# Preferences in AI Project Report

Pratik Deshmukh

October 7, 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 an explicit *leximin OWA* limit.

### 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 Free-Riding Notion Used in the Experiments

We adopt the paper's definition and apply the following operationalization: (i) a voter can attempt to free-ride on issue $i$ only if she *originally approved the winning candidate* on $i$; (ii) to test manipulation, we replace the ballot on issue $i$ by a *restricted* deviation that drops that single approval (all other approvals are kept fixed), and we require that the winner on issue $i$ *remains unchanged* (the voter is non-pivotal on $i$); (iii) the election outcome is recomputed using the manipulated profile, but the voter's gain/loss is assessed with her *truthful* utilities. This deviation class is a subset of all free-riding deviations discussed in [1] but is sufficient for our experiments.

### 2.5 Risk Metrics

We evaluate manipulation opportunities using the following metrics:

- **Trials:** total number of voter–issue pairs considered ($n \times k$).
- **Eligible:** # pairs where the voter originally approved the winner.
- **Possible:** # eligible pairs where dropping that single approval leaves the winner unchanged.
- **Successes:** # possible pairs where the voter's truthful utility increases.
- **Harms:** # possible pairs where the voter's truthful utility decreases.
- **Success rate:** successes/trials.

- **Harm rate:** harms/trials.
- **Risk:** harms/possible (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 $(p, \phi)$-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 above.

All configurations were executed using a 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, the utilitarian rule is immune (zero across all metrics). Within Thiele and OWA families, success rates are small but increase with the family parameter $x$; harms remain tiny, and risk (harms/possible) is near zero. In our runs, the leximin extreme shows the largest success within OWA, consistent with the trend that moving away from utilitarian slightly increases exploitable opportunities.

