Preferences in AI Project Report

Pratik Deshmukh

September 26, 2025

1 Introduction

This report presents experiments on free-riding in sequential decision-making under different *statistical cultures*, following the framework of [1].

2 Background

2.1 Multi-Issue Model

We study sequential decision-making in *multi-issue elections*, where a set of voters must make a series of binary (yes/no) or multi-candidate decisions, one per issue. Each voter submits approval preferences for all candidates on each issue. A voting rule is then applied issue by issue to determine the collective outcome.

2.2 Voting Rules

We focus on two families of rules:

- **Sequential Utilitarian Rule:** selects in each issue the candidate with the highest total number of approvals.
- **Thiele-Based Rules:** a general class where voter satisfaction decreases marginally as more of their approved candidates are selected. Examples include:
  - *Proportional Approval Voting (PAV)* with harmonic weights $(1, \frac{1}{2}, \frac{1}{3}, \dots)$,
  - *Chamberlin–Courant (CC)* with weights $(1, 0, 0, \dots)$.
- **OWA-Based Rules:** aggregate voter satisfaction using Ordered Weighted Averages (OWAs). For instance:
  - *Leximin OWA:* maximizes the welfare of the worst-off voter,
  - *Mean OWA:* averages utilities across voters.

2.3 Statistical Cultures

The way preferences are generated strongly influences experimental outcomes. We consider four cultures:

- **p-IC**: per-issue impartial culture, sampling preferences independently with probability p.
- **Disjoint Groups**: voters are divided into g groups with internally aligned preferences.
- **Resampling Model**: preferences are generated by resampling with parameters (p, ϕ) controlling randomness and correlation.
- **Hamming Noise**: preferences are first generated from another culture and then perturbed by flipping approvals with small probability.

2.4 Welfare and Risk Metrics

We evaluate outcomes using two families of metrics:

- **Welfare Metrics:**
  - *Utilitarian welfare:* sum of utilities across all voters.
  - *Egalitarian welfare:* minimum utility among all voters.
  - *Nash welfare:* product of voter utilities (geometric balance).
- **Risk Metrics:**
  - *Successes:* number of manipulations that improved the manipulator's outcome.
  - *Harms:* number of manipulations that backfired on the manipulator.
  - *Success rate:* proportion of trials with a successful manipulation.
  - *Harm rate:* proportion of trials with a harmful manipulation.

3 Methodology

We repeat the experiments from the paper but use several different statistical cultures: impartial culture (p-IC), disjoint groups, the (p, ϕ)-resampling model, and the Hamming-noise model. For each, we run multiple seeds and compare welfare and risk metrics under sequential utilitarian, seq-PAV, and seq-CC rules.

3.1 Parameters

For transparency, we explicitly document the parameters used in our experiments:

- **p-IC**: resampling probabilities $p \in \{0.25, 0.5, 0.75\}$, with each issue sampled independently.
- **Disjoint**: voters are partitioned into $g = 4$ fixed groups, each with aligned preferences across all k issues.
- **Resampling**: combinations of (p, ϕ) include $(0.25, 0.25)$ and $(0.5, 0.75)$, where p is the resampling probability and ϕ controls correlation strength.
- **Hamming noise**: perturbations are introduced by flipping a voter's approval with probability 0.1 or 0.2 per issue, to model random noise in preferences.

For all cultures, we fix the number of voters $n = 50$, issues $k = 10$, and committee size $m = 5$. Each configuration is run with 10 different random seeds.

4 Results

The combined results table is automatically generated by the experiment pipeline. The table below is included directly from the output file:

4.1 Global Comparisons

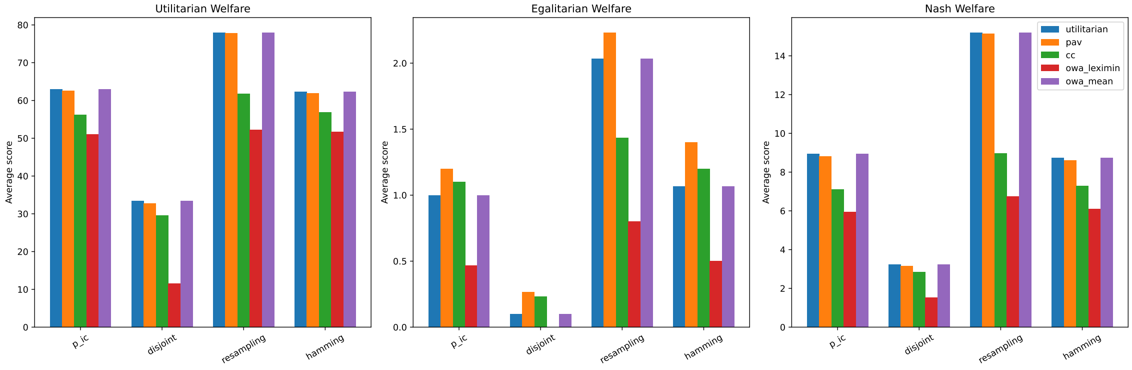

Figure 1: Welfare metrics across cultures and rules. Each subplot shows utilitarian, egalitarian, or Nash welfare.

