Preferences in AI Project Report

Pratik Deshmukh

October 4, 2025

1 Introduction

This report presents experiments on free-riding in sequential decision-making under different *statistical cultures*, following the framework of [1]. Our goal is to replicate the structure of Section 5 of that work while extending it with additional statistical cultures and updated parameter settings.

2 Background

2.1 Multi-Issue Model

We study sequential decision-making in *multi-issue elections*, where a set of voters must decide on several issues, each with multiple candidates. 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 major families of rules, following [1]:

- **Sequential Utilitarian Rule:** selects in each issue the candidate with the highest total number of approvals. This rule is equivalent to the mean OWA and is known to be immune to manipulation.
- **Thiele-Based Rules:** a general class where voter satisfaction decreases marginally as more of their approved candidates are selected. We evaluate sequential Thiele rules with parameters $x \in \{1, 5, 7\}$, where $x = 0$ corresponds to utilitarian aggregation.
- **OWA-Based Rules:** aggregate voter satisfaction using Ordered Weighted Averages (OWAs). Following [1], we use normalized positive weights (no zeros) interpolating between utilitarian and leximin behavior. We evaluate parametric OWA rules with $x \in \{1, 5, 10, 15\}$, and include the explicit *leximin OWA* as the limiting case.

2.3 Statistical Cultures

The way preferences are generated strongly influences manipulation risks. We consider four cultures:

- **p-IC**: per-issue impartial culture, sampling approvals independently with probability $p = 0.5$.
- **Disjoint Groups**: voters are divided into $g = 2$ groups with internally aligned preferences.
- **Resampling Model**: preferences are generated by resampling with parameters $(p, \phi) = (0.5, 0.5)$ controlling randomness and correlation.
- **Hamming Noise**: preferences are first generated from another culture and then perturbed by flipping approvals with small probability $\epsilon = 0.1$ per issue. This noise model captures robustness under small random perturbations.

2.4 Risk Metrics

We evaluate manipulation opportunities using the following metrics:

- **Trials:** total number of manipulation attempts.
- **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.
- **Risk:** ratio of harms to successes (conditional probability of harmful manipulation).

3 Methodology

We replicate the experiments from Section 5 of [1], using four statistical cultures: impartial culture (p-IC), disjoint groups, the (p, ϕ)-resampling model, and the Hamming-noise model. For each culture, we run multiple random seeds and compare risk metrics under sequential utilitarian, sequential Thiele rules ($x = 1, 5, 7$), and OWA rules ($x = 1, 5, 10, 15$, plus leximin).

3.1 Parameters

The main parameters used in our experiments are:

- Number of voters: $n = 20$
- Number of issues: $k = 5$
- Candidates per issue: $c = 4$
- Random seeds: 200
- Cultures and hyperparameters: as defined in the previous subsection.

All configurations were executed using our unified experimental pipeline, which produces both tabular summaries and risk plots.

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 Per-Culture Comparisons

To visualize the trends, we show manipulation risks by rule family (Thiele and OWA) for each statistical culture.

p-IC. Under p-IC, both Thiele and OWA families show mild manipulation rates. OWA rules exhibit slightly lower success and harm rates, while leximin remains largely immune. This matches the expected robustness of random independent preferences.

