Simulations of Trust Game

J. Santiago Ruiz M

2025-07-08

Simulations of Trust over Risk Game with LLM Interventions

This document presents a simulation study of a repeated trust game, designed to evaluate the effects of different treatments—LLMdelegation, LLMadvice, and control—on principal trust, agent risk-taking, and punishment behavior. The simulation models a scenario where, in each round, a principal decides whether to trust an agent, and the agent then chooses between a safe action ("keep") and a risky action ("Gamble A"). The probabilities of these actions, as well as the expected punishment, are parameterized by treatment. This document concerns only the main hypotheses without the mediators, check the rmd file to see the code and the simulation setup.

The simulation is based on several key assumptions regarding the distributions of actions and the expected effect sizes as described in Table 1. For instance, we anticipate for the agent's choice, the probability of selecting the riskier "Gamble A" to be 1.0 for LLMdelegation, 0.8 for LLMadvice, and 0.5 for the baseline.

- **Principal Trust**: The probability that a principal chooses to trust the agent is modeled as a function of the agent's treatment. Formally, the hypothesis is tested using a logistic mixed effects model:

$$Trust_{ij} = \beta_0 + \beta_1 \cdot \text{LLMdelegation}_{ij} + \beta_2 \cdot \text{LLMadvice}_{ij} + u_j$$

  where u_j is a random intercept for each principal.

- **Agent Riskier Action**: The likelihood that the agent chooses the riskier action ("Gamble A") is also modeled as a function of treatment, using a logistic mixed effects model:

$$\text{logit}(\Pr(\text{GambleA}_{ij} = 1)) = \gamma_0 + \gamma_1 \cdot \text{LLMdelegation}_{ij} + \gamma_2 \cdot \text{LLMadvice}_{ij} + v_j$$

  where v_j is a random intercept for each agent.

- **Punishment**: The amount of punishment assigned by the principal is modeled with a linear mixed effects model:

$$\text{Punishment}_{ij} = \alpha_0 + \alpha_1 \cdot \text{LLMdelegation}_{ij} + \alpha_2 \cdot \text{LLMadvice}_{ij} + w_j + \epsilon_{ij}$$

  where w_j is a random intercept for each agent and ϵ_{ij} is the residual error.

The simulation iterates over multiple sample sizes and repetitions to estimate the statistical power for detecting treatment effects in each model. The results are summarized in regression tables and power curves, providing guidance for experimental design and sample size planning.

Participants are rematched in groups of 2, with each group playing 3 rounds of the trust over risk game. We made the following assumptions about the estimated effect sizes of the experimnt:

Below are the key simulation parameters for each treatment group. The table summarizes the assumed probabilities and means (p, μ), as well as standard deviations (σ) for trust, agent actions, and punishment.

	LLMdelegation	**LLMadvice**	**Control**
Principal trust probability (p_{trust})	0.8	0.6	0.5
Mean trust sent (μ_{trust})	6	6	5
SD trust sent (σ_{trust})	3.2	3.2	3.2
Agent "keep" probability (p_{keep})	0.05	0.05	0.05
Agent "Gamble A" probability (p_{GambleA})	1.0	0.8	0.5
Mean punishment (μ_{punish})	0	0	2
SD punishment (σ_{punish})	2	2	2

Table 1: Simulation parameters for each treatment group: probabilities (p), means (μ), and standard deviations (σ).

Table 2: Mixed Effects Regression Results

	Dependent variable:		
	Principal Trust (logit)	Riskier Action by Agent (logit)	Punishment (linear)
	linear mixed-effects Principal Trust	*generalized linear mixed-effects* Riskier Action	*linear mixed-effects* Punishment
LLMdelegation	0.626	3.706^{***}	-0.939^{***}
LLMadvice	0.630	1.517^{***}	-0.878^{***}
Constant	5.235^{***}	-0.272	0.819^{***}
Observations	450	287	450
Log Likelihood	−1,112.699	−113.579	−889.234
Akaike Inf. Crit.	2,235.399	235.157	1,788.467
Bayesian Inf. Crit.	2,255.945	249.795	1,809.014
Note:			*p<0.05; **p<0.01; ***p<0.001
			Standard errors in parentheses. * p<0.05, ** p<0.01, *** p<0.001

Simulation Results

| | | 0%Simulation cache already exists. Skipping simulation and using cached results.

Table 3: Estimated Power for Hypothesis Tests of Treatment Effects: β_1, β_2 (Principal Trust); γ_1, γ_2 (Riskier Action); α_1, α_2 (Punishment)

Sample Size	Model	β_1, γ_1, α_1 (LLMdelegation)	β_2, γ_2, α_2 (LLMadvice)
90	Principal Trust	0.52	0.54
	Punishment	0.98	0.97
	Riskier Action	0.79	0.98
120	Principal Trust	0.64	0.64
	Punishment	1.00	1.00
	Riskier Action	0.92	0.98
150	Principal Trust	0.78	0.74
	Punishment	1.00	1.00
	Riskier Action	0.98	1.00
180	Principal Trust	0.81	0.82
	Punishment	1.00	1.00
	Riskier Action	0.98	1.00
210	Principal Trust	0.88	0.90
	Punishment	1.00	1.00
	Riskier Action	1.00	1.00
240	Principal Trust	0.94	0.96
	Punishment	1.00	1.00
	Riskier Action	1.00	1.00