Information Criterion (BIC), relative risk (RRisk), log ratio, and odds ratio.

Following previous work (e.g., Rayson & Garside, 2000; Gregson et al., 2022), we interpret %DIFF as an indicator of effect size and direction. Frequentist statistical significance was determined using log likelihood values, with the following thresholds: LL ≥ 3.84 ($p < .05$), LL ≥ 6.63 ($p < .01$), LL ≥ 10.83 ($p < .001$), and LL ≥ 15.13 ($p < .0001$).

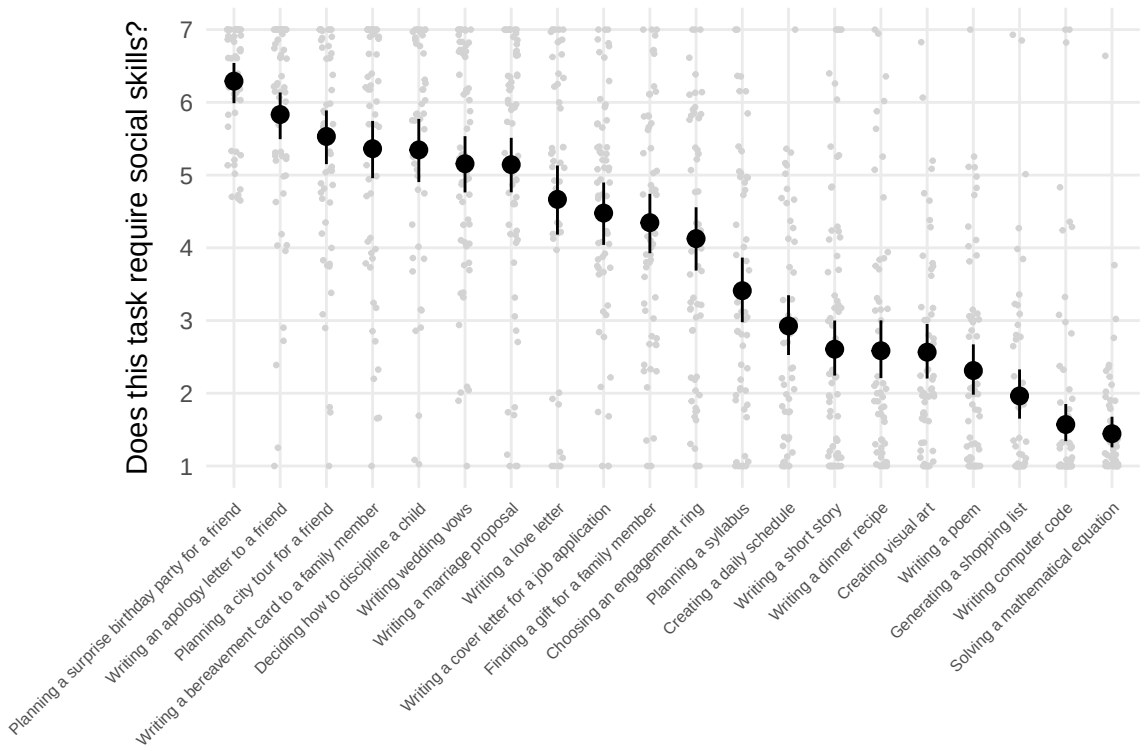
Supplementary Figures



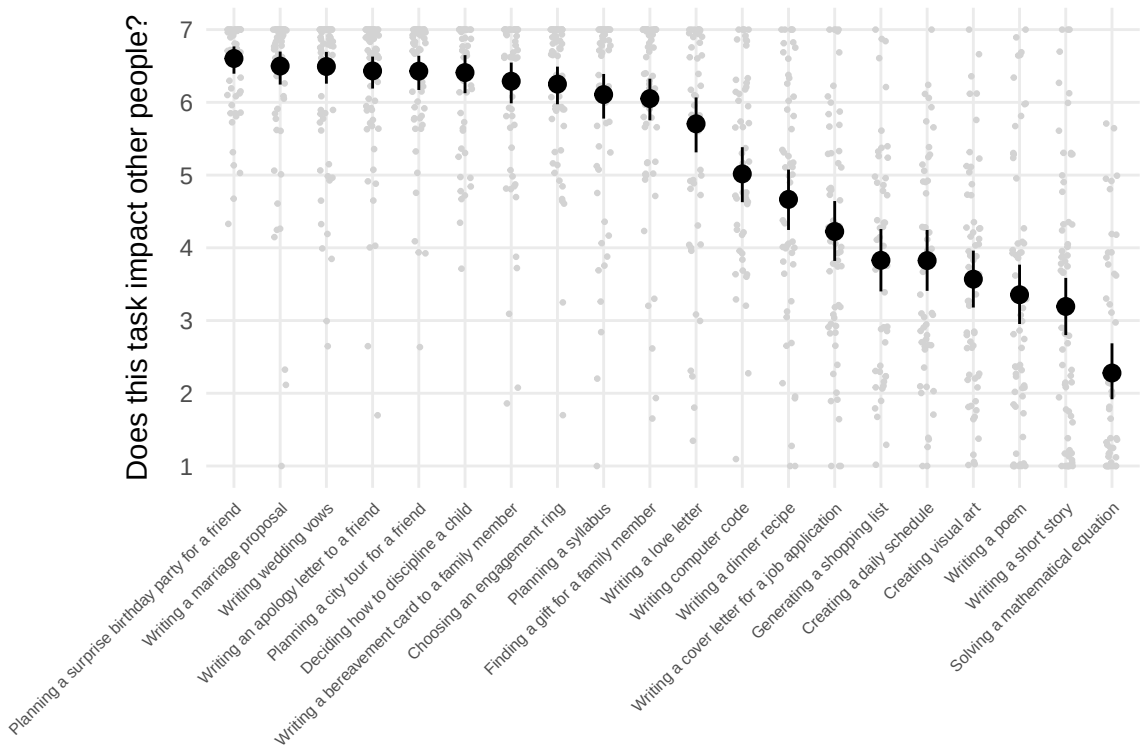
Supplementary Figure 1: Model-estimated task-specific correlations between all six questions. Values are posterior median correlations. A positive correlation indicates that tasks that are rated highly on one question tend to be rated highly on another question.