culture	rule	seeds	Welfare			Risk				
			utilitarian	egalitarian	nash	trials	successes	harms	success_rate	harm_rate
p_ic	utilitarian	30	62.967	1.000	8.954	100.000	0.000	0.000	0.000	0.000
p_ic	pav	30	62.567	1.200	8.828	100.000	1.567	1.800	0.016	0.018
p_ic	cc	30	56.200	1.100	7.124	100.000	2.567	3.000	0.026	0.030
p_ic	owa_leximin	30	51.100	0.467	5.948	100.000	0.000	0.000	0.000	0.000
p_ic	owa_mean	30	62.967	1.000	8.954	100.000	0.000	0.000	0.000	0.000
disjoint	utilitarian	30	33.500	0.100	3.241	100.000	0.000	0.000	0.000	0.000
disjoint	pav	30	32.767	0.267	3.158	100.000	2.967	0.200	0.030	0.002
disjoint	cc	30	29.633	0.233	2.838	100.000	3.500	0.433	0.035	0.004
disjoint	owa_leximin	30	11.600	0.000	1.526	100.000	0.000	0.000	0.000	0.000
disjoint	owa_mean	30	33.500	0.100	3.241	100.000	0.000	0.000	0.000	0.000
resampling	utilitarian	30	78.033	2.033	15.212	100.000	0.000	0.000	0.000	0.000
resampling	pav	30	77.900	2.233	15.144	100.000	0.567	1.100	0.006	0.011
resampling	cc	30	61.767	1.433	8.984	100.000	0.467	0.633	0.005	0.006
resampling	owa_leximin	30	52.300	0.800	6.755	100.000	0.133	0.100	0.001	0.001
resampling	owa_mean	30	78.033	2.033	15.212	100.000	0.000	0.000	0.000	0.000
hamming	utilitarian	30	62.300	1.067	8.734	100.000	0.000	0.000	0.000	0.000
hamming	pav	30	61.867	1.400	8.618	100.000	1.433	1.867	0.014	0.019
hamming	cc	30	56.933	1.200	7.292	100.000	1.467	1.867	0.015	0.019
hamming	owa_leximin	30	51.733	0.500	6.092	100.000	0.000	0.000	0.000	0.000
hamming	owa_mean	30	62.300	1.067	8.734	100.000	0.000	0.000	0.000	0.000

Table 1: Combined results across cultures and rules. Welfare metrics are (utilitarian, egalitarian, Nash), while risk metrics include success and harm rates.

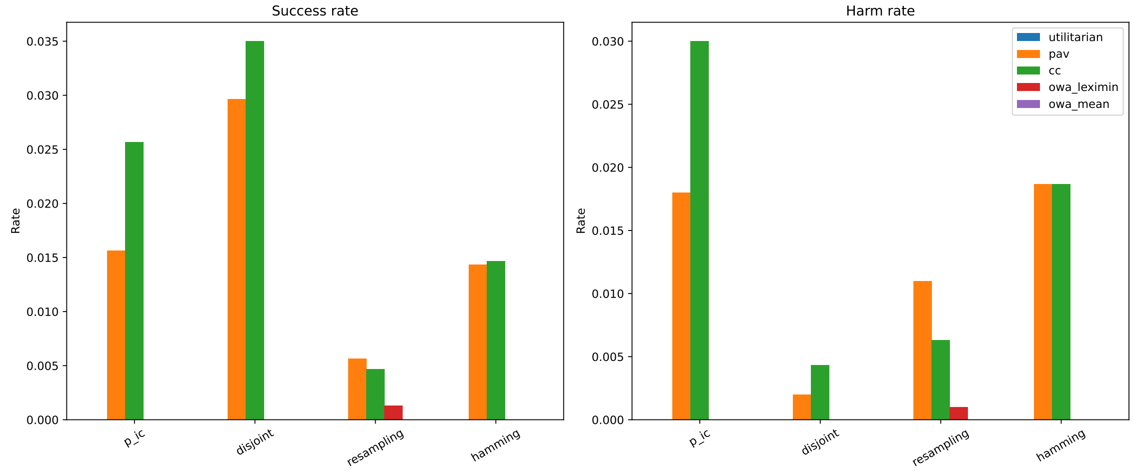

Figure 2: Manipulation risk (success and harm rates) across cultures and rules.

