Preferences in AI Project Report

Pratik Deshmukh

September 28, 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.

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 families of rules, as in [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 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 the case $x = 0$ corresponds to utilitarian aggregation.

- **OWA-Based Rules:** aggregate voter satisfaction using Ordered Weighted Averages (OWAs). We evaluate parametric OWA rules with $x \in \{1, 5, 10, 15\}$, interpolating between utilitarian ($x = 0$) and leximin ($x = n - 1$). For completeness, we also include the explicit *leximin OWA*.

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.
- **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. This noise model is slightly different from that used in [1], but serves the same purpose of modeling robustness under 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.

3 Methodology

We repeat 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, we run multiple 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

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

- **p-IC**: approval probability $p = 0.5$.
- **Disjoint**: voters are partitioned into $g = 2$ groups with aligned preferences.
- **Resampling**: parameters $(p, \phi) = (0.5, 0.5)$, consistent with the disjoint model.
- **Hamming noise**: perturbations introduced by flipping a voter's approval with probability 0.1 per issue.

For all cultures, we fix the number of voters $n = 20$, issues $k = 5$, and $c = 4$ candidates per issue. Each configuration is run with 30 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 Per-Culture Comparisons

p-IC.

Disjoint Groups.

Resampling Model.

Hamming Noise.

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, sometimes increasing manipulation success rates. The *resampling* model yields non-negligible risks depending on ϕ, and the *Hamming noise* experiments show that random perturbations can both increase and decrease manipulation opportunities, depending on the rule.

culture	rule	seeds	trials	successes	harms	success_rate	harm_rate	risk
p_ic	utilitarian	30	100.000	0.000	0.000	0.000	0.000	0.000
p_ic	thiele_x1	30	100.000	1.567	1.800	0.016	0.018	0.557
p_ic	thiele_x5	30	100.000	5.333	4.300	0.053	0.043	1.046
p_ic	thiele_x7	30	100.000	5.167	4.167	0.052	0.042	1.033
p_ic	owa_x1	30	100.000	0.000	0.000	0.000	0.000	0.000
p_ic	owa_x5	30	100.000	2.400	2.000	0.024	0.020	0.114
p_ic	owa_x10	30	100.000	0.233	0.633	0.002	0.006	0.000
p_ic	owa_x15	30	100.000	0.000	0.000	0.000	0.000	0.000
p_ic	owa_leximin	30	100.000	0.000	0.000	0.000	0.000	0.000
disjoint	utilitarian	30	100.000	0.000	0.000	0.000	0.000	0.000
disjoint	thiele_x1	30	100.000	2.967	0.200	0.030	0.002	0.063
disjoint	thiele_x5	30	100.000	4.233	0.533	0.042	0.005	0.112
disjoint	thiele_x7	30	100.000	4.367	0.533	0.044	0.005	0.107
disjoint	owa_x1	30	100.000	0.000	0.000	0.000	0.000	0.000
disjoint	owa_x5	30	100.000	0.200	0.000	0.002	0.000	0.000
disjoint	owa_x10	30	100.000	0.733	0.000	0.007	0.000	0.000
disjoint	owa_x15	30	100.000	3.233	0.000	0.032	0.000	0.000
disjoint	owa_leximin	30	100.000	0.000	0.000	0.000	0.000	0.000
resampling	utilitarian	30	100.000	0.000	0.000	0.000	0.000	0.000
resampling	thiele_x1	30	100.000	0.567	1.100	0.006	0.011	0.317
resampling	thiele_x5	30	100.000	1.300	1.933	0.013	0.019	0.343
resampling	thiele_x7	30	100.000	1.267	1.467	0.013	0.015	0.410
resampling	owa_x1	30	100.000	0.400	1.700	0.004	0.017	0.167
resampling	owa_x5	30	100.000	1.633	1.933	0.016	0.019	0.000
resampling	owa_x10	30	100.000	0.000	0.000	0.000	0.000	0.000
resampling	owa_x15	30	100.000	0.000	0.033	0.000	0.000	0.000
resampling	owa_leximin	30	100.000	0.000	0.000	0.000	0.000	0.000
hamming	utilitarian	30	100.000	0.000	0.000	0.000	0.000	0.000
hamming	thiele_x1	30	100.000	1.433	1.867	0.014	0.019	0.818
hamming	thiele_x5	30	100.000	2.900	3.667	0.029	0.037	1.191
hamming	thiele_x7	30	100.000	2.867	3.800	0.029	0.038	1.216
hamming	owa_x1	30	100.000	0.000	0.000	0.000	0.000	0.000
hamming	owa_x5	30	100.000	1.600	2.400	0.016	0.024	0.187
hamming	owa_x10	30	100.000	0.000	0.467	0.000	0.005	0.000
hamming	owa_x15	30	100.000	0.000	0.000	0.000	0.000	0.000
hamming	owa_leximin	30	100.000	0.000	0.000	0.000	0.000	0.000