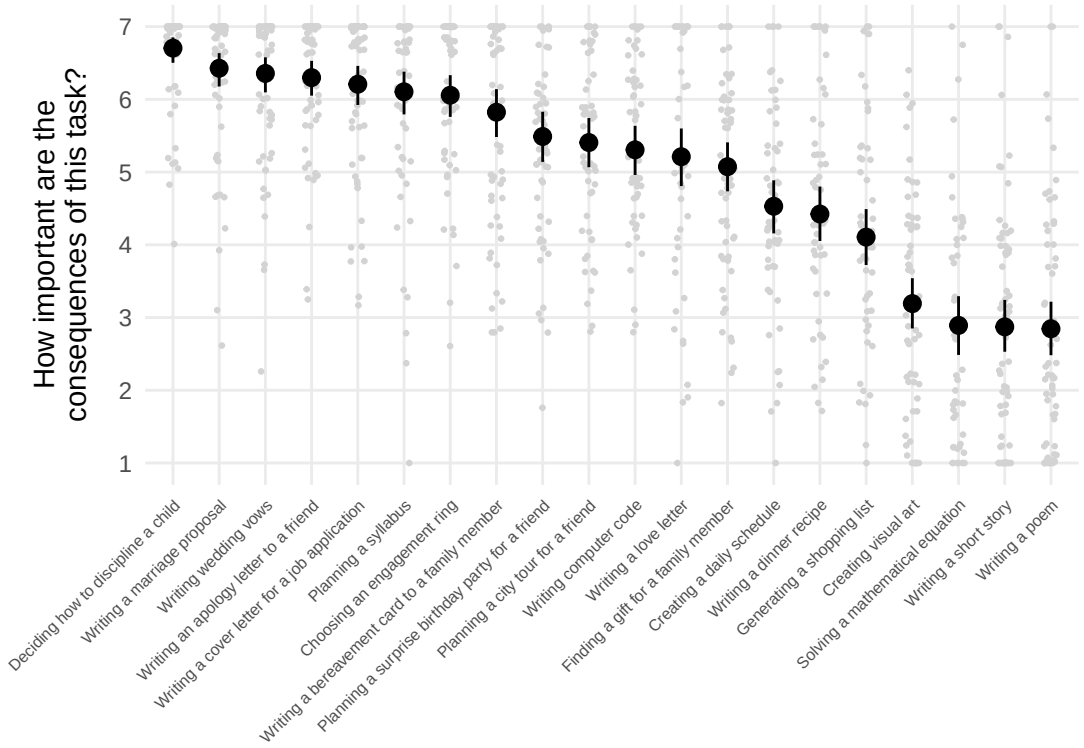
Supplementary Figure 2: Model-estimated means for the question “Is this a social task?” across all 20 tasks. Grey points represent participant responses to the question, jittered for easier viewing. Black points are estimated means from the fitted model, pooling over participants. Black points and line ranges represent posterior medians and 95% credible intervals, respectively.