| culture | rule | seeds | trials | eligible | possible | successes | harms | success_rate | harm_rate | risk |
|---|---|---|---|---|---|---|---|---|---|---|
| p_ic | utilitarian | 200 | 100.000 | 61.785 | 39.480 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| p_ic | thiele_x1 | 200 | 100.000 | 61.355 | 40.085 | 3.185 | 0.010 | 0.032 | 0.000 | 0.001 |
| p_ic | thiele_x5 | 200 | 100.000 | 58.950 | 41.125 | 6.745 | 0.030 | 0.067 | 0.000 | 0.001 |
| p_ic | thiele_x7 | 200 | 100.000 | 58.930 | 41.220 | 6.850 | 0.030 | 0.068 | 0.000 | 0.001 |
| p_ic | owa_x1 | 200 | 100.000 | 61.785 | 39.485 | 0.085 | 0.000 | 0.001 | 0.000 | 0.000 |
| p_ic | owa_x5 | 200 | 100.000 | 61.505 | 40.050 | 1.905 | 0.010 | 0.019 | 0.000 | 0.000 |
| p_ic | owa_x10 | 200 | 100.000 | 60.315 | 39.860 | 5.310 | 0.075 | 0.053 | 0.001 | 0.002 |
| p_ic | owa_x15 | 200 | 100.000 | 59.040 | 41.195 | 7.145 | 0.130 | 0.071 | 0.001 | 0.003 |
| p_ic | owa_leximin | 200 | 100.000 | 57.845 | 41.960 | 7.460 | 0.095 | 0.075 | 0.001 | 0.002 |
| disjoint | utilitarian | 200 | 100.000 | 34.530 | 28.045 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| disjoint | thiele_x1 | 200 | 100.000 | 33.835 | 28.275 | 1.495 | 0.000 | 0.015 | 0.000 | 0.000 |
| disjoint | thiele_x5 | 200 | 100.000 | 33.085 | 29.210 | 1.910 | 0.010 | 0.019 | 0.000 | 0.000 |
| disjoint | thiele_x7 | 200 | 100.000 | 33.085 | 29.210 | 1.920 | 0.010 | 0.019 | 0.000 | 0.000 |
| disjoint | owa_x1 | 200 | 100.000 | 34.530 | 28.190 | 0.140 | 0.000 | 0.001 | 0.000 | 0.000 |
| disjoint | owa_x5 | 200 | 100.000 | 33.550 | 28.625 | 1.675 | 0.000 | 0.017 | 0.000 | 0.000 |
| disjoint | owa_x10 | 200 | 100.000 | 33.110 | 28.875 | 2.070 | 0.000 | 0.021 | 0.000 | 0.000 |
| disjoint | owa_x15 | 200 | 100.000 | 32.880 | 29.245 | 1.995 | 0.010 | 0.020 | 0.000 | 0.000 |
| disjoint | owa_leximin | 200 | 100.000 | 32.600 | 29.005 | 2.030 | 0.025 | 0.020 | 0.000 | 0.001 |
| resampling | utilitarian | 200 | 100.000 | 77.450 | 57.490 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| resampling | thiele_x1 | 200 | 100.000 | 77.335 | 57.000 | 2.180 | 0.000 | 0.022 | 0.000 | 0.000 |
| resampling | thiele_x5 | 200 | 100.000 | 75.610 | 58.280 | 6.030 | 0.000 | 0.060 | 0.000 | 0.000 |
| resampling | thiele_x7 | 200 | 100.000 | 75.325 | 58.380 | 6.430 | 0.005 | 0.064 | 0.000 | 0.000 |
| resampling | owa_x1 | 200 | 100.000 | 77.450 | 57.390 | 0.035 | 0.000 | 0.000 | 0.000 | 0.000 |
| resampling | owa_x5 | 200 | 100.000 | 77.400 | 57.010 | 0.610 | 0.000 | 0.006 | 0.000 | 0.000 |
| resampling | owa_x10 | 200 | 100.000 | 77.075 | 57.840 | 3.290 | 0.040 | 0.033 | 0.000 | 0.001 |
| resampling | owa_x15 | 200 | 100.000 | 76.005 | 58.125 | 6.025 | 0.005 | 0.060 | 0.000 | 0.000 |
| resampling | owa_leximin | 200 | 100.000 | 74.705 | 58.850 | 7.395 | 0.010 | 0.074 | 0.000 | 0.000 |
| hamming | utilitarian | 200 | 100.000 | 61.405 | 38.255 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| hamming | thiele_x1 | 200 | 100.000 | 60.905 | 39.890 | 3.235 | 0.035 | 0.032 | 0.000 | 0.001 |
| hamming | thiele_x5 | 200 | 100.000 | 58.590 | 41.935 | 6.165 | 0.030 | 0.062 | 0.000 | 0.001 |
| hamming | thiele_x7 | 200 | 100.000 | 58.520 | 41.885 | 6.245 | 0.035 | 0.062 | 0.000 | 0.001 |
| hamming | owa_x1 | 200 | 100.000 | 61.405 | 38.335 | 0.115 | 0.005 | 0.001 | 0.000 | 0.000 |
| hamming | owa_x5 | 200 | 100.000 | 61.160 | 38.625 | 1.950 | 0.015 | 0.019 | 0.000 | 0.000 |
| hamming | owa_x10 | 200 | 100.000 | 59.825 | 40.235 | 5.180 | 0.020 | 0.052 | 0.000 | 0.000 |
| hamming | owa_x15 | 200 | 100.000 | 58.580 | 42.070 | 6.805 | 0.055 | 0.068 | 0.001 | 0.001 |
| hamming | owa_leximin | 200 | 100.000 | 57.875 | 42.790 | 6.690 | 0.045 | 0.067 | 0.000 | 0.001 |

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

**Disjoint Groups.** Correlated group structure reduces overall opportunities compared to p-IC. Success rates remain low across both families; harms are close to zero and risk is negligible.

**Resampling Model.** Resampling yields patterns similar to p-IC: utilitarian is immune; success rates increase modestly with $x$ in both families; harms and risk remain very small.