5 Discussion

Our experiments highlight the strong dependence of results on both the choice of voting rule and the statistical culture used to generate preferences.

Effect of Statistical Cultures. The *p-IC* culture yields relatively balanced welfare outcomes, but manipulation risks are non-negligible (see Fig. 1 and Fig. 2). The *disjoint* model shows much lower welfare, with preferences polarized across groups. The *resampling* model produces higher welfare but introduces manipulation opportunities. Finally, *Hamming noise* highlights robustness: small random flips can reduce manipulation opportunities but also slightly lower welfare.

Effect of Voting Rules. The *utilitarian rule* maximizes approvals but is highly exposed to manipulation. Thiele rules provide different trade-offs: *PAV* offers proportionality with some manipulation risk, while *CC* heavily rewards first approvals and thus reduces manipulation success at the cost of lower welfare. OWA rules show further trade-offs: *Leximin* minimizes risk for the worst-off voter, while the *mean OWA* balances welfare and risk.

Risk Metrics. Disjoint cultures create fewer opportunities for free-riding, while resampling and p-IC allow more. Harm rates are never negligible, confirming that voters attempting to free-ride risk worsening their outcomes.

6 Conclusion

Our experiments confirm that the choice of statistical culture strongly influences both welfare outcomes and robustness of voting rules to manipulation. Cultures such as disjoint groups lead to low welfare but reduced manipulation success, while resampling and p-IC generate higher welfare but greater susceptibility to free-riding. The introduction of Hamming noise shows that even small perturbations can alter both welfare and risk profiles, highlighting the fragility of certain rules.

Across voting rules, utilitarian aggregation achieves high welfare but exposes itself to manipulation. Thiele-based rules (PAV, CC) and OWA-based rules (leximin, mean) illustrate trade-offs: reducing manipulation success comes at the expense of utilitarian efficiency.

Future work should extend these experiments by exploring richer grids of parameters (p, ϕ), noise rates, and group structures, as well as alternative OWA weightings. Replicating the visualization style of [1] with plots would also allow direct side-by-side comparison of manipulation risk across rules.

Repository

The full project code and report sources are available at:
github.com/inquisitour/preferences-in-ai

List of Figures

A Per-Culture Figures

A.1 Per-Issue Impartial Culture (p-IC)

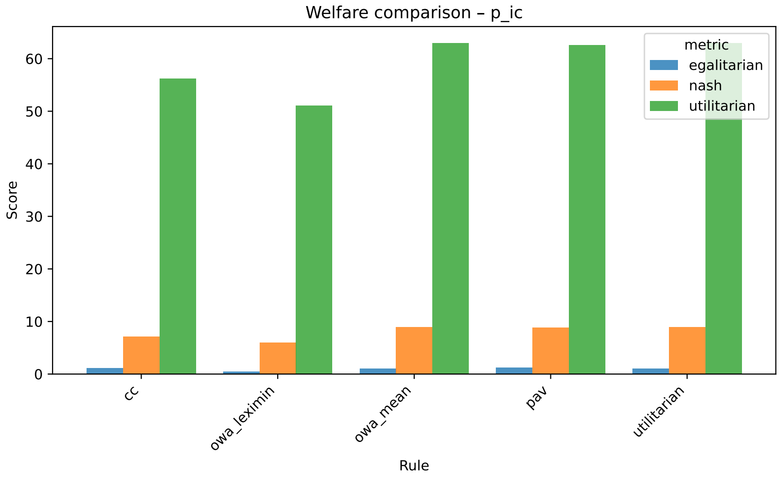

Figure 3: Welfare comparison for p-IC culture across voting rules.