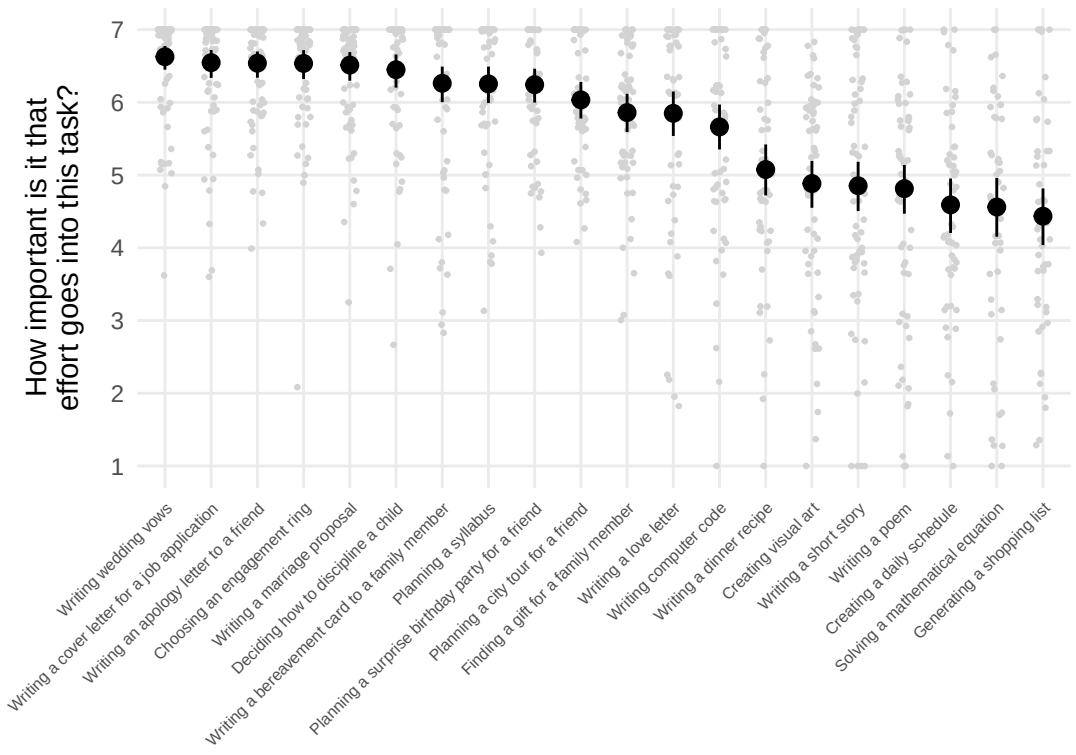
Supplementary Figure 3: Model-estimated means for the question "Does this task require social skills?" across all 20 tasks. Grey points represent participant responses to the question, jittered for easier viewing. Black points are estimated means from the fitted model, pooling over participants. Black points and line ranges represent posterior medians and 95% credible intervals, respectively.



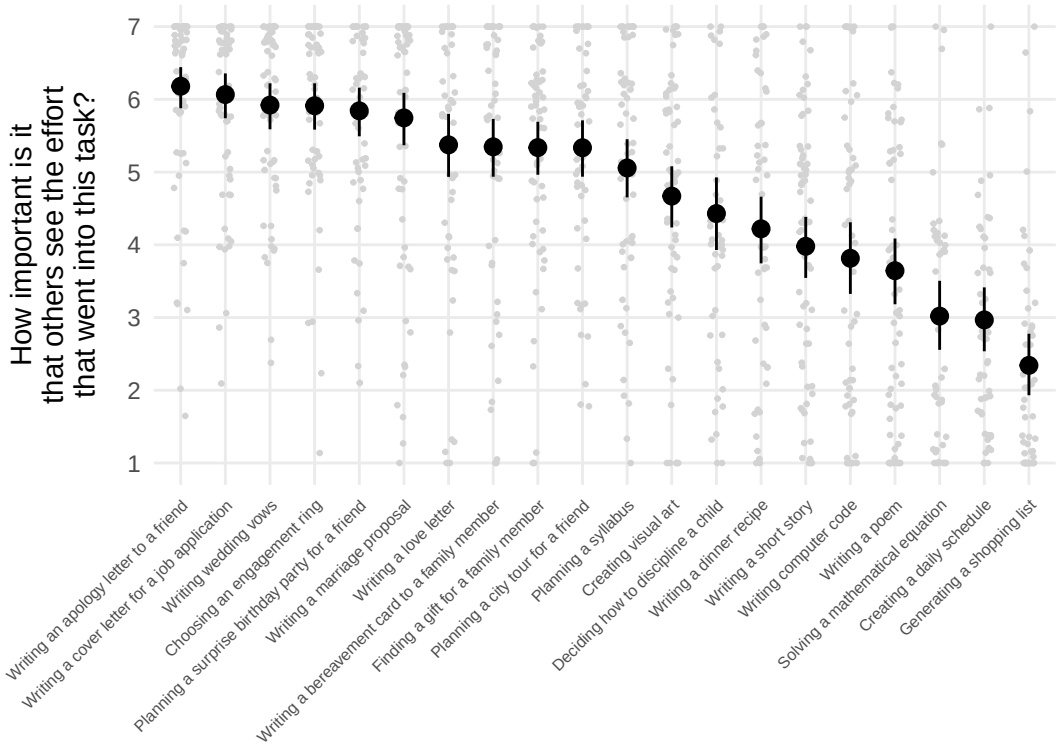
Supplementary Figure 4: Model-estimated means for the question “Does this task impact other people?” across all 20 tasks. Grey points represent participant responses to the question, jittered for easier viewing. Black points are estimated means from the fitted model, pooling over participants. Black points and line ranges represent posterior medians and 95% credible intervals, respectively.



