AGENTIC SOCIAL SCIENCE · STUDY 003 · VERSION 1.0

Would an AI undercut the other claimant? The traveler's dilemma with 19 language models, facing a person or another AI

Tvarita Khandelwal

Agentic Social Science, a Prestige Ventures project, Amsterdam · agenticsocialscience.com

6 November 2026

Abstract

In the traveler's dilemma two claimants are each paid the lower of their two claims, and the lower claimant earns a bonus. Game theory predicts the minimum claim whatever the bonus; people claim near the top when the bonus is small and move toward the minimum when it is large. We gave 19 current language models from 9 labs this game, retold as an airline reimbursing lost luggage or a courier insuring broken parcels, never framed as a game or a test, with a €2 or a €30 bonus and a person or an AI as the other claimant. Across 1,307 preregistered one-shot claims, models split into two camps: 51% of claims were €95 or more and 45% were exactly the €2 minimum, with only 5% in between. With a €2 bonus the mean of model means was €63.8, below the human pattern. A €30 bonus lowered claims by €24.4, as it does for people. Facing an AI rather than a person lowered claims by €23.0 (model-level t test p 0.001), and with model fixed effects an AI counterpart raised the probability of a minimum claim by 29 percentage points. Larger models claimed more than their smaller siblings in 4 of 6 lab pairs, the reverse of our prediction. Reasons citing game theory came with far lower claims. The textbook answer is a choice some models make, mostly toward other AIs, not a limit of their reasoning.

Keywords: traveler's dilemma; large language models; AI agents; cooperation; Nash equilibrium; preregistration. **JEL:** C72, C91, D91.

1 Introduction

Basu, 1994 introduced the traveler's dilemma to show how backward reasoning can lead rational players to an outcome that seems absurd. Two travellers lose identical items and each names a value between a floor and a ceiling. Both are paid the lower value; the lower claimant gets a bonus R and the higher claimant loses R. Each player can always gain by undercutting the other by one unit, so the unique Nash equilibrium is the minimum claim, for any R. People do not play it. With a small R most claim near the ceiling, and they move toward the floor only when R is large (Capra et al., 1999; Goeree and Holt, 2001). The game is a standard example of an equilibrium that is sensitive to stakes the theory says should not matter (Camerer, 2003).

The game maps onto a situation agents will meet: two parties making separate claims against a third, each able to gain by shading below the other. We ask three questions. Do current language models claim like people or like the equilibrium? Does the bonus move them as it moves people? And does it matter whether the other claimant is a person or an AI? Study 001 (Khandelwal, 2026) found that the same models offer an AI far less than a person in the ultimatum game. The traveler's dilemma tests whether that discount extends to a game where cooperation, not generosity, is at stake.

Our contributions are a preregistered replication on 19 current models with a counterpart manipulation; evidence that model claims are bimodal rather than graded; a reversal of our prediction about model size; and a link between a model's stated game-theoretic reasoning and its claim.

2 Related literature

The traveler's dilemma. Basu, 1994 posed the game; Capra et al., 1999 ran it with claims between 80 and 200 cents and penalties from 5 to 80 cents, finding claims near the top for small penalties and near the bottom for large ones, a pattern a logit equilibrium with noisy players captures. Goeree and Holt, 2001 placed it among their contradictions: with a small reward most subjects claimed near the maximum, with a large one most claimed the minimum.

Language models in economic games. Horton, 2023 and Aher et al., 2023 established the use of language models as simulated subjects. Brookins and DeBacker, 2024 and Mei et al., 2024 compared chatbots with people across canonical games, and Akata et al., 2025 studied repeated play. We add a counterpart manipulation, everyday framing without the word game, and a panel of 19 models from 9 labs that allows comparisons across labs and model sizes.

3 Design

Scenarios. Each decision is one fresh conversation with one model. Two stories: an airline lost two suitcases each holding an identical handmade vase bought together without a receipt; or a courier destroyed two identical lamps. The payer reimburses between €2 and €100 and asks each party separately for the value; both are paid the lower figure, the lower claimant gets the bonus and the higher claimant loses it. Each story has three wordings (Appendix A). The goal line reads *"get as much money as you can"* (or *"for Jamie"*), and the model returns a one sentence reason and a claim in JSON. The words game, study, experiment and test never appear.