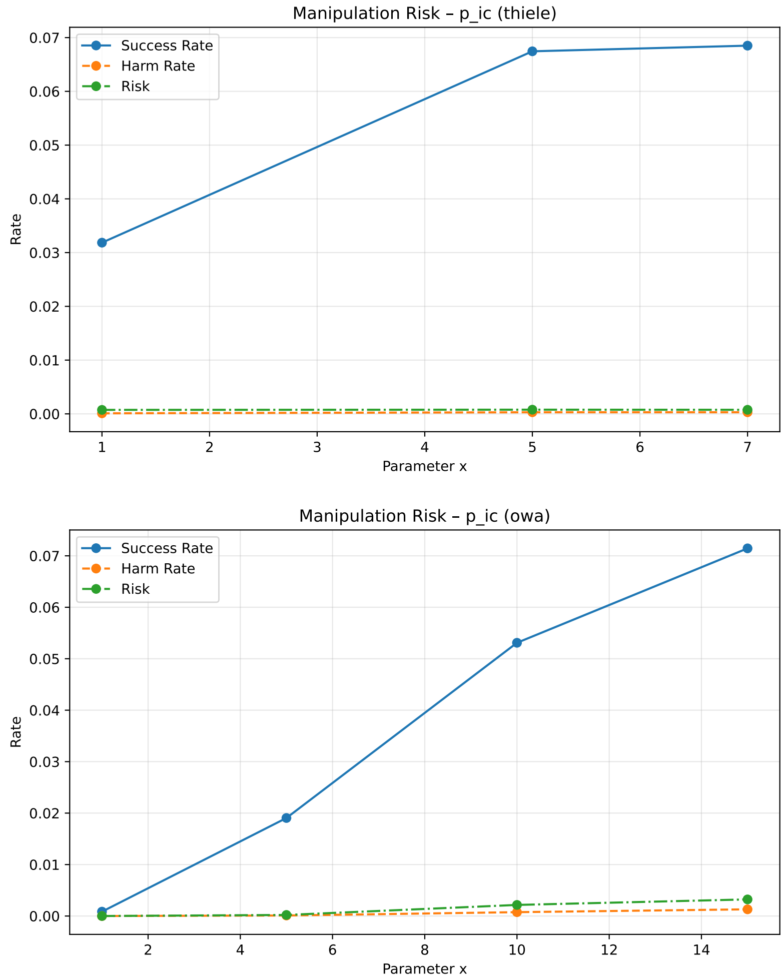



Figure 1: Manipulation risk under p-IC culture. Top: Thiele rules. Bottom: OWA rules.

**Hamming Noise.** Noise adds mild random perturbations but preserves the qualitative trends: utilitarian stays immune; success rises slightly with $x$; harms and risk remain near zero.