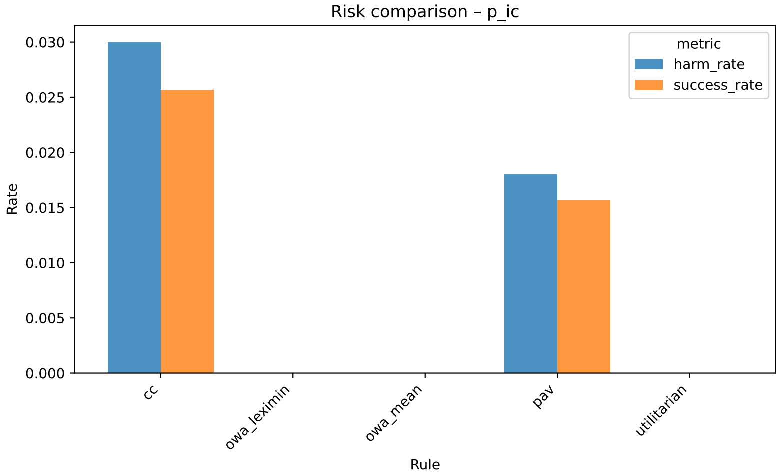

Figure 4: Manipulation risk under p-IC culture.

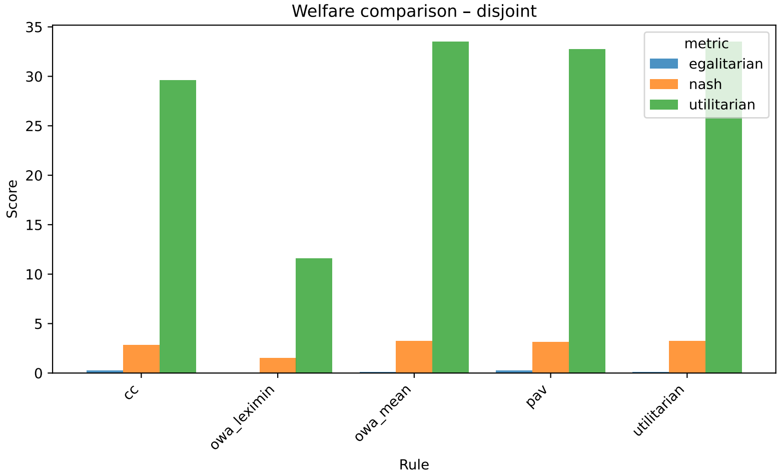

Figure 5: Welfare comparison for disjoint-group culture.

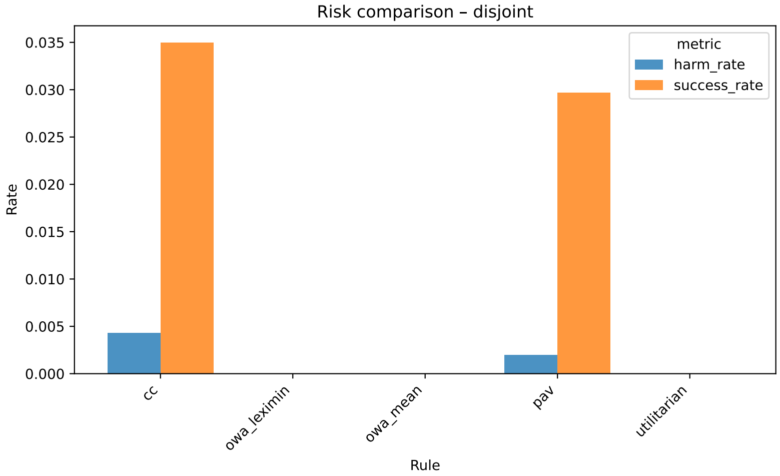

Figure 6: Manipulation risk under disjoint-group culture.

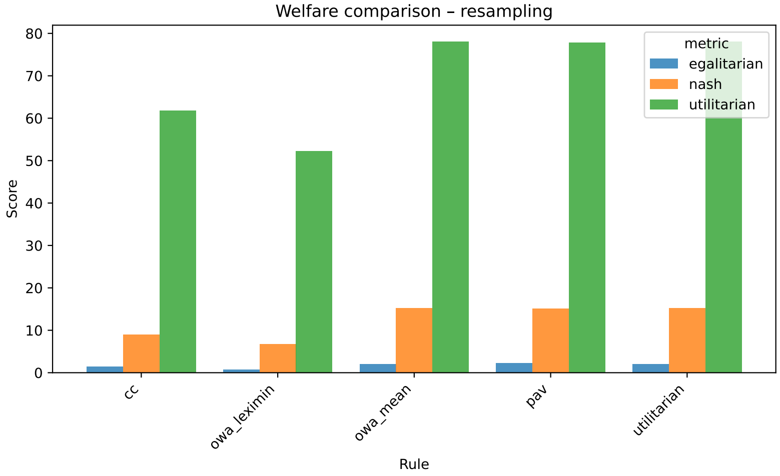

Figure 7: Welfare comparison for resampling culture.

