Impartial Selection with Predictions*

Javier Cembrano
Department of Algorithms and Complexity
Max-Planck-Institut für Informatik
Saarbrücken, Germany
Department of Industrial Engineering
Universidad de Chile
Santiago, Chile
jcembran@mpi-inf.mpg.de

Felix Fischer
School of Mathematical Sciences
Queen Mary University of London
London, UK
felix.fischer@qmul.ac.uk

Max Klimm
Institute of Mathematics
Technische Universität Berlin
Berlin, Germany
klimm@tu-berlin.de

Abstract

We study the selection of agents based on mutual nominations, a theoretical problem with many applications from committee selection to AI alignment. As agents both select and are selected, they may be incentivized to misrepresent their true opinion about the eligibility of others to influence their own chances of selection. Impartial mechanisms circumvent this issue by guaranteeing that the selection of an agent is independent of the nominations cast by that agent. Previous research has established strong bounds on the performance of impartial mechanisms, measured by their ability to approximate the number of nominations for the most highly nominated agents. We study to what extent the performance of impartial mechanisms can be improved if they are given a prediction of a set of agents receiving a maximum number of nominations. Specifically, we provide bounds on the consistency and robustness of such mechanisms, where consistency measures the performance of the mechanisms when the prediction is accurate and robustness its performance when the prediction is inaccurate. For the general setting where up to k agents are to be selected and agents nominate any number of other agents, we give a mechanism with consistency $1 - O\big(\frac{1}{k}\big)$ and robustness $1 - \frac{1}{e} - O\big(\frac{1}{k}\big)$. For the special case of selecting a single agent based on a single nomination per agent, we prove that 1-consistency can be achieved while guaranteeing $\frac{1}{2}$-robustness. A close comparison with previous results shows that (asymptotically) optimal consistency can be achieved with little to no sacrifice in terms of robustness.

1 Introduction

Majority voting is a simple but very important mechanism for collective decision making. Its use dates back at least to ancient Athens, where it was employed for example to decide on the expulsion of citizens from the city [18]. A much more recent proposal uses majority voting to aggregate the solutions of multiple calls to large language models (LLMs) [14]. Some proposals even go so far as

*The full version of the paper can be accessed on arXiv:2510.19002.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

using it in AI alignment, and destroying AI entities if they are perceived as unaligned with human ethics by other AI entities [24]; see also Aaronson [1], Irving et al. [20]. We may formalize this idea by considering a fixed number of different AI entities that can nominate other entities for being incompatible with human values. The entity that receives the most nominations in that way would then be destroyed.

The motivation for using majority decisions in these applications is their superior robustness to outliers compared to decisions made by a single entity. This argument requires, of course, that each entity is incentivized to reveal its true opinion about others rather than following its selfish interests. This is true for voting in general, but even more so in settings like those described above where the set of candidates and the set of voters overlap or are the same. Indeed, it is reasonable to assume that an Athenian citizen in fear of expulsion would have cast their vote for someone they considered likely to receive a large number of nominations, rather than someone they considered worthy of expulsion, in order to minimize their own risk of being expelled. Similarly, it is naïve to assume that AI entities risking destruction due to misalignment will truthfully report on the misalignment of other entities if this negatively affects their own chances of survival. What is needed are voting mechanisms for which the probability that an entity is selected is independent of the nominations cast by that entity. Such mechanisms are called *impartial* in the literature.

While impartiality is obviously appealing, previous work has established strong impossibility results for mechanisms that satisfy it. Deterministic impartial mechanisms that select a fixed number k of entities must fail natural axioms [12, 19], and the overall number of nominations for the selected entities cannot provide a constant approximation to the maximum number of nominations for any set of k entities [3]. Even randomized impartial mechanisms are relatively limited; for example, for the selection of a single entity they can only approximate the maximum number of nominations to a factor of $\frac{1}{2}$ [3, 17].

To improve the performance of impartial mechanisms, we will assume that the mechanism has access to a *prediction* of the entities most suitable for selection. Depending on the application, the prediction could for example come from another LLM not participating in the voting process or from expert advice. The prediction should not be thought of as a prediction about how the votes will turn out, but rather about who is most appropriate for selection. In particular, the prediction is independent of the votes and, thus, following the prediction is impartial. This assumption, while being crucial for our analysis, is satisfied in many scenarios. For a concrete example, consider a situation where a group of agents cast votes on one another about who should be considered for a promotion. The group feeds all CVs to an LLM, asking it for its opinion on who is the best candidate and taking its output as a prediction. In a second step, they can use one of the mechanisms in this paper to do a formal vote. This two-step process has the advantage that when the output of the LLM does not align at all with the opinions of the group, they can overrule its decision. In addition, the process is impartial in the sense that nobody can influence their own chance of being promoted. The guarantee of impartiality applies regardless of whether the agents know the prediction before casting their votes or not.

Our work is part of a growing literature on algorithms and mechanisms with advice; a website maintained by Lindermayr and Megow [22] provides an excellent overview of the area. The area is motivated by the fact that LLMs often provide astonishingly accurate answers, but also sometimes fail spectacularly. Mechanisms with advice therefore need to be able to cope with good as well as bad predictions, without a clear way to distinguish between the two. This trade-off is studied formally by considering the *consistency* and *robustness* of a mechanism. The consistency of a mechanism describes its ability to produce good outcomes when the predictions are accurate; the robustness its ability to produce reasonable results even when the predictions are inaccurate. The ability of a mechanism to move gracefully between these extremes is referred to as *smoothness*.

We will specifically consider deterministic and randomized impartial selection mechanisms with predictions. As it is standard in the literature on impartial selection, we formalize nominations among entities as a directed graph, where the set $[n] = \{1, \dots, n\}$ of vertices represent the entities and an edge from i to j indicates that i casts a nomination for j. A deterministic k-selection mechanism with predictions is given such a graph and a prediction $\hat{S} \subseteq [n]$ with $|\hat{S}| = k$, and returns a set of at most k vertices. A randomized k-selection mechanism is a lottery over deterministic mechanisms. Letting Δ_k denote the maximum sum of indegrees of any k vertices in the graph, a mechanism is called α-consistent for some $\alpha \in [0, 1]$ if the (expected) sum of indegrees of the selected vertices is at least $\alpha\Delta_k$ when the prediction is accurate, i.e., when the total indegree of the vertices in $\hat{S}$