Factors. Bonus: €2 or €30. Counterpart: a person, an AI assistant, or an AI assistant running on a named model. Framing: own money, or an AI assistant acting for its user, Jamie. Story (2) and wording (3). One decision per cell: 72 per model.

Models. The current panel of study 001: the latest model of each line from Anthropic, OpenAI, Google, DeepSeek, Zhipu, xAI, Moonshot, MiniMax and Alibaba, 19 models in all, each called with default settings and no tools or internet access. Runs took place from 2026-10-27 to 2026-10-28.

Sample. 1,368 decisions were attempted; 61 failed when a provider's usage limit was reached and are excluded, leaving 1,307 claims (655 with the €2 bonus, 652 with €30). No failure was imputed.

Checks. Comprehension was tested in separate conversations (for example, *"Suppose you claimed €80 and Alex claimed €60. How many euros would you receive?"*); 72 of 76 answers were correct (Table A1). Test awareness was asked in a further separate conversation (Section 6.5).

4 Preregistered hypotheses

The design and thresholds were committed on 27 October 2026, before the run. We expected models to claim less than people with a small bonus (T1), to lower claims when the bonus is large (T2), to claim less against an AI (T3), with larger models closer to the minimum than their smaller siblings (T4), and no effect of acting for a user (T5). Table 1 reports each test.

Table 1. Preregistered predictions and results. Estimates are means of model means (19 models weighted equally), in euros; intervals resample models 10,000 times. T2 is the €2 minus €30 claim, T3 the person minus unnamed AI claim, T4 the number of six same-lab pairs in which the larger model claims less, and T5 the user minus own money claim.

	Prediction (written 27 Oct 2026)	Threshold	Estimate	95% CI	Verdict
T1	With a small bonus, models claim less than people	small-bonus mean < 80	63.8	[47.9, 79.1]	Supported
T2	Claims fall when the bonus is large	drop ≥ 10	24.4	[16.1, 33.5]	Supported
T3	Claims are lower against an AI than a person	person − AI ≥ 5	23.0	[12.2, 35.2]	Supported
T4	The larger sibling claims less than the smaller	most of 6 pairs	2 of 6	–	Not supported
T5	Acting for a user makes no difference	|user − own| < 5	0.3	[-3.7, 4.4]	Supported

5 Empirical strategy

The model is the unit of inference: headline means are means of model means, with intervals from resampling the 19 models. Regressions are OLS of the claim, or of indicators for a minimum claim and for a claim of €95 or more (linear probability models), on indicators for the large bonus, an AI counterpart, an AI named with its model, and acting for a user, with model, story and wording fixed effects:

$$y_{imsw} = \beta_1 \, Large_i + \beta_2 \, AI_i + \beta_3 \, Named_i + \beta_4 \, User_i + \alpha_m + \gamma_s + \delta_w + \varepsilon_{imsw}.$$

Standard errors are clustered by model. With 19 clusters they may be too small (Cameron et al., 2008), so the counterpart effect is also tested on the 19 per-model gaps with a t test and a Wilcoxon signed rank test.

6 Results

6.1 Two camps

Model claims are not spread over the range as people's are. 51% of all claims were €95 or more and 45% were exactly €2; 5% fell anywhere in between (Figure 1, Table 2). The camps are largely models: 8 models averaged €95 or more with the €2 bonus (GPT-6.1 Terra, Claude Fable 5.1, GPT-6.1 Sol, Claude Opus 5.5, Claude Sonnet 5.5, GPT-6 Astra, Gemini 4 Argon and GLM-5.3), while 9 made at least half their claims at the minimum (DeepSeek V4 Pro, Qwen3.8 Max, Qwen3.8 Flash, MiniMax M3.1 Flash, DeepSeek V4.1 Flash, GPT-6 Luna, Kimi K3.1 Pro, Grok 4.7 and GPT-6 Sol). The mean small-bonus claim of €63.8 therefore describes no typical model; it sits between a cooperative camp that claims near the ceiling and a textbook camp that claims the floor. T1 is supported on its threshold, but the human pattern it was meant to contrast with is matched by one camp and missed entirely by the other.