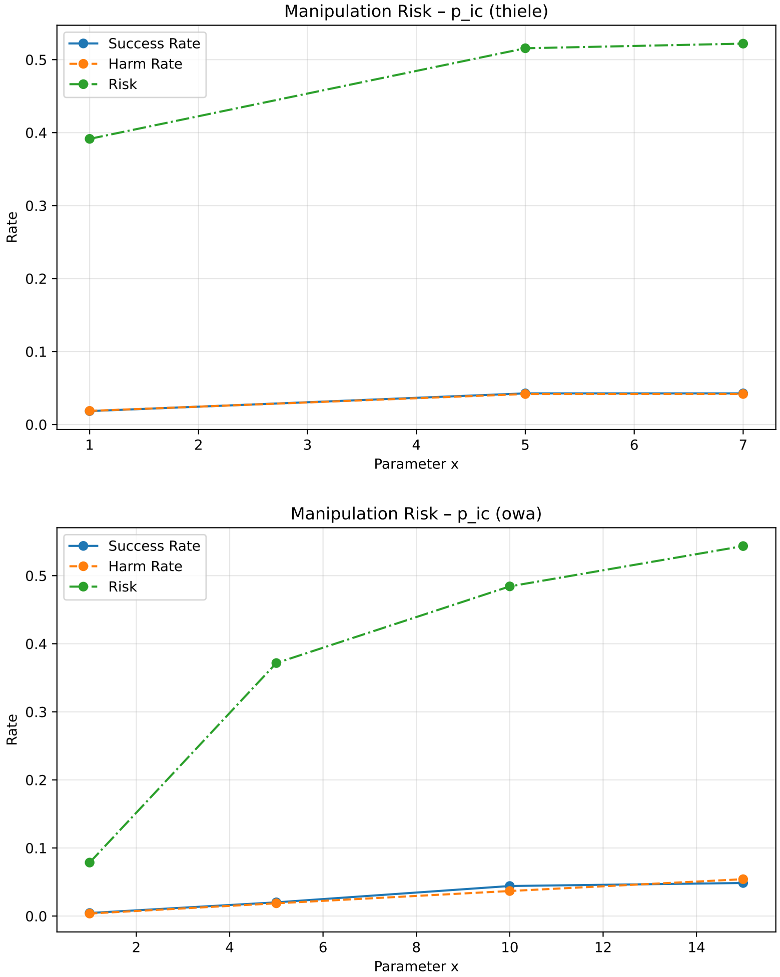
Disjoint Groups. Disjoint group structures introduce correlated preferences. Manipulation success increases slightly for Thiele rules, but harm rates stay very low, indicating that free-riding is relatively safe when preferences are clustered.

Resampling Model. The resampling culture yields intermediate results between p-IC and disjoint. Thiele and OWA rules exhibit low but nonzero manipulation potential. Higher x values slightly increase risk, consistent with greater proportionality.

Hamming Noise. Hamming perturbations add random noise to structured preferences. As expected, this causes mild increases in both success and harm rates. Leximin remains mostly stable, confirming its resilience to random perturbations.