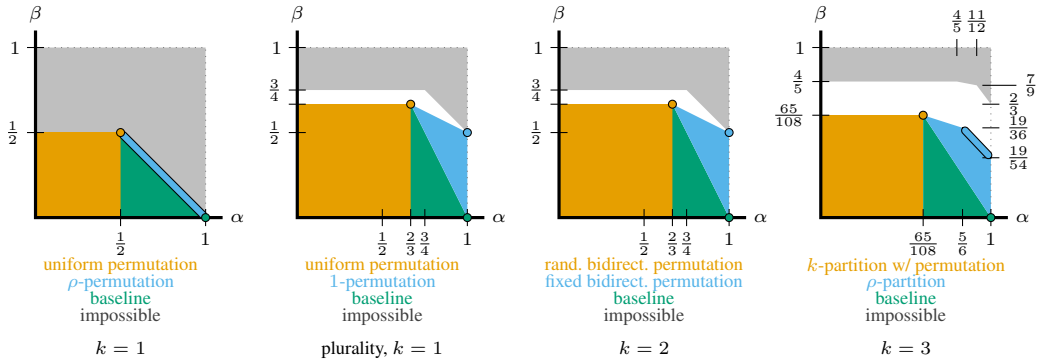

Figure 1: Trade-off between α-consistency and β-robustness of impartial k-selection mechanisms. Orange dots are the best mechanisms from previous work; orange areas are the whole ranges of possible consistency–robustness combinations implied by them. Green dots are the trivial mechanisms always selecting the predicted set; green areas are new ranges of consistency–robustness combinations implied by lotteries of them with previous work. Blue dots and blue rounded rectangles are new mechanisms introduced in this paper; blue areas are new ranges of consistency–robustness combinations implied by them or by lotteries of them with previous work. Gray areas are impossible consistency–robustness combinations as shown in Theorem 6.1. Whether the combinations in the white areas are achievable by impartial mechanisms is left for future research.

is indeed equal to Δ_k. While it is trivial to achieve 1-consistency in an impartial way, by simply returning the predicted set $\hat{S}$, this would lead to arbitrarily bad performance when the predictions are inaccurate. To measure the performance of a mechanism in such cases, a mechanism is called β-robust for some $\beta \in [0,1]$ if the (expected) sum of indegrees of the selected vertices is at least $\beta\Delta_k$ regardless of the quality of the prediction. We will be interested in the largest possible values of α and β for which impartial α-consistent and β-robust mechanisms can be found.

Our Results. We study impartial mechanisms with predictions in different settings; Figure 1 summarizes our results and compares them with previous work. As we initiate the study of impartial mechanisms with predictions, all previous mechanisms are unable to deal with predictions and consequently have equal robustness and consistency. Comparing our results with the baseline mechanism defined as a lottery between the best known mechanisms from the literature and the trivial mechanism that always selects the predicted set shows significant improvements.

We first study the classic setting of randomized impartial 1-selection mechanisms for approval voting. We propose a family of mechanisms we call ρ-permutation mechanisms that are parametrized by a confidence parameter $\rho \in \left[\frac{1}{2}, 1\right]$. The mechanisms build upon the so-called (uniform) permutation mechanism [7, 17], which does not use any predictions and is $\frac{1}{2}$-robust. In a nutshell, this mechanism permutes the vertices uniformly at random and carefully selects a vertex with maximum indegree from vertices that appear previously in the permutation. Our mechanisms favor permutations where the predicted vertex appears towards the end so that most of its incoming edges are likely observed, with a bias that depends on the confidence parameter. We show that for any $\rho \in \left[\frac{1}{2}, 1\right]$ the resulting mechanism is ρ-consistent and $(1-\rho)$-robust (Proposition 3.1), and that this trade-off between consistency and robustness is best-possible (Theorem 6.1). While this optimal consistency–robustness trade-off can also be achieved by the baseline mechanism that randomizes between the uniform permutation mechanism and the mechanism that selects the predicted vertex, such a mechanism would fail basic fairness notions, as discussed in Section 3.

We then study 1-selection mechanisms for plurality voting, where each vertex has exactly one outgoing edge. In this setting, we establish that the 1-permutation mechanism that puts the predicted vertex at the end of the permutation is 1-consistent and $\frac{1}{2}$-robust (Theorem 3.3). Prior work had established that the uniform permutation mechanism is $\frac{2}{3}$-robust, which also implies $\frac{2}{3}$-consistency [11]. By an appropriate lottery between both mechanisms, we achieve $\left(\frac{2}{3} + \frac{1}{3}\rho\right)$-consistency and $\left(\frac{2}{3} - \frac{1}{6}\rho\right)$-robustness for all $\rho \in [0,1]$ (Corollary 3.4). We further show in Theorem 6.1 that for any α-consistent and β-robust impartial mechanism, $\beta \leq \frac{3}{4}$ and $\alpha + \beta \leq \frac{3}{2}$.

We next consider 2-selection mechanisms. In this setting, the bidirectional permutation mechanism [7] was shown to achieve the optimal robustness guarantee of $\frac{1}{2}$. We show that, by placing the predicted

vertices at both ends of the permutation, we obtain the best-possible consistency guarantee of 1 without any sacrifice of robustness (Theorem 4.1). For randomized mechanisms, an appropriate lottery between this mechanism and the randomized permutation mechanism [7] achieves $\left(\frac{2}{3}+\frac{1}{3}\rho\right)$-consistency and $\left(\frac{2}{3}-\frac{1}{6}\rho\right)$-robustness for all $\rho \in [0,1]$; see Proposition 4.2. We further show in Theorem 6.1 that, for any α-consistent and β-robust impartial mechanism, $\beta \leq \frac{3}{4}$ and $\alpha+\beta \leq \frac{3}{2}$.

We finally study randomized k-selection mechanisms for an arbitrary number $k \in \mathbb{N}$ and obtain Theorem 5.2, our most challenging result in terms of technical difficulty. Bjelde et al. [7] proposed the k-partition mechanism with permutation, which partitions the vertices randomly into k sets and selects one vertex from each set in a similar way to the permutation mechanism, but also accounting for edges from outside the set. We propose the ρ-partition mechanism for $\rho \in \left[\frac{1}{2}, 1\right]$, that partitions the vertices randomly into k sets but enforces that each set contains exactly one of the predicted vertices. In each set, the predicted vertex is put at position ρ while all other vertices obtain a position drawn uniformly from the unit interval. We then select one vertex from each set, as the k-partition mechanism with permutation. The mechanism achieves higher consistency by avoiding that more than one of the predicted vertices ends up in the same set, but the analysis requires new techniques because the probabilities of two optimal vertices being in the same set are no longer independent.

In the realm of mechanisms with predictions, it is common to also study approximation guarantees as a function of the prediction error, commonly referred to as smoothness. In our context, a natural notion of error of a predicted set of vertices is the difference between the maximum indegree of a set of k vertices and the indegree of the predicted set, normalized by the maximum indegree so it lies in the interval $[0,1]$. Since all our α-consistent and β-robust mechanisms provide an α-approximation of the indegree of the predicted set, independently of whether this set is or is not optimal, they immediately yield a smoothness guarantee of $\max\left\{\alpha(1-\eta), \beta\right\}$ for an error $\eta \in [0,1]$.