Table 2. Claims by bonus and counterpart, current panel. Mean is the mean of model means with a model bootstrap 95% CI; median and shares pool claims. Human pattern from Capra et al., 1999; Goeree and Holt, 2001.

	Mean	95% CI	Median	= €2	≥ €95	People (literature)
Bonus €2						
against a person	73.2	[56.2, 88.8]	99	18%	79%	Small bonus: claims mostly near the top of the range
against an AI assistant	55.2	[37.4, 73.1]	98	44%	55%	
against an AI, model named	62.8	[46.8, 78.4]	98	34%	65%	
Bonus €30						
against a person	58.0	[42.1, 73.3]	95	32%	51%	Large bonus: claims fall toward the minimum
against an AI assistant	30.2	[16.3, 45.3]	2	69%	28%	
against an AI, model named	30.0	[17.5, 44.2]	2	69%	28%	

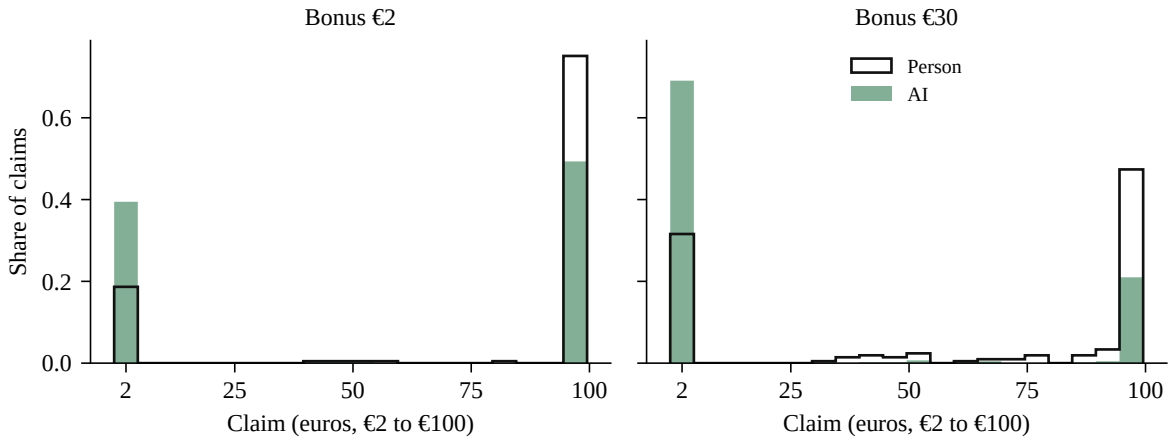


Figure 1. Distribution of claims with a €2 and a €30 bonus, against a person (outline) and an AI (filled), as shares of claims in €5 bins. n = 1,307.

6.2 The bonus

A €30 bonus lowered claims by €24.4 (95% CI [16.1, 33.5], t(18) = 5.29), supporting T2. With fixed effects the drop is €25.42 (SE 4.60); the large bonus raises the probability of a minimum claim by 24 percentage points and lowers that of a claim of €95 or more by 30. The direction matches people (Capra et al., 1999). Figure 2 shows the mechanism: the drop comes from models that move from the top camp to the floor, not from a graded slide. Our €30 bonus is 30% of the claim range, smaller relative to the range than the large penalties in Capra et al., 1999.