culture	rule	seeds	trials	successes	harms	success_rate	harm_rate	risk
p_ic	utilitarian	200	100.000	0.000	0.000	0.000	0.000	0.000
p_ic	thiele_x1	200	100.000	1.840	1.870	0.018	0.019	0.391
p_ic	thiele_x5	200	100.000	4.270	4.180	0.043	0.042	0.516
p_ic	thiele_x7	200	100.000	4.270	4.185	0.043	0.042	0.522
p_ic	owa_x1	200	100.000	0.445	0.365	0.004	0.004	0.079
p_ic	owa_x5	200	100.000	2.010	1.870	0.020	0.019	0.372
p_ic	owa_x10	200	100.000	4.410	3.675	0.044	0.037	0.484
p_ic	owa_x15	200	100.000	4.855	5.410	0.049	0.054	0.543
p_ic	owa_leximin	200	100.000	4.935	4.920	0.049	0.049	0.508
disjoint	utilitarian	200	100.000	0.000	0.000	0.000	0.000	0.000
disjoint	thiele_x1	200	100.000	2.265	0.190	0.023	0.002	0.031
disjoint	thiele_x5	200	100.000	2.715	0.380	0.027	0.004	0.081
disjoint	thiele_x7	200	100.000	2.725	0.380	0.027	0.004	0.081
disjoint	owa_x1	200	100.000	0.815	0.045	0.008	0.000	0.012
disjoint	owa_x5	200	100.000	3.435	0.375	0.034	0.004	0.070
disjoint	owa_x10	200	100.000	3.885	0.600	0.039	0.006	0.103
disjoint	owa_x15	200	100.000	3.200	0.655	0.032	0.007	0.118
disjoint	owa_leximin	200	100.000	3.745	0.920	0.037	0.009	0.139
resampling	utilitarian	200	100.000	0.000	0.000	0.000	0.000	0.000
resampling	thiele_x1	200	100.000	0.465	0.735	0.005	0.007	0.209
resampling	thiele_x5	200	100.000	1.430	1.625	0.014	0.016	0.376
resampling	thiele_x7	200	100.000	1.450	1.505	0.015	0.015	0.391
resampling	owa_x1	200	100.000	0.000	0.020	0.000	0.000	0.005
resampling	owa_x5	200	100.000	0.495	0.280	0.005	0.003	0.082
resampling	owa_x10	200	100.000	0.790	1.440	0.008	0.014	0.263
resampling	owa_x15	200	100.000	1.950	3.625	0.020	0.036	0.383
resampling	owa_leximin	200	100.000	1.765	2.415	0.018	0.024	0.477
hamming	utilitarian	200	100.000	0.000	0.000	0.000	0.000	0.000
hamming	thiele_x1	200	100.000	2.120	2.185	0.021	0.022	0.448
hamming	thiele_x5	200	100.000	3.870	3.975	0.039	0.040	0.530
hamming	thiele_x7	200	100.000	3.940	3.970	0.039	0.040	0.521
hamming	owa_x1	200	100.000	0.335	0.430	0.003	0.004	0.098
hamming	owa_x5	200	100.000	2.030	2.295	0.020	0.023	0.420
hamming	owa_x10	200	100.000	3.830	4.065	0.038	0.041	0.498
hamming	owa_x15	200	100.000	4.890	4.810	0.049	0.048	0.501
hamming	owa_leximin	200	100.000	4.735	5.130	0.047	0.051	0.560

Table 1: Combined results across cultures and rules. Risk metrics include trials, successes, harms, success and harm rates, and risk (harms/successes).


Manipulation Risk – p_ic (thiele)

Manipulation Risk – p_ic (owa)

Figure 1: Manipulation risk under p-IC culture. Top: Thiele rules. Bottom: OWA rules.

5 Discussion

Our experiments highlight the dependence of manipulation risk on both the voting rule and the statistical culture.

Effect of Statistical Cultures. The *p-IC* culture exhibits moderate manipulation opportunities. The *disjoint* model produces clearer group structures, increasing the chance that one group can manipulate effectively. The *resampling* model shows behavior similar to p-IC but slightly more correlated outcomes, while *Hamming noise* introduces random fluctuations that may both increase or reduce risks depending on the rule.