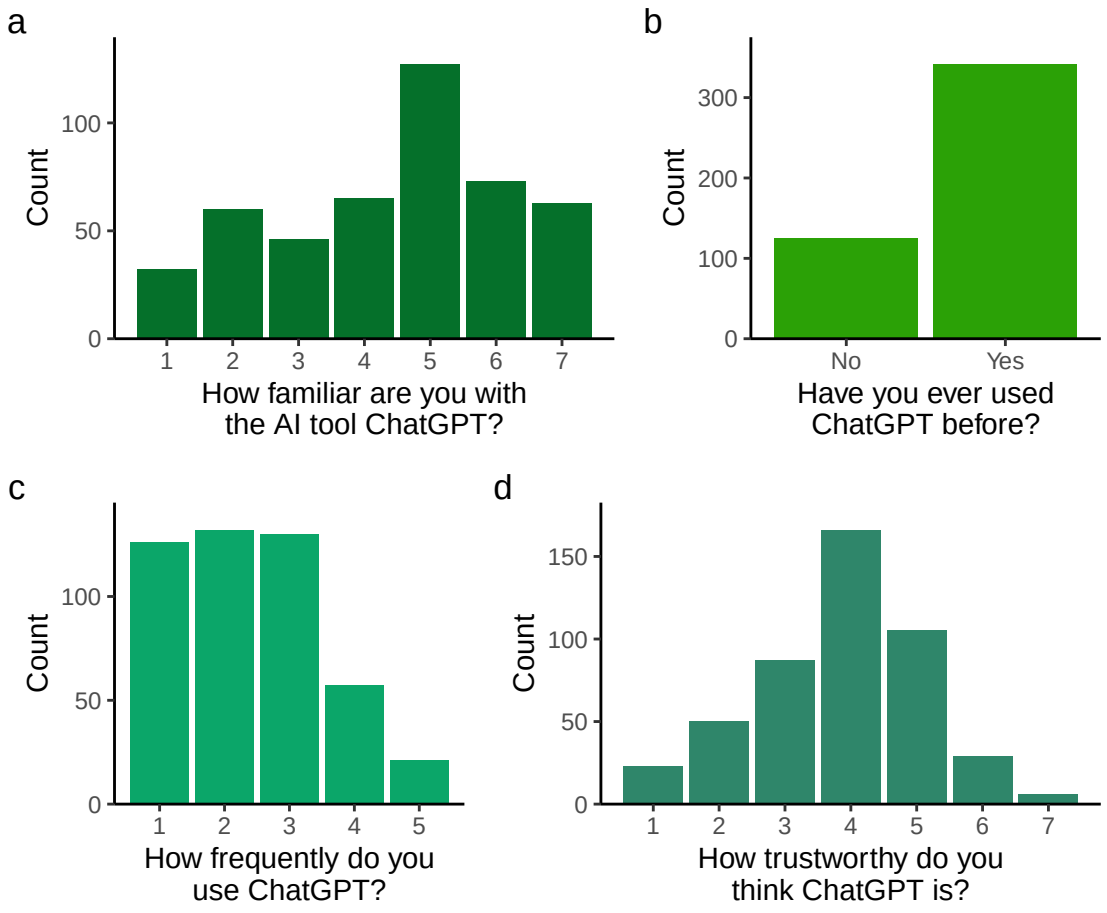
Supplementary Figure 5: Model-estimated means for the question “How important are the consequences of this task?” across all 20 tasks. Grey points represent participant responses to the question, jittered for easier viewing. Black points are estimated means from the fitted model, pooling over participants. Black points and line ranges represent posterior medians and 95% credible intervals, respectively.



