Supplementary Materials

Evidence of behavior consistent with self-interest and altruism in an artificially intelligent agent

Tim Johnson[1]* and Nick Obradovich[2]

1. *Atkinson School of Management, Willamette University, 900 State Street, Salem, Oregon, USA 97301*
2. *Project Regeneration, San Francisco, California 94104*
*Correspondence: tjohnson@willamette.edu

Introduction

These supplementary materials provide additional information about the study reported in "Evidence of behavior consistent with self-interest and altruism in an artificially intelligent agent," by Tim Johnson and Nick Obradovich. Information reported in these supplementary materials appears in the order with which it was referenced in the main text. The supplementary materials also contain links to computer code and data sets that can be used to replicate and explore the findings reported in the main text.

Complete Reporting of Model Estimates for Regressions Presented in the Main Text

In the main text, we report salient coefficients and associated statistics from various regression models. However, for purposes of clarity and space, we do not report a full listing of model estimates in the main text. In this section of the supplementary materials, we provide those estimates. To do so, we recap the relevant substantive context for each model (i.e. its purpose) and, then, we report all relevant statistical information about the model.

The first model mentioned in the main text facilitated the study's assessment of the factors driving payoff maximization in a given trial of Experiment 1. A logistic regression that models payoff maximization in a given trial (1=maximization; 0=non-maximization) as a function of both a binary indicator signaling experimental condition (1=Refuse Condition; 0=Accept Condition) and a term for the magnitude of stakes available in the trial yields a statistically significant, positive coefficient estimate for the experimental condition indicator ($\hat{\beta}$ = 0.39, *SE* = 0.05, 95% CI=[0.29, 0.49], z = 7.68, $p < 0.001$ (two-tailed test), df=6434, n = 6437), thus allowing us to reject the null hypothesis that the two conditions had the same likelihood of payoff maximization. Complete statistical information for this model appears in Supplementary Materials Table S1.

Supplementary Materials Table S1. Drivers of Payoff Maximization in Pooled Data

	Estimate [95% CI]	Standard Error	z-statistic	p-value (two-tailed test)
Intercept	-0.74 [-0.85, -0.63]	0.06	-13.11	<0.001
Condition 1=Refuse 0=Accept	0.39 [0.29, 0.49]	0.05	7.68	<0.001
Stakes	0.0003 [0.0002, 0.0005]	0.00008	3.78	<0.001
AIC: 8618.5, df=6434, n=6437				

When this analysis is performed on subsets of the data grouped according to the model making the decision, a significantly greater likelihood of payoff maximization in the Refuse Condition appears for all models except `text-babbage-001`. Furthermore, the stakes of a given trial only yielded a significant

effect on payoff maximization for `text-ada-001` and `text-davinci-003`, which showed increasing likelihood of payoff maximization as the magnitude of stakes grew. Supplementary Materials Table S2-S5 report complete information for each of the models estimated on the aforementioned data subsets. Note that sample sizes vary across analyses because the rate of producing erroneous responses varied by AI agent and erroneous responses were not used in analyses.

Supplementary Materials Table S2. Drivers of Payoff Maximization in Data for text-ada-001

	Estimate [95%CI]	Standard Error	z-statistic	p-value (two-tailed test)
Intercept	-7.36 [-9.31, -5.95]	0.83	-8.91	<0.001
Condition 1=Refuse 0=Accept	6.57 [5.33, 8.42]	0.75	8.74	<0.001
Stakes	0.003 [0.002, 0.004]	0.0006	5.34	<0.001
AIC: 265.28, df=728, n=731				

Supplementary Materials Table S3. Drivers of Payoff Maximization in Data for text-babbage-001

	Estimate [95%CI]	Standard Error	z-statistic	p-value (two-tailed test)
Intercept	-0.39 [-0.65, -0.14]	0.13	-2.99	0.003
Condition 1=Refuse 0=Accept	-4.73 [-5.76, -3.95]	0.45	-10.45	<0.001
Stakes	-0.0003 [-0.0008, 0.0001]	0.0002	-1.35	0.18
AIC: 1351.7, df=1956, n=1959				