Related Work. Impartiality, as we study it here, was first considered by de Clippel et al. [15] for the division of a divisible resource among members of a set of agents based on divisions proposed by the agents. In the context of selection, it was first studied by Holzman and Moulin [19] and Alon et al. [3]. Holzman and Moulin studied deterministic mechanisms for the special case of *plurality voting*, where each member casts exactly one nomination for another member of the set. They showed that, even in this restricted setting, impartiality is incompatible with the axioms of negative and positive unanimity, where the former requires that a member receiving no nomination is never selected and the latter that a member nominated by all members except themselves is always selected. Alon et al. studied the more general setting of approval voting, where members may nominate an arbitrary number of other members and a fixed number k of members is to be selected. Call a mechanism an *exact* k-selection mechanism if it always selects exactly k members, and α-optimal for $\alpha \in [0,1]$ if the (expected) number of nominations that the selected members receive is always at least an α-fraction of the total number of nominations that the k best members receive. In this terminology, Alon et al. showed that no deterministic, impartial, and exact k-selection mechanism can be α-optimal for any fixed $\alpha > 0$. They further provided a randomized impartial $\frac{1}{4}$-optimal 1-selection mechanism, and a randomized impartial $(1-o(1))$-optimal k-selection mechanism for $k \to \infty$. Fischer and Klimm [17] proposed and analyzed the permutation mechanism and showed that it is $\frac{1}{2}$-optimal, which is best-possible for 1-selection. They further showed that for plurality votes, the same mechanism is α-optimal for $\alpha = \frac{67}{108} \approx 0.620$. Cembrano et al. [11] gave a tight analysis of the permutation mechanism for plurality votes, showing that it is even $\frac{2}{3}$-optimal. They further proposed a new mechanism that is $\frac{2105}{3147}$-optimal, where $\frac{2105}{3147} \approx 0.669$. Bjelde et al. [7] showed that deterministic impartial k-selection mechanisms that are allowed to sometimes select fewer than k members can perform better than exact k-selection mechanisms, and bounded the approximation guarantees of randomized mechanisms that select $k > 1$ members. Caragiannis et al. [9] studied the additive approximation guarantees of impartial selection mechanisms, and Cembrano et al. [12] gave a deterministic mechanism with an improved additive guarantee for plurality votes. Caragiannis et al. [10] considered the additive approximation guarantees of impartial mechanisms that receive *prior information* as additional input. They looked at two different models where members choose their nominations based on a known probability distribution or based on the popularity of a member. We note that this approach differs from ours, since it does not bound the approximation guarantees if the prior information is inaccurate.

The robustness–consistency framework was first used by Purohit et al. [25] to study the performance of online algorithms with predictions. Predictions have been recently incorporated by Berger et al. [6] into the voting setting of metric distortion, where a candidate is to be selected based on rankings

cast by voters with costs given by distances on a common metric space, and the goal is to minimize the ratio between the social cost of the selected candidate and that of the optimal one. More broadly, mechanisms with predictions were first studied by Agrawal et al. [2] for facility location, which has been further considered by Balkanski et al. [5] for randomized mechanisms and different types of predictions, by Fang et al. [16] for a restricted set of candidate locations, and by Istrate and Bonchis [21] for the case where agents' objective is to maximize rather than minimize their distance to the facilities. Balkanski et al. [4] incorporated predictions into the design of strategyproof mechanisms for makespan minimization in scheduling. Xu and Lu [26] also studied a range of mechanism design problems with and without money, including facility location, scheduling, and auction design.

2 Preliminaries

For $n \in \mathbb{N}$, let $[n] = \{1, \dots, n\}$ and let

$$\mathcal{G}_n = \left\{ ([n], E) : E \subseteq ([n] \times [n]) \setminus \bigcup_{i \in [n]} \{(i,i)\} \right\}$$

denote the set of simple graphs with vertex set $[n]$ and without self-loops. For $S, T \in 2^{[n]}$, we denote the edges from vertices in S to vertices in T by

$$N_S^-(T, G) = \{(j,i) \in E : G = ([n], E), j \in S, i \in T\},$$

and the number of such edges by $\delta_S^-(T, G)$. We omit S from the previous notation when $S = [n]$, and we write $N^-(i, G)$ instead of $N^-(\{i\}, G)$ and $\delta^-(i, G)$ instead of $\delta^-(\{i\}, G)$. For $k \in [n]$, we write $\Delta_k(G) = \max_{T \subseteq [n] : |T| = k} \delta^-(T, G)$. We omit k when it is equal to 1 and G whenever it is clear from the context. We refer to the graphs $G = ([n], E) \in \mathcal{G}_n$ such that $|\{(i,j) \in E : j \in [n]\}| = 1$ for every $i \in [n]$, in which all vertices have outdegree exactly one, as *plurality graphs*.

We consider selection mechanisms that obtain a prediction for the set of vertices with maximum indegrees. A *k-selection mechanism with predictions* is a family of functions $f \colon \binom{[n]}{k} \times \mathcal{G}_n \to [0,1]^n$ with $\sum_{i \in [n]} f_i(\hat{S}, G) \leq k$ for all $G \in \mathcal{G}_n$, where $f_i(\hat{S}, G)$ denotes the probability assigned by the mechanism to agent i.[2] For a graph $G \in \mathcal{G}_n$ and $i \in [n]$, the number $f_i(\hat{S}, G)$ is the probability that f selects vertex i when $(\hat{S}, G)$ is the input. A mechanism is called *deterministic* if it only assigns probabilities 0 and 1, and is called *impartial* if $f_i(\hat{S}, G) = f_i(\hat{S}, G')$ whenever for two graphs $G = ([n], E)$ and $G' = ([n], E')$ we have $E \setminus \bigcup_{j \in [n]} \{(i,j)\} = E' \setminus \bigcup_{j \in [n]} \{(i,j)\}$.

For $\alpha \in [0,1]$, we call a k-selection mechanism with predictions α-consistent if it achieves an α-approximation when the predictions are accurate, i.e., $\sum_{i \in [n]} f_i(\hat{S}, G) \delta^-(i, G) \geq \alpha \Delta_k(G)$ for all $n \in \mathbb{N}$, $G \in \mathcal{G}_n$, and $\hat{S} \in \binom{[n]}{k}$ with $\delta^-(\hat{S}, G) = \Delta_k(G)$. For $\beta \in [0,1]$, we call a k-selection mechanism with predictions β-robust if it achieves a β-approximation regardless of the predictions' quality, i.e., $\sum_{i \in [n]} f_i(\hat{S}, G) \delta^-(i, G) \geq \beta \Delta_k(G)$ for all $n \in \mathbb{N}$, $G \in \mathcal{G}_n$, and $\hat{S} \in \binom{[n]}{k}$.