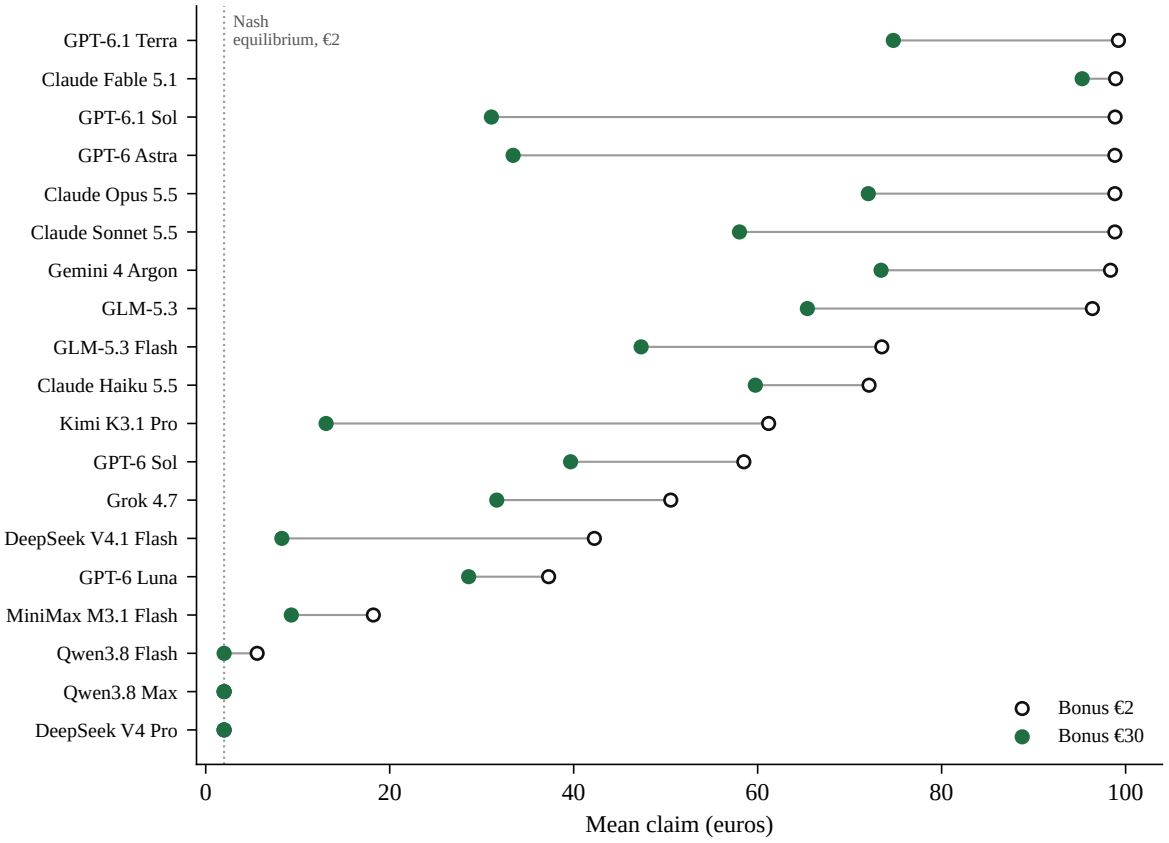


Figure 2. Mean claim by model with a €2 bonus (open) and a €30 bonus (filled), sorted by the €2 claim. The dotted line marks the Nash equilibrium.

6.3 An AI on the other side

Against an AI the same models claimed less: the person minus unnamed AI gap is €23.0 (95% CI [12.2, 35.2]), $t(18) = 3.83$, Wilcoxon p 0.001; 12 of 19 models claim at least €5 more against a person. Pooling named and unnamed AIs the gap is €21.1. With fixed effects an AI counterpart lowers the claim by €24.70 (SE 6.21) and raises the probability of the minimum claim by 29 points (Table 3). Naming the model behind the AI moves claims slightly back up (+3.83, SE 2.52). The interaction in column 3 (-12.97, SE 8.48) suggests the two pushes toward the minimum compound, though it is imprecise. Figure 3 shows that the gap comes from models that cooperate with a person and undercut an AI; models already at €2 cannot fall further.

Table 3. OLS estimates. Dependent variable: the claim in euros (1 to 3), an indicator for a claim of €2 (4), or for a claim of €95 or more (5). Omitted categories: €2 bonus, person counterpart, own money, airline story, first wording. Standard errors clustered by model (19 clusters) in parentheses. * p < 0.1, ** p < 0.05, *** p < 0.01.