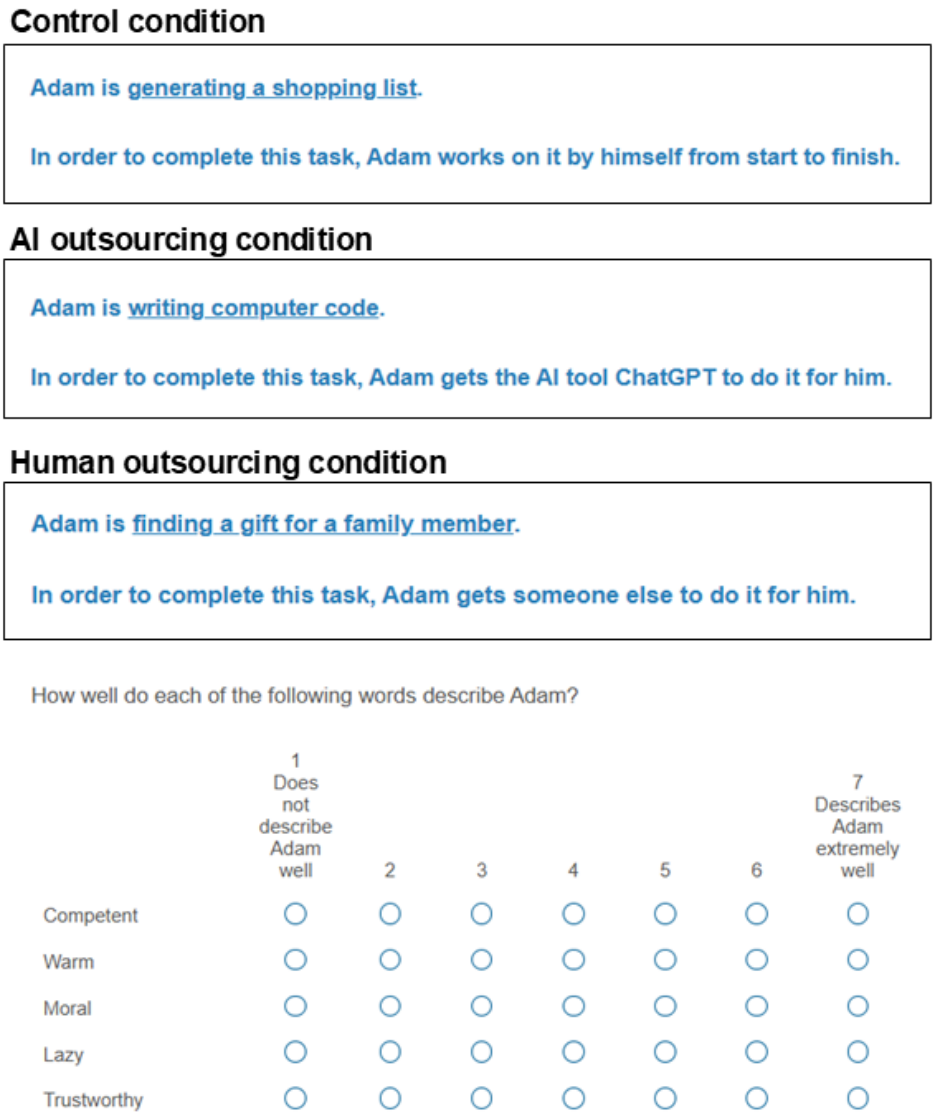
Supplementary Figure 6: Model-estimated means for the question "How important is it that effort goes into this task?" across all 20 tasks. Grey points represent participant responses to the question, jittered for easier viewing. Black points are estimated means from the fitted model, pooling over participants. Black points and line ranges represent posterior medians and 95% credible intervals, respectively.



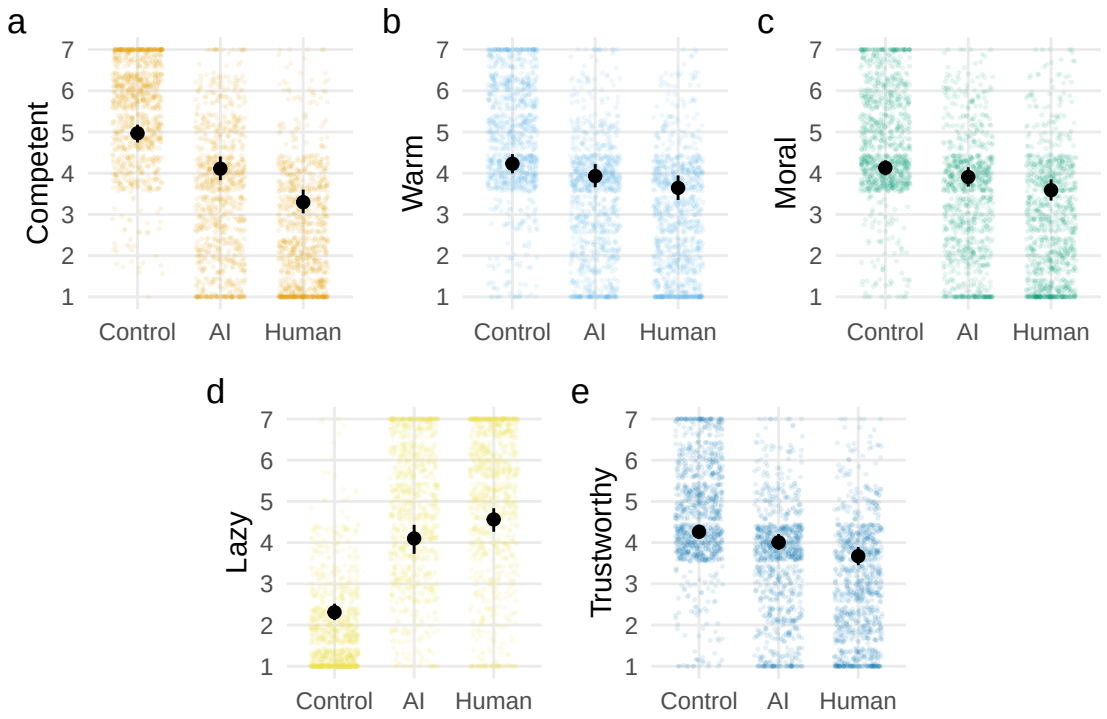
Supplementary Figure 7: Model-estimated means for the question "How important is it that others see the effort that goes into this task?" across all 20 tasks. Grey points represent participant responses to the question, jittered for easier viewing. Black points are estimated means from the fitted model, pooling over participants. Black points and line ranges represent posterior medians and 95% credible intervals, respectively.