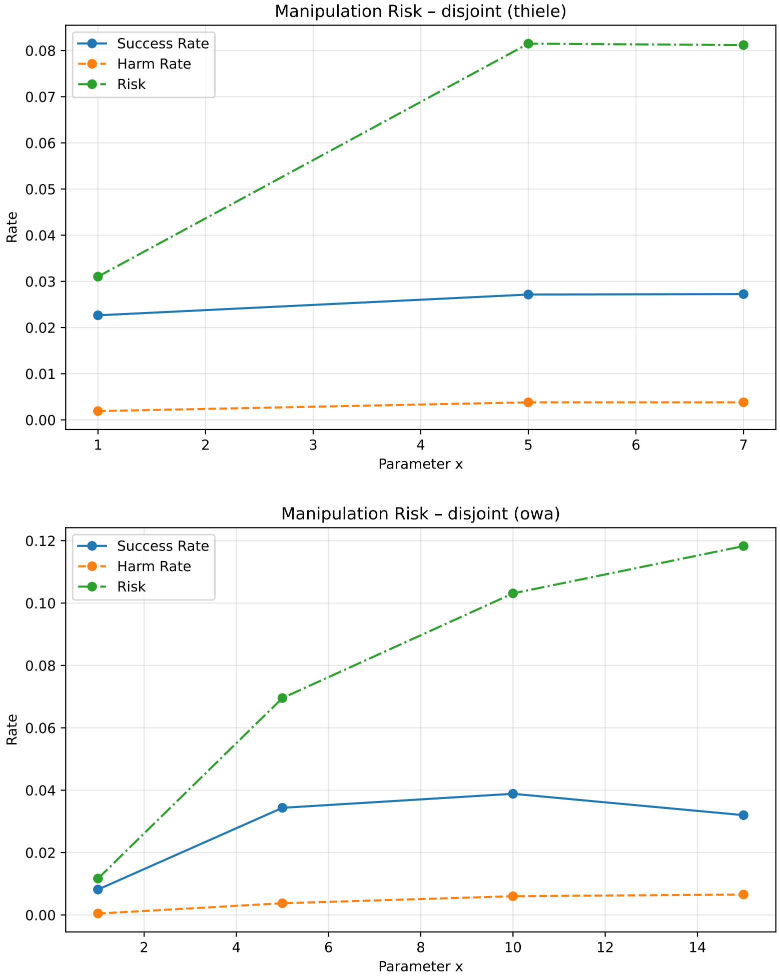


Figure 2: Manipulation risk under disjoint-group culture. Top: Thiele rules. Bottom: OWA rules.

Effect of Voting Rules. The *utilitarian rule* (equivalently mean OWA) is immune to manipulation. Thiele rules with larger x increase proportionality but also marginally increase free-riding potential. OWA rules interpolate between utilitarian and leximin: intermediate parameters show moderate manipulation risks, while the *leximin rule* remains the most resistant to manipulation, in line with our results (very low success and harm rates).

Risk Metrics. Across all settings, harm rates remain small but nonzero, confirming that attempts at manipulation carry measurable risks. The ratio of harms to successes (*risk*)

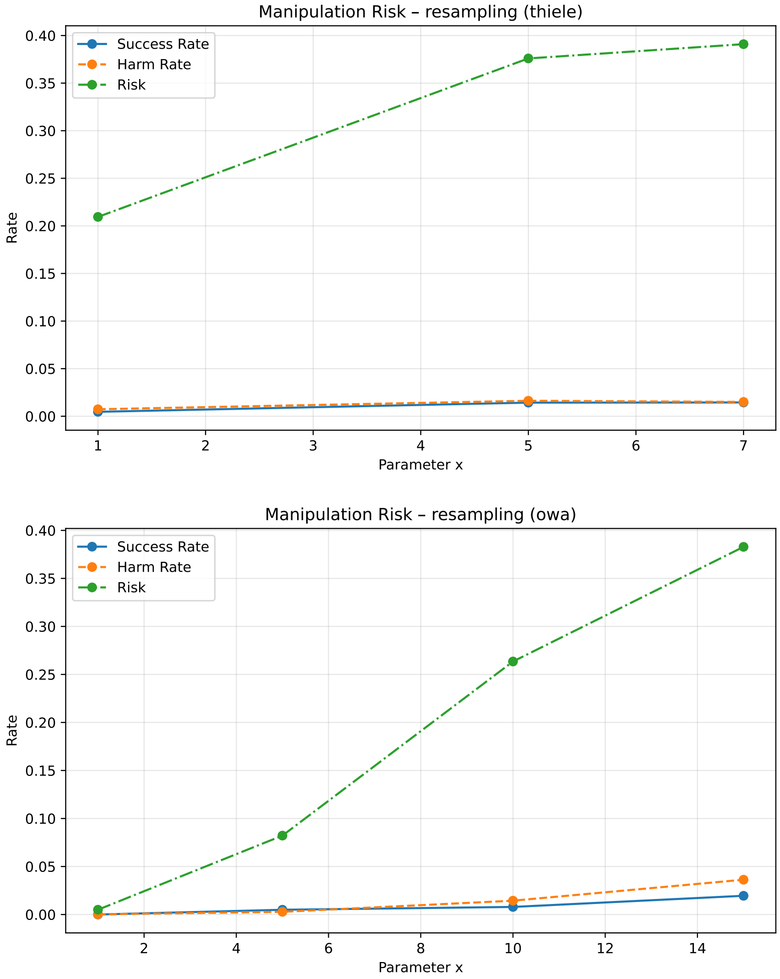


Figure 3: Manipulation risk under resampling culture. Top: Thiele rules. Bottom: OWA rules.