We finally require some notation regarding permutations. For a (vertex) set S, we let $\Pi_S \subset S^{|S|}$ denote the set of permutations of the set S; we refer to the order induced by a permutation as an order *from left to right* for ease of notation. We write Π_n as a shorthand for $\Pi_{[n]}$. For a permutation $\pi \in \Pi_S$, a set $S' \subseteq S$, and a vertex $i \in S$, we write $\pi_{<i} = \{j \in S : j = \pi_r, i = \pi_t \text{ for some } r < t\}$ for the set of vertices that appear to the left of i, $\pi(S') \in \Pi_{S'}$ for the restriction of π to S', and $\bar{\pi} \in \Pi_S$ for the reverse of π. Sometimes we fix the position of some vertices in the permutation. For a set of distinct vertices $\{i_j : j \in [m]\}$ and distinct positions $\{r_j : j \in [m]\}$, we write $\Pi_S(i_1 \to r_1, \dots, i_m \to r_m)$ for the set of permutations $\pi \in \Pi_S$ such that $i_j = \pi_{r_j}$ for every $j \in [m]$.

3 Selecting a Single Vertex

In this section, we study 1-selection mechanisms with predictions. For ease of notation, we denote the predicted set by $\hat{S} = \{\hat{\imath}\}$ and write $\Delta(G)$ instead of $\Delta_1(G)$ for the maximum indegree.

It is well known that deterministic mechanisms cannot achieve any constant approximation in the classic setting without predictions, which for our setting has the direct implication that no deterministic

[2] It is not hard to see that such a distribution over vertices can be translated into a probability over sets of size at most k via the Birkhoff-von Neumann Theorem; see Bjelde et al. [7, Lemma 2.1] for the details.

Algorithm 1 Permutation mechanism $\mathrm{Pm}(G, S, x)$

Input: graph $G = ([n], E)$, set $S \subseteq [n]$, $x \in [0,1]^S$.
Output: vertex $i^{\mathrm{Pm}} \in [n]$.

```
π ← π(x) ∈ Π_S
initialize i^Pm ← π_1 and d ← δ^-_{[n]\S}(π_1)
for r ∈ {2, ..., |S|} do
    i ← π_r
    if δ^-_{([n]\S)∪(π_{<i}\{i^Pm})}(i) ≥ d then
        update i^Pm ← i and d ← δ^-_{([n]\S)∪π_{<i}}(i)
return i^Pm
```

Algorithm 2 ρ-permutation mechanism $\mathrm{Pm}^{\rho}(\hat{\imath}, G)$

Input: graph $G = ([n], E)$, predicted vertex $\hat{\imath} \in [n]$.
Output: vertex $i^{\mathrm{Pm}} \in [n]$.

```
x_î ← ρ
sample x_i ∈ [0,1] uniformly at random ∀i ∈ [n]\{î}
return Pm(G, [n], x)
```

mechanism with predictions can be β-robust for a constant $\beta > 0$. Thus, the trivial answer to the best-possible trade-off between consistency and robustness is given by the mechanism that selects the predicted vertex $\hat{\imath}$ and achieves 1-consistency and 0-robustness.

The problem becomes more interesting with randomization, as the best-known mechanism for the setting without predictions achieves a $\frac{1}{2}$-approximation. We refer to the mechanism achieving this approximation, introduced by Fischer and Klimm [17], as the *uniform permutation mechanism*. This mechanism sorts the vertices uniformly at random and considers them one by one according to this order while maintaining a candidate vertex, initially the first vertex. A vertex is taken as the new candidate if its observed indegree is larger than that of the current candidate, where *observed indegree* refers to the indegree when only considering incoming edges from previous vertices and omitting a potential edge from the current candidate. The vertex that is the candidate in the end is selected.

We define a more general version of this mechanism, where in addition to the graph $G = ([n], E)$, the mechanism receives a subset of vertices $S \subseteq [n]$ and a vector $x \in [0,1]^S$. Vertices in S are those taken into account for the permutation, while all other vertices in $[n] \setminus S$ are not eligible for selection and the incoming edges from these vertices are always considered. The vector $x \in [0,1]^S$ defines the permutation $\pi \in \Pi_S$: i comes before j if its associated value x_i is smaller than x_j. Formally, for every $i, j \in S$ we have $i \in \pi_{<j}$ if and only if either $x_i < x_j$ or both $x_i = x_j$ and $i < j$ hold (we break ties in favor of vertices with smaller indices). We denote the permutation $\pi \in \Pi_S$ constructed in this way from $x \in [0,1]^S$ by $\pi(x)$.

The permutation mechanism for a fixed set S and vector $x \in [0,1]^S$ is formally described in Algorithm 1; we refer to its output for a graph G, a set S, and a vector x by $\mathrm{Pm}(G, S, x)$. The uniform permutation mechanism, providing the best-possible guarantee among randomized 1-selection mechanisms without prediction, corresponds to the mechanism that receives a graph G and returns $\mathrm{Pm}(G, [n], x)$, where $x_i \in [0,1]$ is taken uniformly at random for each $i \in [n]$.

Instead of the uniform permutation mechanism, we consider in the setting with predictions the ρ-permutation mechanism, given in Algorithm 2. This mechanism receives a graph $G = ([n], E)$ and a predicted vertex $\hat{\imath} \in [n]$, and returns $\mathrm{Pm}(G, [n], x)$, where now $\hat{\imath}$ has an associated value $x_{\hat{\imath}} = \rho$ and all values x_i for $i \in [n] \setminus \{\hat{\imath}\}$ are sampled uniformly at random. The value ρ then has the natural interpretation of a confidence parameter: Taking $\rho = 1$ ensures seeing all incoming edges of the predicted vertex, while smaller values of ρ increase the probability of seeing potential outgoing edges of $\hat{\imath}$. This mechanism attains any convex combination of α-consistency and β-robustness between the points $(\alpha, \beta) \in \{(1,0), (\frac{1}{2}, \frac{1}{2})\}$. In Section 6, we will see that this trade-off is actually best-possible.