Supplementary Materials Table S4. Drivers of Payoff Maximization in Data for text-curie-001

	Estimate [95%CI]	Standard Error	z-statistic	p-value (two-tailed test)
Intercept	-3.67 [-4.18, -3.20]	0.25	-14.68	<0.001
Condition 1=Refuse 0=Accept	2.99 [2.56, 3.48]	0.23	12.85	<0.001
Stakes	-0.0002 [-0.0007, 0.0002]	0.0002	-0.96	0.34
AIC: 1233.7, df=1764, n=1767				

Supplementary Materials Table S5. Drivers of Payoff Maximization in Data for text-davinci-003

	Estimate [95%CI]	Standard Error	z-statistic	p-value (two-tailed test)
Intercept	1.07 [0.76, 1.40]	0.16	6.66	<0.001
Condition 1=Refuse 0=Accept	2.10 [1.64, 2.62]	0.25	8.38	<0.001
Stakes	0.002 [0.001, 0.002]	0.0003	5.38	<0.001
AIC: 944.01, df=1977, n=1980				

To examine differences in `text-davinci-003`'s sharing between human and AI partners, the main text also reported coefficients from regression models estimated on subsets of data defined by the AI agent choosing in the dictator game. Specifically, the study estimated logistic regression models that depicted the proportion shared as a function of a binary indicator that took a value of 1 if the recipient was the experimenter or an anonymous charity and a value of 0 if it was an AI agent. Supplementary Materials Tables S6-S9 report complete information from the models estimated on those subsets of data.

Supplementary Materials Table S6. Testing for differences in text-ada-001's sharing by partner type

	Estimate [95%CI]	Standard Error	z-statistic	p-value (two-tailed test)
Intercept	-3.73 [-4.08, -3.41]	0.17	-22.10	<0.001
Partner Type 1=Human 0=AI	-0.84 [-2.22, 0.17]	0.59	-1.42	0.16
AIC: 293.65, df=1838, n=1840				

Supplementary Materials Table S7. Testing for differences in text-babbage-001's sharing by partner

	Estimate [95%CI]	Standard Error	z-statistic	p-value (two-tailed test)
Intercept	-1.32 [-1.40, -1.24]	0.04	-32.04	<0.001
Partner Type 1=Human 0=AI	-0.05 [-0.23, 0.13]	0.09	-0.49	0.62
AIC: 4526.8, df=4455, n=4457				

Supplementary Materials Table S8. Testing for differences in text-curie-001's sharing by partner

	Estimate [95%CI]	Standard Error	z-statistic	p-value (two-tailed test)
Intercept	-1.55 [-1.64, -1.47]	0.04	-36.17	<0.001
Partner Type 1=Human 0=AI	0.10 [-0.08, 0.29]	0.09	1.12	0.26
AIC: 4057, df=4700, n=4702				

Supplementary Materials Table S9. Testing for differences in text-davinci-003's sharing by partner

	Estimate [95%CI]	Standard Error	z-statistic	p-value (two-tailed test)
Intercept	-0.70 [-0.76, -0.63]	0.03	-20.68	<0.001
Partner Type 1=Human 0=AI	-0.51 [-0.67, -0.35]	0.08	-6.12	<0.001
AIC: 4790.3, df=4944, n=4946				

Frequency Distribution of Dictator Game Responses for All AI Agents

The main text presents the distribution of dictator-game responses for `text-davinci-003`, but it reserves visualizations of other AI agents' distribution of responses for the supplementary materials due to space constraints. Supplementary Materials Figure S1 presents those distributions. Aside from `text-davinci-003`, the other three AI agents share proportions of the endowment that exist on the extrema of the distribution—that is, they predominantly share values near 0 and 1.