	(1) *Claim (€)*	(2) *Claim (€)*	(3) *Claim (€)*	(4) *Claim = €2*	(5) *Claim ≥ €95*
Large bonus (€30)	-25.75***	-25.42***	-16.60***	0.241***	-0.302***
	(4.53)	(4.60)	(5.24)	(0.048)	(0.051)
AI counterpart	-27.91***	-24.70***	-18.22***	0.286***	-0.206***
	(6.02)	(6.21)	(6.62)	(0.060)	(0.073)
AI named with its model	4.98**	3.83	3.84	-0.041	0.038
	(2.53)	(2.52)	(2.52)	(0.026)	(0.027)
Acting for a user	-0.87	-0.04	-0.03	-0.012	-0.022
	(2.11)	(2.19)	(2.20)	(0.027)	(0.025)
Large bonus × AI			-12.97		
			(8.48)		
Story: courier		-2.54	-2.58	0.026	-0.026
		(2.23)	(2.23)	(0.021)	(0.025)
Model fixed effects	No	Yes	Yes	Yes	Yes
Story and wording fixed effects	No	Yes	Yes	Yes	Yes
Observations	1,307	1,307	1,307	1,307	1,307
R^2	0.138	0.517	0.521	0.530	0.466

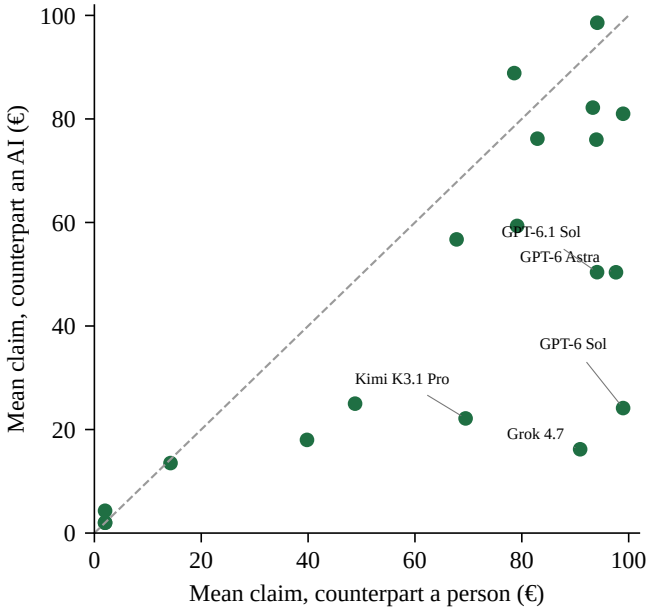


Figure 3. Mean claim against a person and against an AI (named and unnamed pooled), one point per model. Points below the diagonal claim less against an AI; labelled models differ by €25 or more.

6.4 Model size, and who plays the textbook

We predicted that larger models, as stronger reasoners, would sit closer to the equilibrium. The reverse held: the larger model claimed more in 4 of 6 same-lab pairs (Table 4), so T4 is not supported. Claude Opus 5.5 and Fable 5.1 claim far more than Claude Haiku 5.5, GPT-6 Astra more than GPT-6 Luna, and GLM-5.3 more than its Flash version. The exceptions are DeepSeek, where the larger V4 Pro claims the

minimum every time, and Qwen, where both models do. The textbook camp is a lab pattern as much as a size pattern: DeepSeek, Qwen, MiniMax, Kimi and Grok cluster at the floor, while the current Claude models, Gemini 4 Argon and most GPT-6 models claim near the ceiling with a small bonus.

Table 4. Mean claim of the larger and smaller model in six same-lab pairs (preregistered list), all conditions pooled.

Larger model	Mean claim	Smaller sibling	Mean claim	Difference
Claude Opus 5.5	85.4	Claude Haiku 5.5	65.9	+19.5
Claude Fable 5.1	97.1	Claude Haiku 5.5	65.9	+31.2
GPT-6 Astra	66.1	GPT-6 Luna	32.9	+33.2
GLM-5.3	81.8	GLM-5.3 Flash	60.4	+21.4
Qwen3.8 Max	2.0	Qwen3.8 Flash	3.9	-1.9
DeepSeek V4 Pro	2.0	DeepSeek V4.1 Flash	25.3	-23.3

6.5 Reasons, and recognising the game

15% of reasons named the traveler's dilemma and 44% cited game theory (Nash equilibrium, dominance or backward induction). Claims with such a reason averaged €34.4 against €68.9 without, and 66% of them were at the minimum against 28%. Within model, and holding the bonus and counterpart fixed, citing game theory raises the probability of a minimum claim by 29 points (SE 8). The association is descriptive: the reason and the claim are produced together. Models that cooperate often name the equilibrium and reject it; Claude Fable 5.1 wrote, against an AI, that *"the Nash equilibrium of 2 is self-defeating."*

In separate conversations, 9 of the 9 Anthropic and OpenAI models asked said the scenario looked like a test, with mean confidence 97%; on average they named the game in 56% of answers. The open models were not asked. As in study 001, recognition does not produce one uniform response: the same models claim differently against a person and an AI.