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

Effect of Voting Rules. The *utilitarian rule* (equivalently mean OWA) is immune to manipulation. Thiele rules with larger x increase proportionality but are more exposed to free-riding opportunities. OWA rules interpolate between utilitarian and leximin: interme-

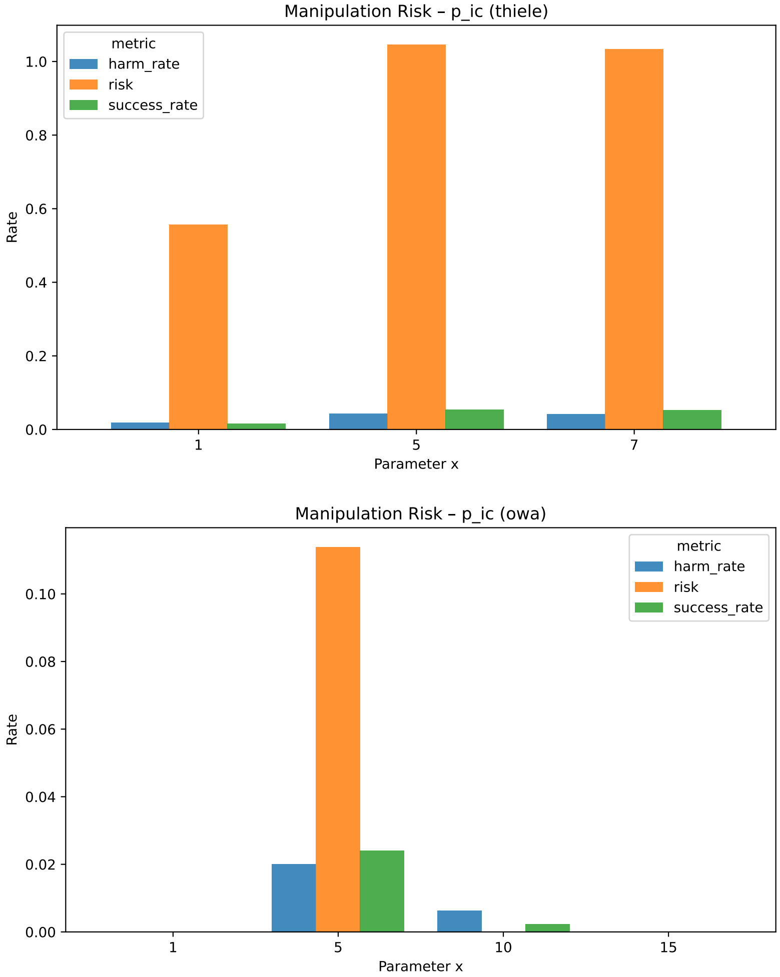

Figure 1: Manipulation risk under p-IC culture. Top: Thiele rules. Bottom: OWA rules.

diate parameters show manipulation risks, while leximin tends to reduce success rates but sometimes increases harms.

Risk Metrics. Across all settings, harm rates are non-negligible, confirming that attempts at manipulation carry real risks. In particular, when manipulation succeeds, it is often accompanied by harmful attempts that worsen outcomes for manipulators.