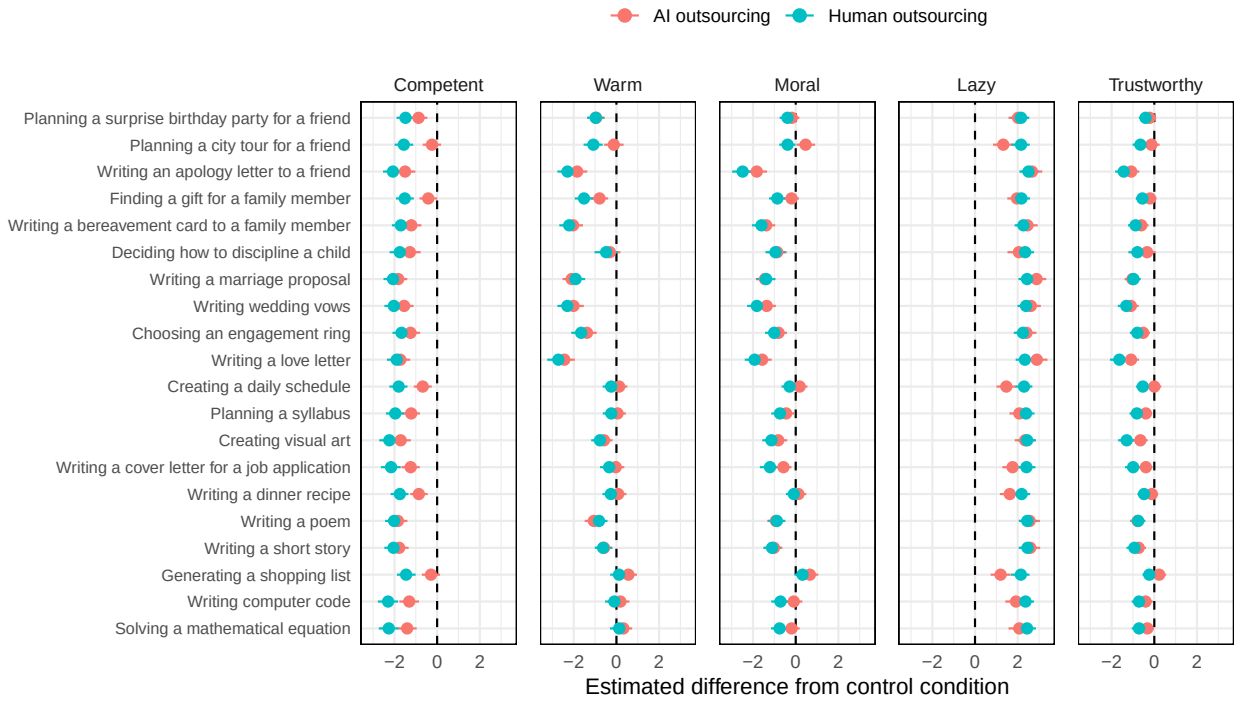
Supplementary Figure 8: Responses to the questions about ChatGPT in the second pilot study.