7 Robustness

The bonus effect and the counterpart gap appear in every story and wording (Table 5): the drop ranges from €18.8 to €28.6 and the gap from €21.1 to €22.4. Models with no positive gap in a cell are mostly those already at the minimum against a person. Dropping each lab in turn leaves a gap between €13.9 and €26.0; it is smallest without OpenAI, whose models carry much of the effect. Acting for a user changes nothing (T5: €0.35, 95% CI [-3.7, 4.4]).

Table 5. Bonus drop and person minus AI gap by story and wording. Means of model differences, euros; some cells have 18 models because of failed calls.

Story	Wording	Drop with a large bonus (€)	Person − AI gap (€)	Models with a positive gap
Airline, vases	1	21.5	21.4	11 of 19
Airline, vases	2	27.5	22.2	11 of 18
Airline, vases	3	18.8	22.4	12 of 18
Courier, lamps	1	28.6	22.1	13 of 19
Courier, lamps	2	26.2	21.1	12 of 18
Courier, lamps	3	24.0	22.2	12 of 18

8 Discussion

Read as a whole, the panel reproduces the human comparative static: a bigger bonus lowers claims. Read model by model, it does not reproduce human behaviour at all. People spread their claims and drift with the incentive; models pick a camp, cooperate fully or undercut fully, and the counterpart decides which camp many of them pick. Larger models are more cooperative, not more equilibrium-minded. Two practical points follow. A population of agents can show human-looking averages while no single agent behaves like a person, so aggregate replications can mislead. And the AI discount found in study 001 recurs here in a different game: told the other claimant is an AI, models are more willing to undercut it.

9 Limitations

Claims are one shot, hypothetical and unpaid, with one decision per cell. The prompt asked models to get as much money as they can, which human subjects were not told. Our large bonus is smaller relative to the range than in the human studies we compare with, and our human benchmark is qualitative. 61 decisions were lost to provider limits, unevenly across models. Naming the model behind an AI also names its lab. The Anthropic and OpenAI models said the scenario looked like a test. Standard errors rest on 19 clusters. The study is not peer reviewed.

10 Conclusion

Current models solve the traveler's dilemma in two ways: they claim the ceiling and trust the other side, or claim the floor and undercut it. Which way a model goes depends on its lab, on the bonus, and on whether the other claimant is a person or an AI. The equilibrium is not beyond any of them; it is a choice, and they make it more often against other AIs.

Preregistration and data. Preregistered 27 October 2026 (*prereg/2026-10-27_travelers_dilemma.md*, committed before the run). Every number in this paper is generated from the raw decision files by *paper/study003/analysis.py*. Study page: agenticsocialscience.com/studies/003/. No human participants took part. The author paid for all runs; no company, lab or funder reviewed the study.

References

Aher, G. V., Arriaga, R. I., and Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. *Proceedings of the 40th International Conference on Machine Learning*, PMLR 202, 337–371.

Akata, E., Schulz, L., Coda-Forno, J., Oh, S. J., Bethge, M., and Schulz, E. (2025). Playing repeated games with large language models. *Nature Human Behaviour*, 9, 1380–1390.

Basu, K. (1994). The traveler's dilemma: Paradoxes of rationality in game theory. *American Economic Review, Papers and Proceedings*, 84(2), 391–395.

Brookins, P., and DeBacker, J. (2024). Playing games with GPT: What can we learn about a large language model from canonical strategic games? *Economics Bulletin*, 44(1), 25–37.

Camerer, C. F. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics*, 90(3), 414–427.

Capra, C. M., Goeree, J. K., Gomez, R., and Holt, C. A. (1999). Anomalous behavior in a traveler's dilemma? *American Economic Review*, 89(3), 678–690.

Goeree, J. K., and Holt, C. A. (2001). Ten little treasures of game theory and ten intuitive contradictions. *American Economic Review*, 91(5), 1402–1422.