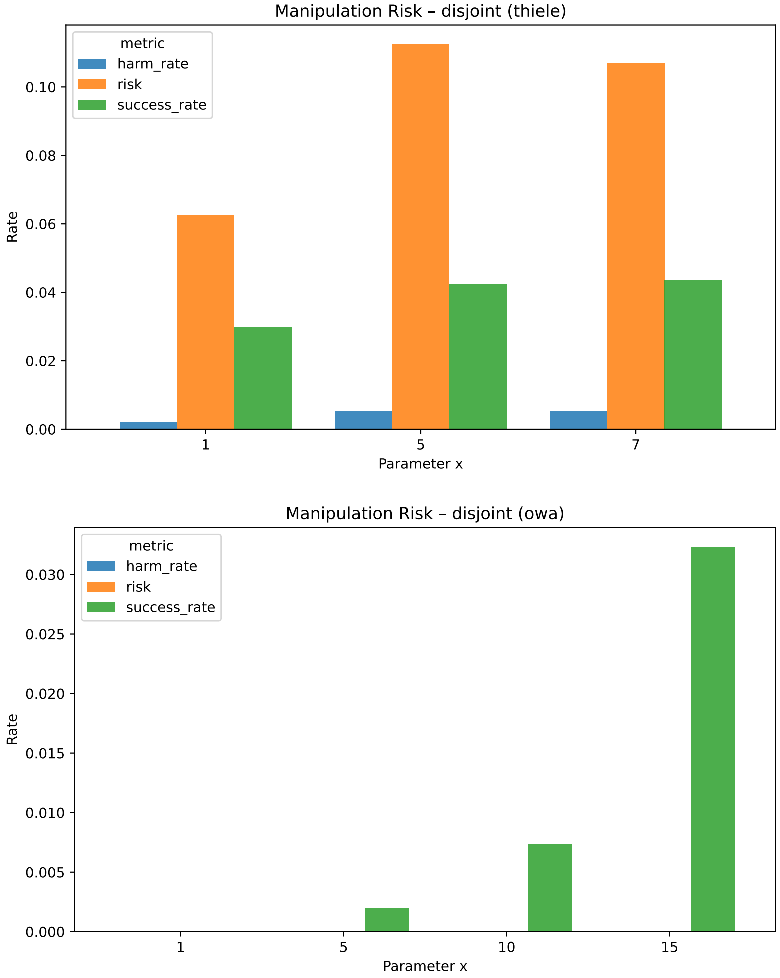

Figure 2: Manipulation risk under disjoint-group culture. Top: Thiele rules. Bottom: OWA rules.

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.

Across voting rules, utilitarian aggregation is immune. Thiele-based rules and OWA-based rules illustrate trade-offs: increasing proportionality or fairness typically comes at the expense of higher manipulation risk.

Future work should extend these experiments by exploring richer grids of parameters