Proposition 3.1. *For any confidence parameter $\rho \in [\frac{1}{2}, 1]$ the ρ-permutation mechanism is impartial, ρ-consistent and $(1-\rho)$-robust.*

We need some notation. For a fixed graph $G = ([n], E) \in \mathcal{G}_n$, set $S \subseteq [n]$, and vector $x \in [0,1]^S$, we let $i^{\mathrm{Pm}}(G, S, x)$ denote the outcome of $\mathrm{Pm}(G, S, x)$. Whenever x is fixed, we write π for the induced permutation instead of $\pi(x)$. As a key property for the analysis of the (uniform) permutation mechanism, Bousquet et al. [8] showed that, for any fixed permutation, it selects a vertex with maximum indegree from the left. Bjelde et al. [7] extended this result to the case where we restrict to a set of vertices and consider all incoming edges from other vertices. We phrase the latter result with our notation as the following lemma, which we apply in the full version to prove Proposition 3.1.

Lemma 3.2 (Bjelde et al. [7])**.** *For every $G = ([n], E) \in \mathcal{G}_n$, $S \subseteq [n]$, and $x \in [0,1]^S$, it holds that $i^{\mathrm{Pm}}(G, S, x) \in \arg\max\{\delta_{([n]\setminus S)\cup\pi_{<i}}(i, G) : i \in [n]\}$.*

It is worth noting that the consistency and robustness guarantees of Proposition 3.1 are also achieved by a baseline mechanism that returns the predicted vertex with probability ρ and runs the uniform permutation mechanism with probability $1-\rho$. However, the baseline mechanism fails a basic *unanimity* notion introduced by Holzman and Moulin [19]: If a vertex v is such that all other vertices have a single outgoing edge to v, then v should be selected. Whenever v is not the predicted vertex, the baseline mechanism fails to select v with constant probability, while the ρ-permutation mechanism returns v as long as it is not first or second in the permutation, i.e., with probability $1 - O\big(\frac{1}{n}\big)$.

Plurality Voting. A usual restriction in voting is that each member nominates one other member, which in our graph representation implies having vertices with outdegree one. This paradigm of *plurality voting*, extensively considered in the impartial selection literature [11, 19, 23], has been shown to enable better approximation guarantees for randomized mechanisms.[3] In particular, Cembrano et al. [11] proved that the uniform permutation mechanism provides an improved approximation ratio of $\frac{2}{3}$ in this case.

In our setting, we show that the ρ-permutation mechanism with $\rho = 1$, where the predicted vertex is deterministically placed at the end of the permutation and all other vertices are sorted uniformly at random, achieves 1-consistency and $\frac{1}{2}$-robustness. The following theorem provides a more fine-grained bound on the robustness of this mechanism as a function of the maximum indegree Δ of the input graph; the bound of $\frac{1}{2}$ follows by taking the worst case over Δ.

Theorem 3.3. *The 1-permutation mechanism is impartial, 1-consistent, and $\beta(\Delta)$-robust on plurality graphs with maximum indegree $\Delta \geq 2$, where*

$$\beta(\Delta) = \begin{cases} \frac{3\Delta-2}{4\Delta} & \text{if } \Delta \text{ is even,} \\ \frac{3\Delta^2-2\Delta-1}{4\Delta^2} & \text{if } \Delta \text{ is odd.} \end{cases}$$

Moreover, this function β is increasing, implying that this mechanism is impartial, 1-consistent, and $\frac{1}{2}$-robust on plurality graphs.

We prove this theorem in the full version using a strengthened version of a lemma of Cembrano et al. [11] that establishes a negative correlation between the indegree from the left of the maximum-indegree vertex and that of all other vertices. We show that this result remains true for the non-uniform distribution over permutations induced by the vector x defined in the 1-permutation mechanism, and that it holds not only for the maximum-indegree vertex but for any fixed vertex. The proof adapts that of Cembrano et al. [11], defining an injective function between sets of permutations to couple the probabilities that certain indegrees are observed in the permutation taken in the mechanism. We then use the lemma to prove Theorem 3.3. The most challenging case, which ultimately leads to a worse robustness guarantee than in the setting without predictions, is when the maximum-indegree vertex has an incoming edge from the predicted vertex, as this edge is never considered by the mechanism when observing the indegrees from the left. However, since all outdegrees are 1, we can still obtain a lower bound on the probability of selecting this maximum-indegree vertex or another vertex with high indegree.

We now state the implications of Theorem 3.3, in terms of the trade-off between consistency and robustness we can achieve by combining the 1-permutation mechanism with the uniform permutation mechanism. The proof of this result can be found in the full version. In Section 6, we will see that this trade-off is not far from tight.

Corollary 3.4. *For every $\rho \in [0,1]$, there exists a randomized 1-selection mechanism with predictions that is impartial, α-consistent, and β-robust on plurality graphs with maximum indegree Δ, where*

$$\alpha(\Delta) = \begin{cases} \frac{3\Delta+2}{4(\Delta+1)} + \frac{\Delta+2}{4(\Delta+1)}\rho & \text{if } \Delta \text{ is even,} \\ \alpha(\Delta-1) & \text{if } \Delta \text{ is odd,} \end{cases} \qquad \beta(\Delta) = \begin{cases} \frac{3\Delta+2}{4(\Delta+1)} - \frac{\Delta+2}{4\Delta(\Delta+1)}\rho & \text{if } \Delta \text{ is even,} \\ \frac{3\Delta-1}{4\Delta} - \frac{\Delta+1}{4\Delta^2}\rho & \text{if } \Delta \text{ is odd.} \end{cases}$$

In particular, for every $\rho \in [0,1]$, there exists a randomized 1-selection mechanism with predictions that is impartial, $\big(\frac{2}{3} + \frac{1}{3}\rho\big)$-consistent, and $\big(\frac{2}{3} - \frac{1}{6}\rho\big)$-robust on plurality graphs.

[3] The impossibility of providing a constant approximation of the maximum indegree with deterministic mechanisms remains true in this restricted setting [19].

Algorithm 3 Fixed bidirectional permutation mechanism, $\text{Pm}_{\text{bi}}(\hat{S}, G)$

Input: graph $G = ([n], E)$, predicted set $\hat{S} = \{\hat{\imath}_1, \hat{\imath}_2\} \subseteq [n]$.
Output: set $S \subseteq [n]$ with $|S| \leq 2$.
Fix $x_{\hat{\imath}_1} \leftarrow 0$ and $x_{\hat{\imath}_2} \leftarrow 1$
fix $x_i \in (0, 1)$ arbitrarily for each $i \in [n] \setminus \hat{S}$
$\bar{x}_i \leftarrow 1 - x_i$ for every $i \in [n]$
return $\text{Pm}(G, [n], x) \cup \text{Pm}(G, [n], \bar{x})$

4 Selecting Two Vertices

In this brief section, we state our results for the selection of two vertices. In terms of mechanisms without predictions, the best-known deterministic and randomized impartial mechanisms achieve $\frac{1}{2}$- and $\frac{2}{3}$-optimality, respectively. While the bound for deterministic mechanisms is best-possible, only an upper bound of $\frac{3}{4}$ is known for randomized mechanisms [7]. For compactness, throughout this section we denote the predicted set by $\hat{S} = \{\hat{\imath}_1, \hat{\imath}_2\}$.