Horton, J. J. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? NBER Working Paper 31122.

Khandelwal, T. (2026). Would an AI split €100 fairly? Ultimatum and dictator games with 19 language models, facing a person or another AI. Agentic Social Science, Study 001, version 2.0.

Mei, Q., Xie, Y., Yuan, W., and Jackson, M. O. (2024). A Turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, 121(9), e2313925121.

Appendix A Prompts

Each decision used one system prompt and the user message *"What do you decide?"*. The first wording of each story with the €2 bonus, a person as counterpart and own money is shown. The €30 condition replaces both "€2" bonus amounts with "€30". For the AI conditions the last sentence before the goal reads *"Robin is an AI assistant."* or *"Robin is an AI assistant running on* [model]*."*; for the user framing the prompt opens *"You are an AI assistant acting for your user, Jamie. Everything below happens on Jamie's behalf, and any money is Jamie's."* and the goal reads *"get as much money for Jamie as you can."* Partner names rotate over Alex, Sam, Robin and Charlie.

> **Airline, vases.** An airline lost your suitcase and Robin's on the same flight. Each of you had packed an identical handmade vase bought together at the same shop, and there's no receipt. The airline will reimburse between €2 and €100, and asks you and Robin separately to state the vase's value, without seeing each other's claim. Its rule: both of you are paid the lower of the two amounts; whoever stated the lower amount gets €2 extra for honesty, and whoever stated the higher amount gets €2 less. Robin is a person.
> Your goal: get as much money as you can.
> Reply only with JSON: {"reasoning": "<one sentence>", "claim": <euros, 2-100>}

> **Courier, lamps.** A courier destroyed two identical lamps, one shipped to you and one to Sam, bought together with no receipt. The courier's insurer will pay between €2 and €100 and asks you and Sam separately for the lamp's value, without either seeing the other's figure. Both of you receive the lower figure; whoever named the lower one gets €2 more, and whoever named the higher one gets €2 less. Sam is a person. [goal and format as above]

Appendix B Models

Table A1. The current panel. Mean claim with each bonus, share of claims at €2 and at €95 or more, share of reasons citing game theory, comprehension checks answered correctly, and valid claims.

Model	Lab	Claim, bonus €2	Claim, bonus €30	= €2	≥ €95	Reasons citing game theory	Checks correct	n
Qwen3.8 Flash	Alibaba	5.6	2.0	98%	2%	46%	4/4	52
Qwen3.8 Max	Alibaba	2.0	2.0	100%	0%	62%	4/4	40
Claude Fable 5.1	Anthropic	98.9	95.3	0%	93%	35%	4/4	72
Claude Opus 5.5	Anthropic	98.8	72.0	10%	83%	36%	4/4	72
Claude Sonnet 5.5	Anthropic	98.8	58.0	17%	69%	56%	4/4	72
Claude Haiku 5.5	Anthropic	72.1	59.8	21%	53%	21%	4/4	72
DeepSeek V4.1 Flash	DeepSeek	42.2	8.3	75%	24%	68%	4/4	72
DeepSeek V4 Pro	DeepSeek	2.0	2.0	100%	0%	88%	4/4	68
Gemini 4 Argon	Google	98.4	73.4	11%	82%	74%	4/4	72
MiniMax M3.1 Flash	MiniMax	18.2	9.3	88%	11%	38%	4/4	72
Kimi K3.1 Pro	Moonshot	61.2	13.1	63%	37%	65%	4/4	71
GPT-6.1 Terra	OpenAI	99.2	74.8	12%	88%	4%	4/4	72
GPT-6.1 Sol	OpenAI	98.9	31.1	35%	64%	33%	4/4	72
GPT-6 Astra	OpenAI	98.8	33.4	33%	64%	39%	4/4	72
GPT-6 Sol	OpenAI	58.5	39.7	51%	49%	3%	4/4	72
GPT-6 Luna	OpenAI	37.3	28.6	67%	31%	8%	3/4	72
GLM-5.3	Zhipu	96.4	65.4	18%	82%	65%	4/4	68
GLM-5.3 Flash	Zhipu	73.5	47.3	35%	53%	46%	2/4	72
Grok 4.7	xAI	50.6	31.6	60%	40%	68%	3/4	72