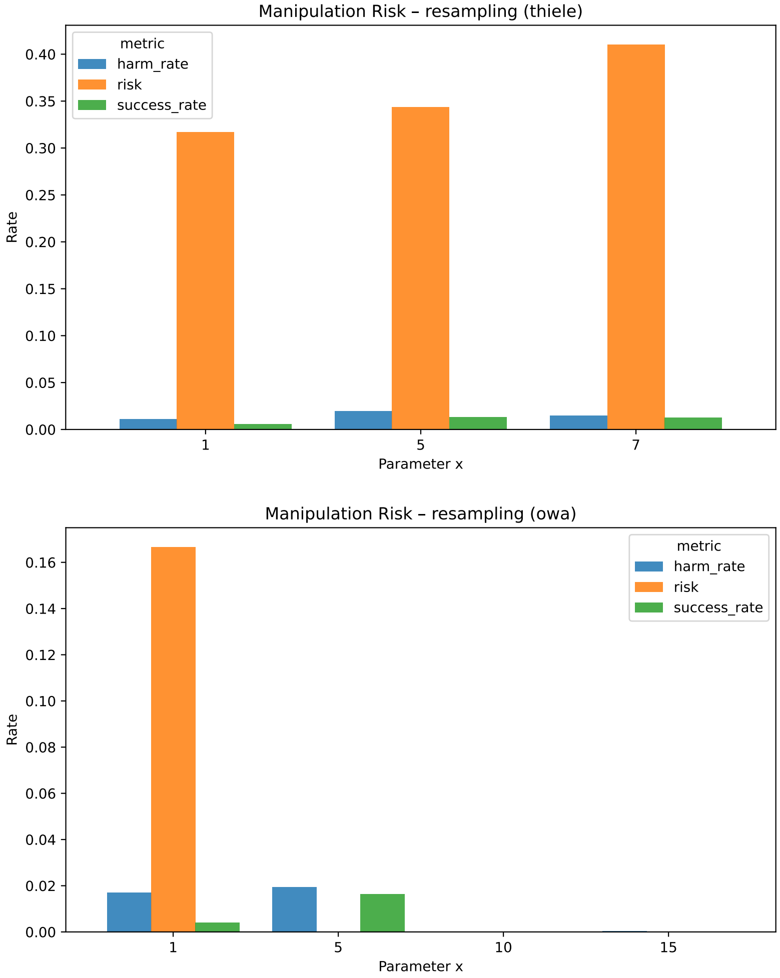

Figure 3: Manipulation risk under resampling culture. Top: Thiele rules. Bottom: OWA rules.

(p, ϕ), noise rates, and group structures. Replicating the visualization style of [1] with further plots would also allow direct side-by-side comparison of manipulation risks across rules.

Repository

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

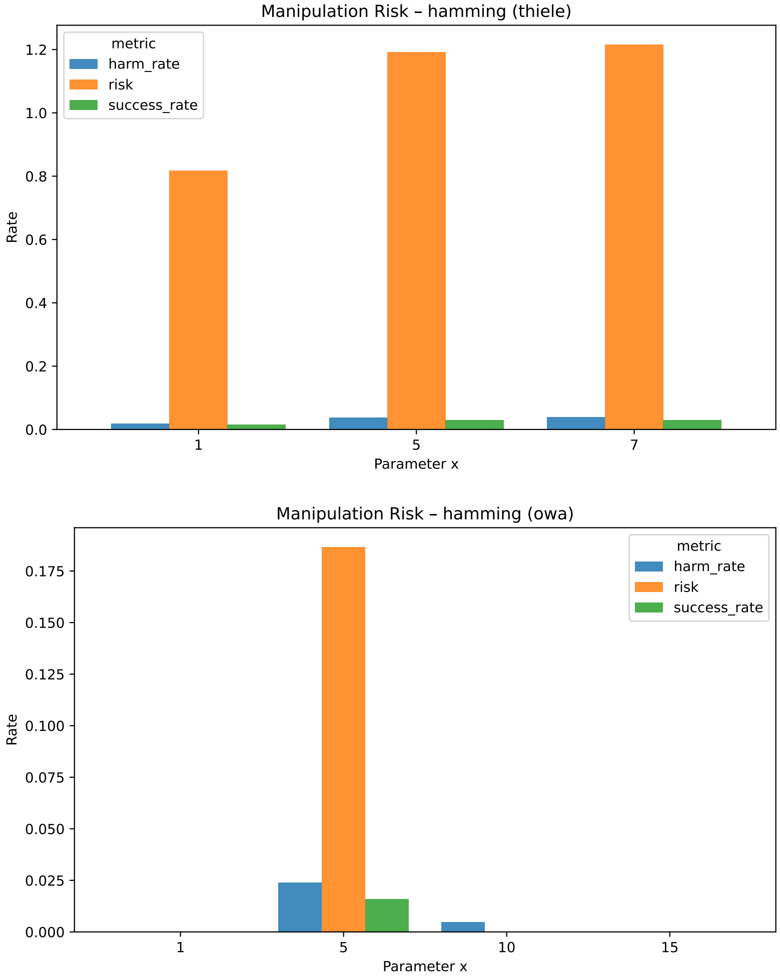

Figure 4: Manipulation risk under Hamming-noise culture. Top: Thiele rules. Bottom: OWA rules.

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.