# 5 Discussion

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

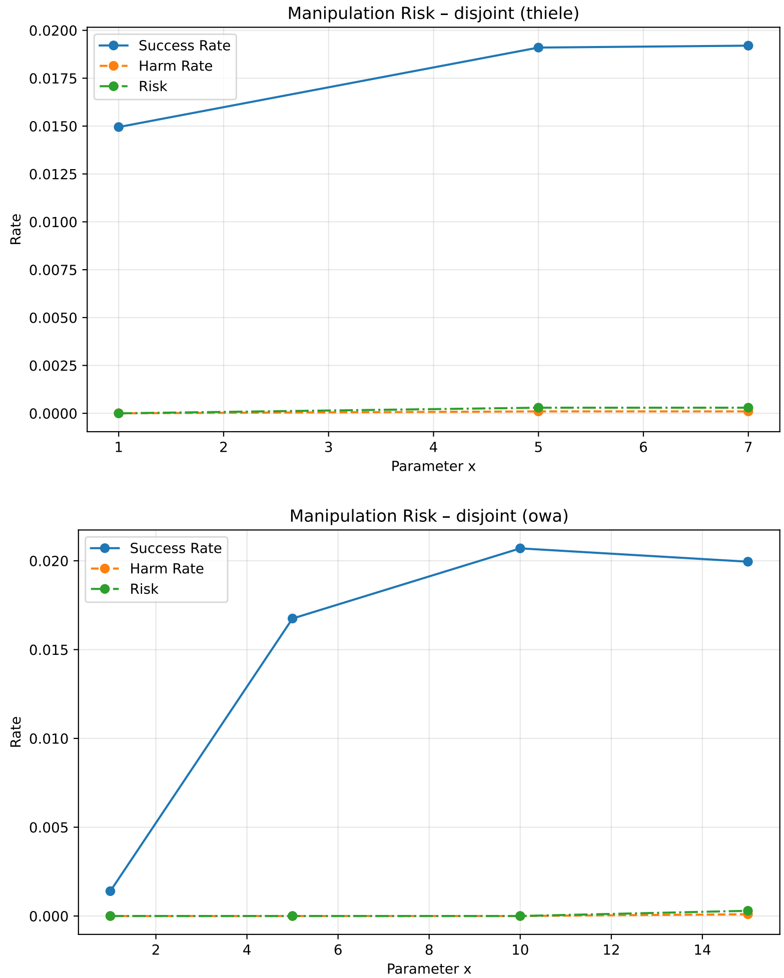



Figure 2: Manipulation risk under disjoint-group culture. Top: Thiele rules. Bottom: OWA rules.

**Effect of Statistical Cultures.** The *p-IC* and *resampling* cultures show modest manipulation opportunities. The *disjoint* model has fewer opportunities overall. *Hamming noise* perturbs preferences slightly but does not qualitatively change the picture.

**Effect of Voting Rules.** The *utilitarian rule* (equivalently mean OWA) is immune to manipulation. For both Thiele and OWA families, moving away from the utilitarian endpoint (flarger $x$ in our parameterization) increases the fraction of successful free-rides, while harm remains rare; consequently, risk (harms/possible) stays very small.

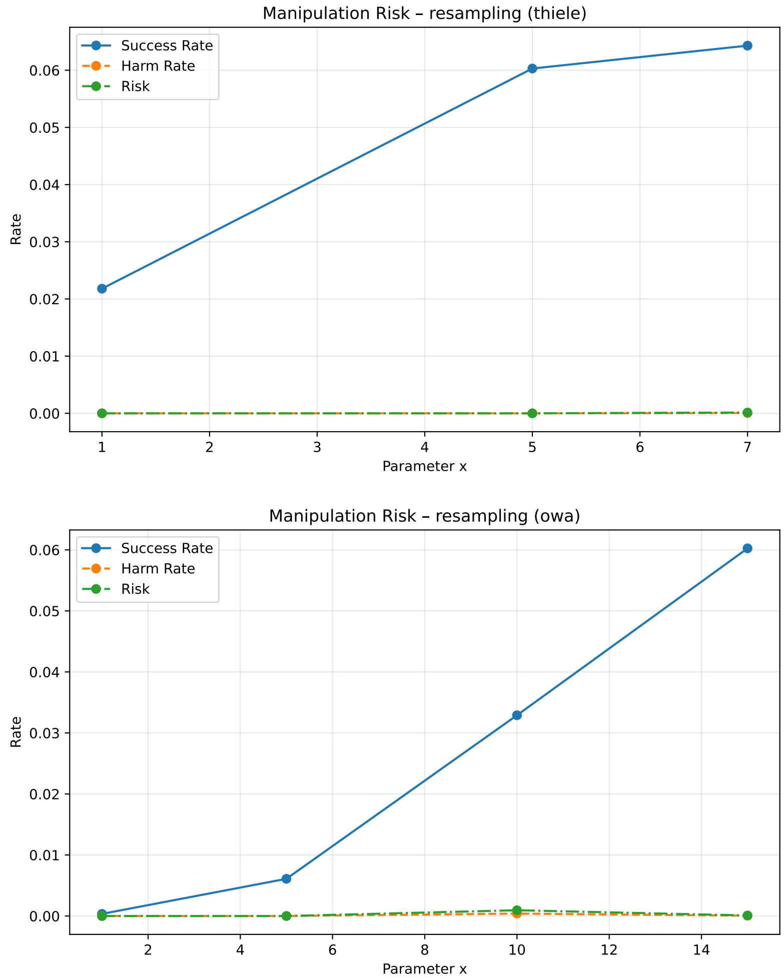
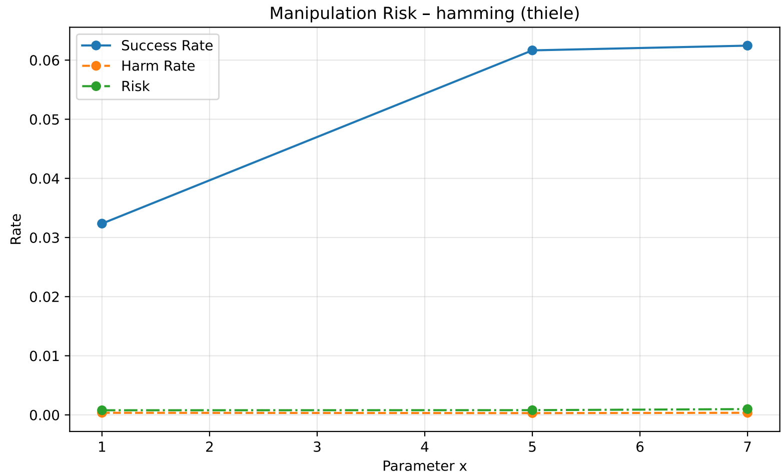