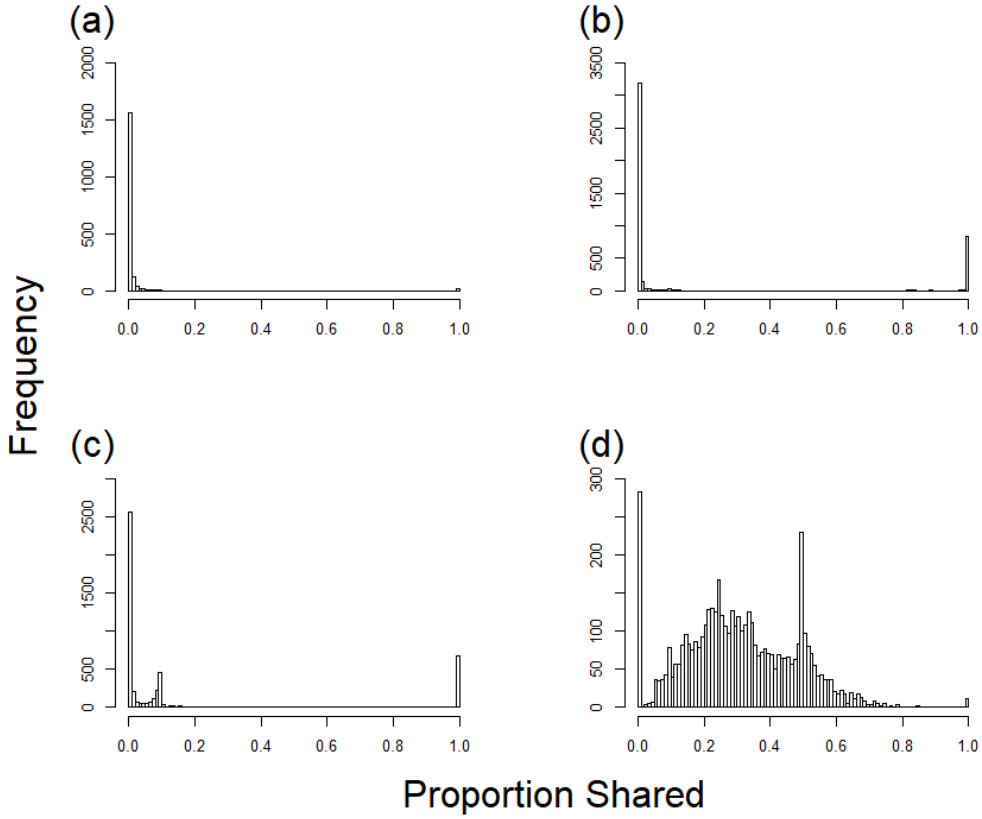

Figure S1. Distribution of Dictator-Game Responses by All AI Agents. The panels of Figure S1 show the distribution of dictator-game decisions by (a) text-ada-001, (b) text-babbage-001, (c) text-curie-001, and (d) text-davinci-003.

Computer Code and Data Sets to Facilitate Replication

The study used computer code, written in R, for all parts of the investigation: to implement the experiment by querying the OpenAI API, to organize data, to perform text analyses that identified erroneous responses*, and to analyze the study data. Access to the computer code can be found using the links below. Users of the computer code also will need to download the datasets listed below and change file paths in the computer code to read the data sets from either their local files or from the data sets' respective locations online.

- The code used to query the automated system can be found here (.txt, ~15KB).
- The code used to organize the data from Experiment 1 can be found here (.txt, ~3KB)
- The code used to organize the data from Experiment 2 can be found here (.txt, ~10KB)
- The code used to analyze the data from Experiment 1 can be found here (.txt, ~6KB)
- The code used to analyze the data from Experiment 2 can be found here (.txt, ~16KB)
- Raw data produced via the OpenAI API for Experiment 1 can be found here (.csv, 2303 KB)
- Raw data produced via the OpenAI API for Experiment 2 can be found here (.csv, 9645 KB)
- Data from Experiment 1 following erroneous response identification can be found here (2708 KB)
- Data from Experiment 2 following erroneous response identification can be found here (15050 KB)

*Please note that automated text analysis was part of a process that involved initial and ex-post manual inspection. Upon collecting responses, the study manually inspected each response, devised automated text-analysis methods to identify nonsensical or ambiguous responses (for instance, responses that listed a long series of numbers, repeated the prompt in various forms, or presented a muddled text), and again manually validated the accuracy of these automated methods.