A.2 Disjoint Groups

A.3 Resampling Model

A.4 Hamming Noise

B Code Documentation Summary

- **core/** – Base types, welfare functions, and voting rule interface.

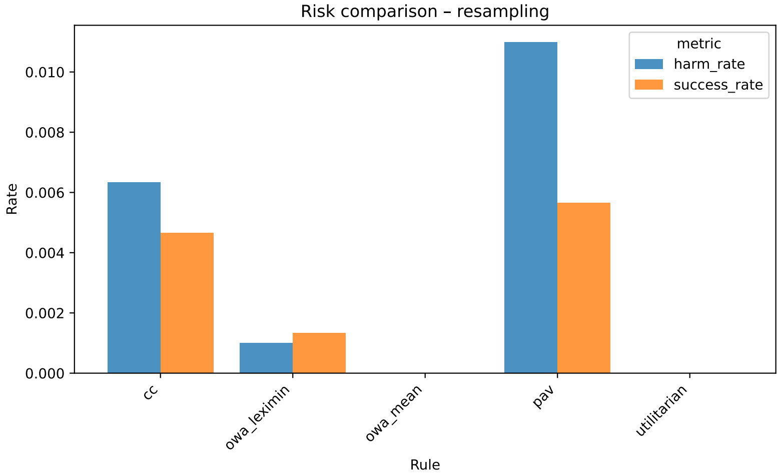

Figure 8: Manipulation risk under resampling culture.

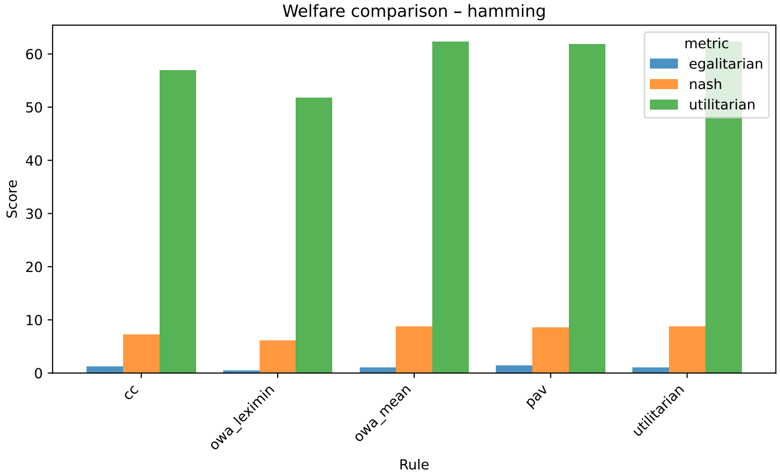

Figure 9: Welfare comparison for Hamming-noise culture.

- **statistical_cultures/** – Preference generators (p-IC, disjoint, resampling, hamming noise).
- **voting_rules/** – Implementations of utilitarian, PAV, CC, and OWA rules.
- **free_riding/** – Manipulation detector, welfare/risk analysis tools.
- **experiments/** – Main experiment runner for end-to-end evaluation.

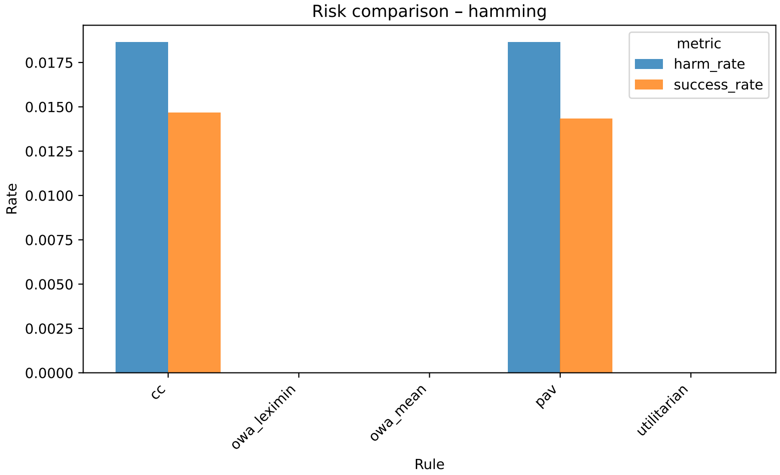

Figure 10: Manipulation risk under Hamming-noise culture.

- **tests/** – Unit tests for cultures, rules, and free-riding checks.

References

[1] Martin Lackner, Jan Maly, and Oliviero Nardi. Free-riding in multi-issue decisions. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 2040–2048. International Foundation for Autonomous Agents and Multiagent Systems, 2023. Also available as arXiv preprint: arXiv:2310.08194.