The deterministic mechanism achieving $\frac{1}{2}$-optimality is based on the permutation mechanism. It runs, for an arbitrarily fixed permutation π, the permutation mechanism for both π and its reverse $\bar{\pi}$, and returns the selected vertices for each direction (potentially the same vertex). A natural approach to incorporate the prediction is to run this mechanism with the predicted vertices at both extremes of the fixed permutation. The resulting mechanism, which we call *fixed bidirectional permutation*, maintains the best-possible robustness of $\frac{1}{2}$ while achieving 1-consistency. The formal description of the mechanism is given in Algorithm 3; the proof of this result is deferred to the full version.

Theorem 4.1. *The fixed bidirectional permutation mechanism is impartial,* 1*-consistent, and* $\frac{1}{2}$*-robust.*

In terms of randomized mechanisms, convex combinations of the best-known mechanism without prediction, achieving $\frac{2}{3}$-robustness [7], and the fixed bidirectional permutation mechanism, achieving 1-consistency and $\frac{1}{2}$-robustness, allows us to attain combinations of α-consistency and β-robustness between $(\alpha, \beta) = \left(\frac{2}{3}, \frac{2}{3}\right)$ and $(\alpha, \beta) = \left(1, \frac{1}{2}\right)$. We state this simple fact in the following proposition; we will see in Section 6 that this combination of consistency and robustness is not far from tight.

Proposition 4.2. *For every* $\rho \in [0, 1]$*, there exists a randomized* 2*-selection mechanism with predictions that is impartial,* $\left(\frac{2}{3} + \frac{1}{3}\rho\right)$*-consistent, and* $\left(\frac{2}{3} - \frac{1}{6}\rho\right)$*-robust.*

5 Selecting $k \geq 3$ Vertices

In this section, we study the impartial selection of $k \geq 3$ vertices when the mechanism is equipped with a prediction on the optimal set.

In terms of deterministic mechanisms, the setting without predictions is far from well understood. Indeed, a large gap remains between the best-known lower and upper bounds of $\frac{1}{k}$ and $\frac{k-1}{k}$ on the approximation guarantee that impartial mechanisms can achieve [7]. Recently, Cembrano et al. [13] improved the lower bound for cases where k is larger than (approximately) $2\sqrt{n}$, but the lower bound of $\frac{1}{k}$ remains the best-known bound for an arbitrary number of agents n. This guarantee comes from the bidirectional permutation mechanism explained in the previous section, whose $\frac{1}{2}$-approximation of the optimal set of two agents translates into a $\frac{1}{k}$-approximation of the optimal committee of k agents. Similarly to the previous section, we can modify this mechanism to maintain its robustness guarantee and achieve 1-consistency. Specifically, we select $k - 2$ vertices from the predicted set and one or two more vertices through our fixed bidirectional permutation mechanism, with the remaining two predicted vertices at the extremes of the permutation. We state the properties of this simple mechanism in the following proposition, proven in the full version.

Proposition 5.1. *There exists a deterministic* k*-selection mechanism with predictions that is impartial,* 1*-consistent, and* $\frac{1}{k}$*-robust.*

Regarding randomized mechanisms, the best-known mechanism for k-selection was developed by Bjelde et al. [7] and provides an approximation guarantee of $\frac{k}{k+1}\left(1 - \left(\frac{k-1}{k}\right)^{k+1}\right)$, which starts at

Algorithm 4 ρ-partition mechanism $\mathrm{Pt}^{\rho}(\hat{S}, G)$

Input: graph $G = ([n], E)$,
predicted set $\hat{S} = \{\hat{\imath}_1, \dots, \hat{\imath}_k\} \in \binom{[n]}{k}$.
Output: set $S^{\mathrm{Pt}} \in \binom{[n]}{k}$.

```
sample j_i ∈ [k] uniformly  ∀i ∈ [n] \ Ŝ
assign i to A_{j_i}  ∀i ∈ [n] \ Ŝ
for j ∈ [k] do
    A_j ← A_j ∪ {î_j}, x_{î_j} ← ρ
    sample x_i ∈ [0,1] uniformly  ∀i ∈ A_j \ {î_j}
    i_j^Pt ← Pm(G, A_j, x)
    S^Pt ← S^Pt ∪ {i_j^Pt}
return S^Pt
```

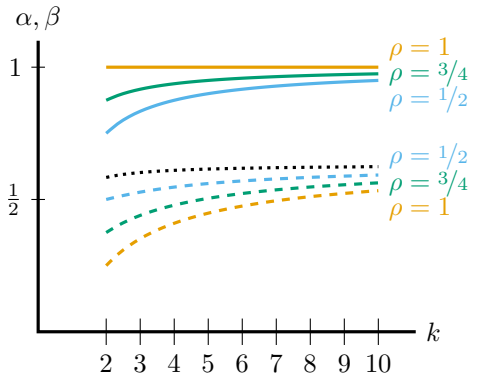

Figure 2: The ρ-partition mechanism (left) and a plot of its α-consistency (solid) and β-robustness (dashed) for the values $\rho = 1$, $\rho = \frac{3}{4}$, and $\rho = \frac{1}{2}$ as a function of k (right). The dotted black line is the consistency and robustness of the k-partition mechanism of Bjelde et al. [7].

$\frac{7}{12} \approx 0.5833$ for $k = 2$, $\frac{65}{108} \approx 0.6019$ for $k = 3$, and approaches $1 - \frac{1}{e} \approx 0.6321$ as k grows. The mechanism assigns each vertex to one out of k sets uniformly at random. It then selects one vertex from each set via the permutation mechanism restricted to that set with an internal permutation taken uniformly at random. While its impartiality is easy to see, the approximation guarantee requires a careful analysis of the expected observed indegree of optimal vertices in each set. In the following, we develop a randomized mechanism with predictions inspired by this mechanism that achieves almost optimal robustness while losing very little in terms of consistency, especially as k grows.

As in the mechanism by Bjelde et al., vertices are assigned to one of k sets, and one vertex is selected from each set by running the permutation mechanism restricted to the set. However, both the assignment to sets and the permutation are not taken independently and uniformly for each vertex anymore. Instead, we assign one predicted vertex to each set; all other vertices are still assigned to a set chosen independently and uniformly at random. Within each set, the permutation is sampled as in the ρ-permutation mechanism from Section 3: For each set A_j with a predicted vertex $\hat{\imath}_j$, we take a vector $x \in [0,1]^{A_j}$ such that $x_{\hat{\imath}_j} = \rho$ and $x_i \in [0,1]$ is taken uniformly at random for each $i \in A_j \setminus \{\hat{\imath}_j\}$. Intuitively, these changes allow the mechanism to see most incoming edges of the predicted vertices while only mildly affecting the distributions to keep a strong robustness guarantee.