Figure 3: Manipulation risk under resampling culture. Top: Thiele rules. Bottom: OWA rules.

**Risk Metrics.** Across all settings, harms are rare relative to both trials and possible cases, and the risk measure (harms/possible) is consistently close to zero. Taken together, manipulators succeed more often than they harm themselves under the restricted deviation class we test.



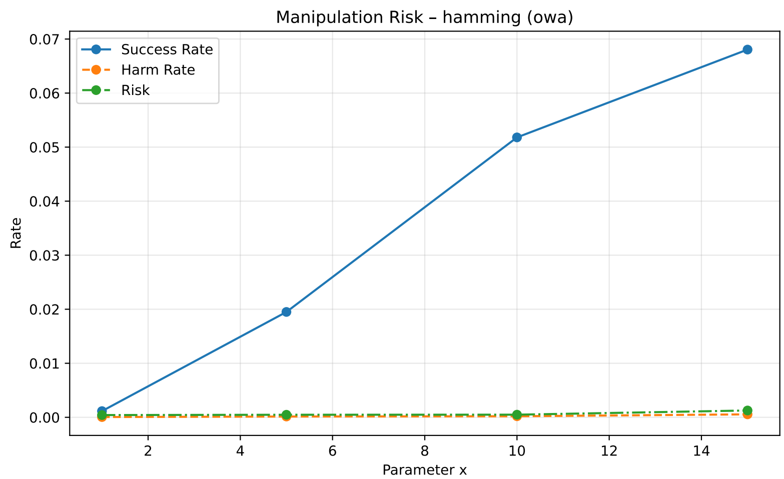




Figure 4: Manipulation risk under Hamming-noise culture. Top: Thiele rules. Bottom: OWA rules.

# 6 Conclusion

The choice of statistical culture influences the frequency of free-riding opportunities, but the qualitative ranking across voting rules is stable in our experiments. Utilitarian is immune; within Thiele/OWA families, higher $x$ values show somewhat more manipulability, yet harms remain rare and risk is small. These findings are compatible with the trends discussed by [1] under our (restricted) deviation model.

Future work could extend these experiments by exploring richer parameter grids for $(p, \phi)$ and $\epsilon$, by scaling to larger electorates, and by testing the full family of admissible free-riding deviations described in [1] (beyond single-approval drops).

## 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.