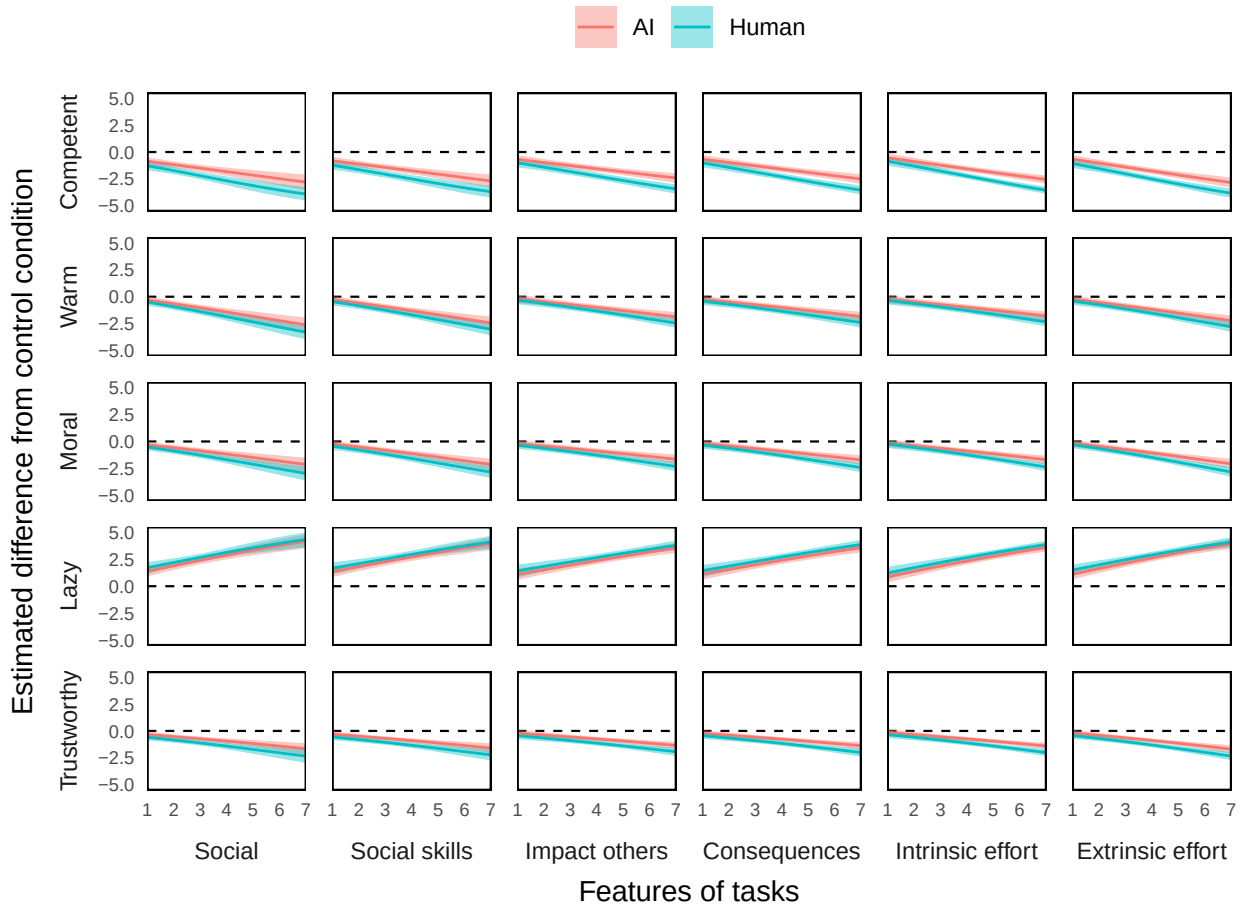
Supplementary Figure 9: Examples of the scenarios presented to participants in the second pilot study.



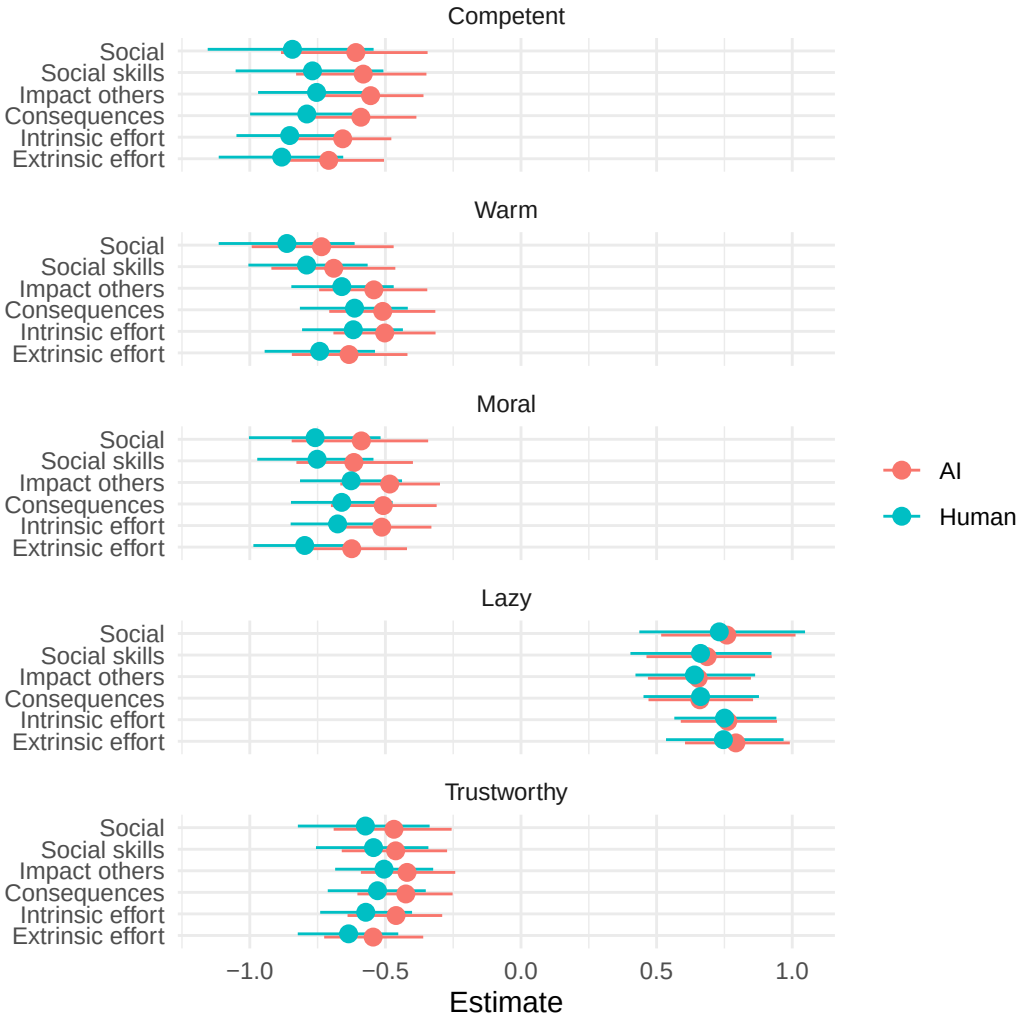
Supplementary Figure 10: Character evaluations in the second pilot study. Participants in the control condition, the AI outsourcing condition, and the human outsourcing condition evaluated people in the scenarios on (a) competence, (b) warmth, (c) morality, (d) laziness, and (e) trustworthiness. Coloured points represent participant responses to the questions, jittered for easier viewing. Black points are estimated marginal means from the fitted model, pooling over participants and tasks. Black points and line ranges represent posterior medians and 95% credible intervals, respectively.