The mechanism, which we refer to as the ρ-partition mechanism, is formally presented in Algorithm 4. For $\rho \in [0,1]$, we denote its output by $\mathrm{Pt}^{\rho}(\hat{S}, G)$ for each graph $G = ([n], E)$ and predicted set $\hat{S}$. By tuning the confidence parameter ρ between $\frac{1}{2}$ and 1, we achieve a consistency between $1 - \frac{1}{2k}$ and 1 while only losing $O\left(\frac{1}{k}\right)$ in robustness compared to the best-known mechanism without prediction.

Theorem 5.2. *For any confidence parameter $\rho \in \left[\frac{1}{2}, 1\right]$, the ρ-partition mechanism is impartial, α-consistent, and β-robust, where $\alpha = 1 - \frac{1-\rho}{k}$ and $\beta = \left(1 - \frac{2\rho}{k+1}\right)\left(1 - \left(\frac{k-1}{k}\right)^k\right)$.*

For example, when taking $\rho = \frac{1}{2}$ to prioritize robustness, our mechanism achieves a robustness guarantee of $\frac{1}{2}$ for $k = 2$, $\frac{19}{36} \approx 0.5278$ for $k = 3$, $\frac{35}{64} \approx 0.5469$ for $k = 4$, and approaching $1 - \frac{1}{e} \approx 0.6321$ for $k \to \infty$. The consistency guarantee for this value of ρ and any $k \geq 2$ is $1 - \frac{1}{2k}$, which is $\frac{3}{4} = 0.75$ for $k = 2$, $\frac{5}{6} \approx 0.8333$ for $k = 3$, $\frac{7}{8} = 0.875$ for $k = 4$, and approaches 1 for $k \to \infty$. When taking $\rho = 1$ to maximize consistency, the mechanism is 1-consistent for any k and achieves a robustness guarantee of $\frac{1}{4} = 0.25$ for $k = 2$, $\frac{19}{54} \approx 0.3519$ for $k = 3$, $\frac{105}{256} \approx 0.0.4102$ for $k = 4$, and again approaching $1 - \frac{1}{e} \approx 0.6321$ for $k \to \infty$. Figure 2 illustrates the performance of the ρ-partition mechanism for $\rho \in \left\{\frac{1}{2}, \frac{3}{4}, 1\right\}$ and $k \in \{2, \dots, 10\}$, and compares it with the k-partition mechanism of Bjelde et al. [7].

The proof of Theorem 5.2 can be found in the full version; here we briefly describe the main ideas behind the robustness guarantee, which constitutes the most difficult part of the proof. For the analysis we consider an optimal set S^* and $j \in [k]$ such that A_j contains an optimal vertex, i.e., $S^* \cap A_j \neq \emptyset$, and sample a vertex i^* from $S^* \cap A_j$ uniformly at random. We then bound the expected indegree of i^* that the mechanism observes by bounding the probability that each in-neighbor i of i^* lies in a set other than A_j or in the set A_j but before i^* according to the internal permutation. What complicates

the analysis is that, unlike in the mechanism without predictions, the events $i^* \in A_j$ and $i \in A_j$ are not independent. However, it is not difficult to see that when $i \notin S^* \cup \hat{S}$, the probability of i being in A_j is the same as in the independent case. We show further that when $i \in S^*$ or $i^* \in \hat{S}$, the probability of i being in A_j cannot increase much, and the only difference is given by the position of the predicted vertex in the internal permutation. The most intricate part of the proof is the case where $i^* \in S^* \setminus \hat{S}$ and $i \in \hat{S} \setminus S^*$, because the events of i^* being sampled in the set A_j and i being in this set can be strongly correlated. Indeed, the probability of the former event conditional on $i \in A_j$ can be as large as 1 if, for example, all predicted vertices except i belong to S^*, as in this case $i \in A_j$ implies that i^* is the unique vertex in $S^* \cap A_j$. We tackle this difficulty by directly computing a lower bound on the (unconditional) probability of i^* being sampled as the optimal vertex in A_j.

6 Upper Bounds

To put our consistency and robustness results into perspective, we will now give upper bounds on the values α and β for which an impartial selection mechanism with predictions can simultaneously guarantee α-consistency and β-robustness. We do so for k-selection with $k \in \{1, 2, 3\}$, and for 1-selection from plurality graphs. The upper bounds are shown in Figure 1 alongside the lower bounds obtained in earlier sections.

Theorem 6.1. *The following statements hold:*

- *(i) If a randomized* 1*-selection mechanism with predictions is impartial,* α*-consistent, and* β*-robust, then* $\beta \leq \frac{1}{2}$ *and* $\alpha + \beta \leq 1$*.*
- *(ii) If a randomized* 1*-selection mechanism with predictions is impartial,* α*-consistent, and* β*-robust on plurality graphs, then* $\beta \leq \frac{3}{4}$ *and* $\alpha + \beta \leq \frac{3}{2}$*.*
- *(iii) If a randomized* 2*-selection mechanism with predictions is impartial,* α*-consistent, and* β*-robust, then* $\beta \leq \frac{3}{4}$ *and* $\alpha + \beta \leq \frac{3}{2}$*.*
- *(iv) If a randomized* 3*-selection mechanism with predictions is impartial,* α*-consistent, and* β*-robust, then* $\beta \leq \frac{4}{5}$*,* $4\alpha + 3\beta \leq 6$*, and* $4\alpha + 21\beta \leq 20$*.*

We prove these results in the full version. To this end, we consider appropriate families of graphs and for each vertex in these graphs introduce a variable for the probability with which some impartial, α-consistent, and β-robust k-selection mechanism selects that vertex. We generalize a lemma of Holzman and Moulin [19] to show that one can restrict attention to symmetric mechanisms, and use impartiality, consistency, robustness, and the fact that the probabilities for each graph must sum up to k to obtain a set of linear inequalities involving the probability variables, α, and β. We then show that any values of α and β *not* satisfying the statements violate the linear inequalities.

7 Discussion

We have initiated the study of impartial selection mechanisms with predictions. Unlike majority voting, these mechanisms are not prone to strategic manipulation. While we have made substantial progress regarding the approximation guarantees achievable by such mechanisms, in most settings a moderate gap remains between the upper and lower bounds. We leave closing these gaps for future work. In addition, it would be interesting to test the mechanisms we have proposed in practical applications, for example in the aggregation of outputs of different LLMs.

