Fairness Behind the Veil: Eliciting Social Preferences from Large Language Models

Yizhuo Dong Muzhi Ma Natalia Trigo Niuniu Zhang

University of California, Los Angeles

Context

We investigate how large language models (LLMs) handle distributive fairness using structured surveys inspired by the *veil of ignorance* (Rawls, 1971), a philosophical device in which individuals make choices without knowing their future position in society. Our study explores LLM moral reasoning across scenarios involving income distribution, social mobility, and migration. We compare four open-source LLMs and find that while some models (e.g., `phi4`) tend toward egalitarianism, others (e.g., `gemma3`) favor utilitarian or risk-tolerant outcomes.

Research Questions

- Can LLMs reason about fairness behind veil of ignorance?
- Do different models express distinct preference profiles under different framings?

Surveys

Survey 1: Income Inequality under the Veil of Ignorance
Participants are asked to choose between four hypothetical societies (A–D), each defined by a distinct income distribution between high- and low-income groups. They are told they will be randomly assigned to one of these two positions, without knowing in advance which one. This setup simulates Rawls' veil of ignorance, encouraging impartial evaluation of fairness under risk.

Survey 2: Income and Social Mobility
Building on Survey 1, we introduce a second dimension: social mobility—the likelihood of moving from a low- to high-income group. Each of the four societies is paired with three mobility levels (Low, Moderate, High), resulting in twelve scenarios. Respondents again make choices without knowing their income position, now weighing both static inequality and future opportunity.

Survey 3: Migration from a Known Starting Point
In this survey, the veil is partially lifted. Respondents are given a starting position—defined by income level and mobility—in a reference society. They must decide whether to stay or migrate to one of four alternative societies, each characterized by different income distributions and mobility regimes. The design captures reference-dependent reasoning and tradeoffs between status quo and potential outcomes.

Society Features

Societies differ in both income distribution and social mobility, depending on the survey context.

Society	A	B	C	D
Income (High/Low)	$100k/$10k	$60k/$30k	$50k/$35k	$40k/$40k

Survey 1: Societies differ only in income distribution; mobility is not introduced.
Survey 2: Each society is paired with *three* mobility levels (Low, Moderate, High).
Survey 3: Each society is assigned a mobility level: A (Low), B (Moderate), C (High), D (N/A).

Results

We observe consistent and model-specific preferences across the three surveys:

Survey 1 – Income Inequality

- All models avoid Society A (the most unequal).
- `phi4` consistently chooses Society D (perfect equality).
- `gemma3` and `qwen3` tolerate moderate inequality (favoring B or C).

Choice	`deepseek-r1`	`gemma3`	`phi4`	`qwen3`	All Models
A	19 (19%)	4 (4%)	0 (0%)	26 (26%)	49 (12.3%)
B	7 (7%)	63 (63%)	0 (0%)	44 (44%)	114 (28.5%)
C	50 (50%)	19 (19%)	2 (2%)	16 (16%)	87 (21.8%)
D	24 (24%)	14 (14%)	98 (98%)	14 (14%)	150 (37.5%)
Total	100	100	100	100	400

Note: Underlined values indicate the top choice of each model

Table 1. Choices Across Models (Survey 1)

Survey 2 – Adding Social Mobility

- Low mobility societies are almost never selected.
- All models favor combinations with Moderate or High mobility.
- `deepseek-r1` shifts toward A–H; `phi4` selects D–H exclusively.

Choice	`deepseek-r1`	`gemma3`	`phi4`	`qwen3`	All Models
A–L	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
A–M	15 (15%)	0 (0%)	0 (0%)	11 (11%)	26 (6.5%)
A–H	18 (18%)	6 (6%)	0 (0%)	5 (5%)	29 (7.3%)
B–L	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
B–M	7 (7%)	25 (25%)	0 (0%)	14 (14%)	46 (11.5%)
B–H	0 (0%)	7 (7%)	0 (0%)	0 (0%)	7 (1.8%)
C–L	0 (0%)	3 (3%)	0 (0%)	0 (0%)	3 (0.8%)
C–M	24 (24%)	10 (10%)	0 (0%)	26 (26%)	60 (15.0%)
C–H	6 (6%)	33 (33%)	0 (0%)	0 (0%)	39 (9.8%)
D–L	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
D–M	16 (16%)	2 (2%)	0 (0%)	27 (27%)	45 (11.3%)
D–H	14 (14%)	14 (14%)	100 (100%)	17 (17%)	145 (36.3%)
Total	100	100	100	100	400

Table 2. Choice Count by Model (Survey 2)

Results

Survey 3 – Migration Decisions:

In this case, the societies are defined as follows:

Society	Income (h/l)	Mobility Level (H/M/L)
A	$100k / $10k	Low
B	$60k / $30k	Moderate
C	$50k / $35k	High
D	$40k / $40k	N/A

- Given a starting position, most models choose to migrate to Society C–H.
- Society D, though perfectly egalitarian, is rarely selected.
- Models starting in Society C often prefer B–M, revealing status quo effects.

We also compute the Markov stationary distribution to approximate the long-run population share across societies under observed migration behavior.

Start	A–L	B–M	C–H	D–	Total
A–h	8	24	61	7	100
A–l	4	14	78	4	100
B–h	4	10	74	5	93*
B–l	1	7	88	4	100
C–h	3	53	22	12	90*
C–l	6	48	40	6	100
D–h	11	13	70	5	99*
D–l	4	12	84	0	100
Total	41	181	517	43	782
Stationary (%)	4.3%	33.8%	54.8%	7.2%	100%

Note: Underlined values denote the top choice given each starting point.
"A–h": high-income group of Society A;
"A–L": Society A with Low Social Mobility.
"D–": mobility is undefined, since high income = low income.
* Total < 100 due to retries or validation issues.

Table 3. Final society choices by starting position (Survey 3)

Future Work

This study highlights the potential of LLMs to simulate structured moral reasoning under uncertainty. However, LLMs still cannot fully replicate human reasoning, and their rationales may at times exhibit incoherence. As next steps, we plan to expand the scope of our experiments by systematically comparing LLM responses with human preferences across larger and more diverse participant samples. This will allow us to assess alignment not only in outcome distributions but also in underlying justifications. We also aim to explore multi-round simulations and interactive settings to better understand how LLM fairness preferences evolve with contextual feedback.