Supplementary Figure 11: Variation in the effects of outsourcing across tasks in the second pilot study. Tasks are ordered from most social (top) to least social (bottom) according to ratings from the first pilot study. Point ranges are differences in marginal means on a 7-point Likert scale for the AI outsourcing condition (red) and the human outsourcing condition (blue) compared to the control condition. Points and ranges represent posterior medians and 95% credible intervals, respectively.



Supplementary Figure 12: The impact of task-specific features (e.g., being a social task) on the causal effects of outsourcing to AI (red) and humans (blue) compared to the control condition. The y-axis reflects the estimated differences between the experimental conditions and the control condition (dashed line) on a 7-point Likert scale. Lines and shaded areas represent posterior medians and 95% credible intervals, respectively. The patterns indicate, for example, more negative effects of outsourcing on character evaluations for tasks that are rated as more social.



Supplementary Figure 13: Interaction parameters from models including task-specific features as moderators of the causal effects of AI outsourcing (red) and human outsourcing (blue) compared to the control condition. Points and line ranges represent posterior medians and 95% credible intervals, respectively.

Supplementary Tables

Supplementary Table 1: Tasks included in the studies.

Task	Pilot Study 1	Pilot Study 2	Study 1	Study 2	Study 4
Writing wedding vows	✓	✓	✓	✓	✓
Writing a love letter	✓	✓	✓	✓	✓
Writing a marriage proposal	✓	✓	✓	✓	
Choosing an engagement ring	✓	✓	✓		
Finding a gift for a family member	✓	✓	✓		
Deciding how to discipline a child	✓	✓	✓		
Writing a bereavement card to a family member	✓	✓	✓	✓	✓
Writing an apology letter to a friend	✓	✓	✓	✓	✓
Planning a city tour for a friend	✓	✓	✓	✓	
Planning a surprise birthday party for a friend	✓	✓	✓	✓	
Writing a cover letter for a job application	✓	✓	✓	✓	✓
Writing computer code	✓	✓	✓	✓	✓
Solving a mathematical equation	✓	✓	✓	✓	✓
Planning a syllabus	✓	✓	✓	✓	✓
Writing a short story	✓	✓	✓	✓	
Writing a poem	✓	✓	✓	✓	
Creating visual art	✓	✓	✓		
Creating a daily schedule	✓	✓	✓	✓	
Generating a shopping list	✓	✓	✓	✓	
Writing a dinner recipe	✓	✓	✓	✓	

Supplementary Table 2: Pairwise contrasts in the second pilot study. Numbers reflect differences in marginal means on a 7-point Likert scale, pooling over participants and tasks. Main numbers are posterior medians, numbers in the square brackets are 95% credible intervals.

	Response				
	Competent	Warm	Moral	Lazy	Trustworthy
AI - Control	-0.86 [-1.16 -0.55]	-0.30 [-0.57 -0.01]	-0.23 [-0.47 0.02]	1.80 [1.42 2.14]	-0.26 [-0.44 -0.05]
Human - Control	-1.68 [-1.98 -1.35]	-0.59 [-0.87 -0.28]	-0.54 [-0.80 -0.28]	2.26 [1.90 2.58]	-0.59 [-0.81 -0.39]
Human - AI	-0.81 [-1.15 -0.46]	-0.29 [-0.61 0.04]	-0.32 [-0.63 -0.02]	0.46 [0.04 0.90]	-0.33 [-0.58 -0.10]

Supplementary References

Boyd, R. L. (2019). BUTTER: Basic unit-transposable text experimentation resource. Available from https://www.butter.tools/

Gregson, R., Piazza, J., & Boyd, R. L. (2022). 'Against the cult of veganism': Unpacking the social psychology and ideology of anti-vegans. *Appetite*, *178*, 106143. doi:10.1016/j.appet.2022.106143

Rayson, P., & Garside, R. (2000). Comparing corpora using frequency profiling. *Proceedings of the Workshop on Comparing Corpora - Volume 9*, 1–6. Presented in Hong Kong. doi:10.3115/1117729.1117730