Acknowledgements

Research was supported by the Deutsche Forschungsgemeinschaft under project number 431465007, by the Engineering and Physical Sciences Research Council under grant EP/T015187/1, and by a Structural Democracy Fellowship through the Brooks School of Public Policy at Cornell University.

References

[1] S. Aaronson. My AI safety lecture for UT effective altruism, 2022. available under `https://scottaaronson.blog/?m=202211`; last accessed May 16, 2025.

[2] P. Agrawal, E. Balkanski, V. Gkatzelis, T. Ou, and X. Tan. Learning-augmented mechanism design: Leveraging predictions for facility location. *Mathematics of Operations Research*, 49 (4):2626–2651, 2024.

[3] N. Alon, F. Fischer, A. Procaccia, and M. Tennenholtz. Sum of us: Strategyproof selection from the selectors. In *Proceedings of the 13th Conference on Theoretical Aspects of Rationality and Knowledge*, pages 101–110, 2011.

[4] E. Balkanski, V. Gkatzelis, and X. Tan. Strategyproof scheduling with predictions. In *Proceedings of the 14th Conference on Innovations in Theoretical Computer Science*. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2023.

[5] E. Balkanski, V. Gkatzelis, and G. Shahkarami. Randomized strategic facility location with predictions. In A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, editors, *Proceedings of the 38th Annual Conference on Neural Information Processing Systems*, 2024.

[6] B. Berger, M. Feldman, V. Gkatzelis, and X. Tan. Learning-augmented metric distortion via (p, q)-veto core. In *Proceedings of the 25th ACM Conference on Economics and Computation*, pages 984–984, 2024.

[7] A. Bjelde, F. Fischer, and M. Klimm. Impartial selection and the power of up to two choices. *ACM Transactions on Economics and Computation*, 5(4):1–20, 2017.

[8] N. Bousquet, S. Norin, and A. Vetta. A near-optimal mechanism for impartial selection. In *Proceedings of the 10th International Conference on Web and Internet Economics*, pages 133–146. Springer, 2014.

[9] I. Caragiannis, G. Christodoulou, and N. Protopapas. Impartial selection with additive approximation guarantees. In *Proceedings of the 12th International Symposium on Algorithmic Game Theory*, pages 269–283. Springer, 2019.

[10] I. Caragiannis, G. Christodoulou, and N. Protopapas. Impartial selection with prior information. In Y. Ding, J. Tang, J. F. Sequeda, L. Aroyo, C. Castillo, and G. Houben, editors, *Proceedings of the ACM Web Conference 2023*, pages 3614–3624, 2023.

[11] J. Cembrano, F. Fischer, and M. Klimm. Improved bounds for single-nomination impartial selection. In *Proceedings of the 24th ACM Conference on Economics and Computation*, page 449, 2023.

[12] J. Cembrano, F. Fischer, D. Hannon, and M. Klimm. Impartial selection with additive guarantees via iterated deletion. *Games and Economic Behavior*, 144:203–224, 2024.

[13] J. Cembrano, S. M. Griesbach, and M. J. Stahlberg. Deterministic impartial selection with weights. *ACM Transactions on Economics and Computation*, 12(3):10:1–10:22, 2024.

[14] L. Chen, J. Q. Davis, B. Hanin, P. Bailis, I. Stoica, M. A. Zaharia, and J. Y. Zou. Are more LLM calls all you need? towards the scaling properties of compound AI systems. In A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, editors, *Proceedings of the 38th Annual Conference on Neural Information Processing Systems*, 2024.

[15] G. de Clippel, H. Moulin, and N. Tideman. Impartial division of a dollar. *Journal of Economic Theory*, 139(1):176–191, 2008.

[16] J. Fang, Q. Fang, W. Liu, and Q. Nong. Mechanism design with predictions for facility location games with candidate locations. In X. Chen and B. Li, editors, *Proceedings of the 18th Annual Conference on Theory and Applications of Models of Computation*, volume 14637 of *Lecture Notes in Computer Science*, pages 38–49, 2024.

[17] F. Fischer and M. Klimm. Optimal impartial selection. *SIAM Journal on Computing*, 44(5): 1263–1285, 2015.

[18] J. G. Heinberg. History of the majority principle. *The American Political Science Review*, 20 (1):52–68, 1926.

[19] R. Holzman and H. Moulin. Impartial nominations for a prize. *Econometrica*, 81(1):173–196, 2013.

[20] G. Irving, P. Christiano, and D. Amodei. AI safety via debate. arXiv Preprint; available under `https://arxiv.org/abs/1805.00899`, 2018.

[21] G. Istrate and C. Bonchis. Mechanism design with predictions for obnoxious facility location. arXiv preprint, available under `https://arxiv.org/abs/2212.09521`, 2022.

[22] A. Lindermayr and N. Megow. Algorithms with predictions, 2025. Available under `https://algorithms-with-predictions.github.io`; last accessed February 10, 2025.

[23] A. Mackenzie. Symmetry and impartial lotteries. *Games and Economic Behavior*, 94:15–28, 2015.

[24] Moebius314. Multiple AIs in boxes, evaluating each other's alignment, 2022. Available under `https://www.lesswrong.com/posts/biskschef2zSNgKkz/multiple-ais-in-boxes-evaluating-each-other-s-alignment`; last accessed May 15, 2025.

[25] M. Purohit, Z. Svitkina, and R. Kumar. Improving online algorithms via ML predictions. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Proceedings of the 31th Annual Conference on Neural Information Processing Systems*, pages 9684–9693, 2018.

[26] C. Xu and P. Lu. Mechanism design with predictions. In L. D. Raedt, editor, *Proceedings of the 31st International Joint Conference on Artificial Intelligence*, pages 571–577, 2022.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction summarize the theoretical results and explain their significance relative to existing work and applications.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses problems left open and limits to applicability.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Theoretical results are correct as stated. Complete proofs are given in the appendix and sketched in the body of the paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: The paper does not include experiments requiring code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The use of predictions in collaborative decision making has the potential to introduce bias and lead to discrimination. Results in the paper are theoretical and general, whereas the question of ethical use of predictions in voting is a practical one and decisions must be made for each particular application. The theoretical results can, however, inform practical decisions regarding the tradeoff between benefits from using predictions and potential harm through bias and discrimination.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The mechanisms proposed in the paper have the potential to achieve better collaborative decisions through the use of predictions. This is discussed in the introduction. The use of predictions in collaborative decision-making has the potential to introduce bias and lead to discrimination. This is discussed in the answer to the previous questions.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not involve data or models that have a high risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.