stays well below 1, indicating that manipulators more often succeed than fail—but not without potential downsides.

6 Conclusion

Our experiments confirm that the choice of statistical culture strongly influences manipulation risks. Cultures with correlated preferences (e.g., disjoint, resampling) can increase opportunities for manipulation, while random noise can unpredictably alter them.

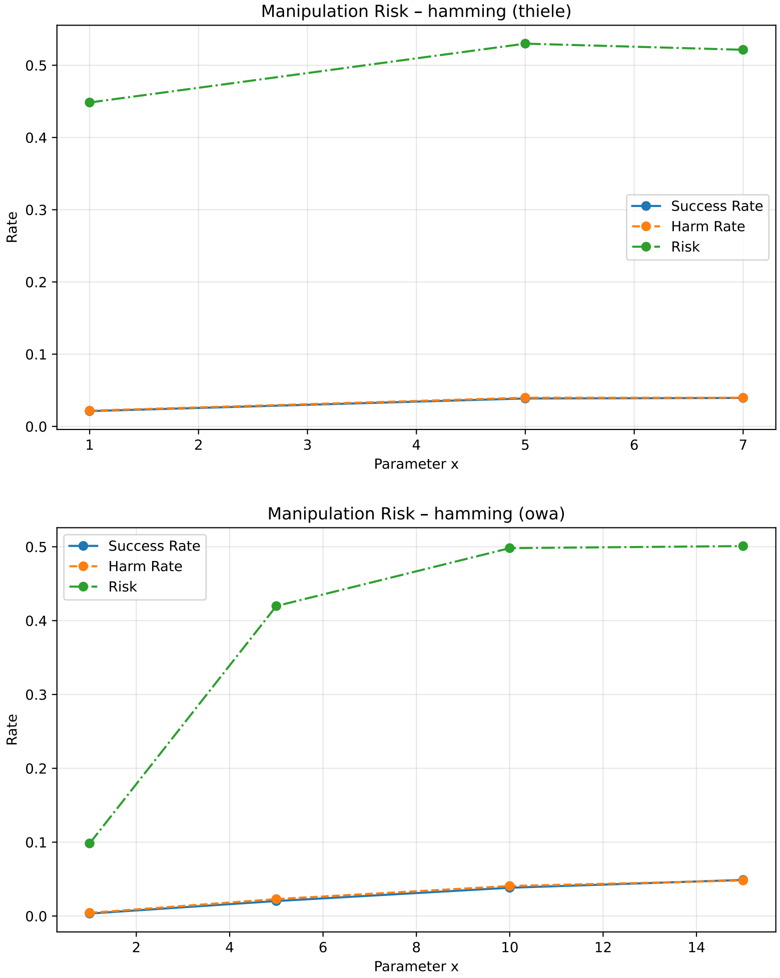


Figure 4: Manipulation risk under Hamming-noise culture. Top: Thiele rules. Bottom: OWA rules.

Across voting rules, utilitarian aggregation is immune. Thiele and OWA rules illustrate clear trade-offs: increasing fairness or proportionality often slightly increases manipulation risk, but leximin consistently shows the highest robustness. These findings align closely with the theoretical and empirical trends observed by [1].

Future work could extend these experiments by exploring richer parameter grids for (p, ϕ) and ϵ, as well as scaling to larger electorates or mixed-issue dependencies. Replicating the visualization style of [1] for welfare and risk together would further improve interpretability.

Repository

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

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.