Procedural Fairness in Multi-Agent Bandits

Joshua Caiata
University of Waterloo
Waterloo, Canada

Carter Blair
Harvard University
Cambridge, United States

Kate Larson
University of Waterloo
Waterloo, Canada

ABSTRACT

In the context of multi-agent multi-armed bandits (MA-MAB), fairness is often reduced to outcomes: maximizing welfare, reducing inequality, or balancing utilities. However, evidence in psychology, economics, and Rawlsian theory suggests that fairness is also about process and who gets a say in the decisions being made. We introduce a new fairness objective, *procedural fairness*, which provides equal decision-making power for all agents, lies in the core, and provides for proportionality in outcomes. Empirical results confirm that fairness notions based on optimizing for outcomes sacrifice equal voice and representation, while the sacrifice in outcome-based fairness objectives (like equality and utilitarianism) is minimal under procedurally fair policies. We further prove that different fairness notions prioritize fundamentally different and incompatible values, highlighting that fairness requires explicit normative choices. This paper argues that procedural legitimacy deserves greater focus as a fairness objective, and provides a framework for putting procedural fairness into practice.

KEYWORDS

Fairness, Multi-armed Bandits, Multi-agent Systems

ACM Reference Format:
Joshua Caiata, Carter Blair, and Kate Larson. 2026. Procedural Fairness in Multi-Agent Bandits. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 18 pages.

1 INTRODUCTION

From the Magna Carta to the words that open constitutions and charters around the world, we have long understood that dignity and fairness require more than simply providing a good outcome. To be fair requires treating each person as an equal, entitled to a voice in the decisions that govern their lives. Yet, in the multi-agent systems we build today, this truth is too often forgotten [9–11]. Fairness is almost always reduced to optimizing for a specific outcome: the sum of utilities, the balancing of welfare, or the smoothing of inequality, echoing the consequentialist tradition of judging actions by their aggregate results [22, 23]. While these notions of fairness may provide elegance and tractability, they miss the very essence of what we consider to be fair. They consider what is gained by a set of decisions, not how they were decided. This is an imposition of values from outside of the system, rather than respecting the very agency from within.

This paper begins from a new conviction, that fairness in multi-agent systems must be grounded not in optimal outcomes, but in the principle of equal voice. To guarantee this is to honour the dignity of participation in multi-agent decision-making; to ignore it is to risk building systems that sacrifice legitimacy, that is, whether the decisions themselves are perceived as rightful and acceptable, for the sake of efficiency. This principle reflects a contractualist view of fairness, which holds that a decision is only legitimate if it cannot be reasonably rejected by those subject to it [19]. We call this principle procedural fairness.

This moral insight is not merely philosophical. Extensive evidence in psychology and economics shows that people consistently value fair process–even if that means outcomes are less than ideal [1, 12, 24]. For example, Lind and Tyler [12] recount an example of a woman whose traffic ticket case was dismissed; however, she still left the courtroom angry because she felt she had compelling evidence and the judge never heard her argument. In fact, many people reported feeling the same way, despite being handed the best possible outcome from the court. This same dynamic appears in collective allocation systems like participatory budgeting, where communities not only want good projects but a voice in which projects are chosen.

Yet existing approaches in multi-agent learning overwhelmingly reduce fairness to outcomes such as utilitarianism, Nash welfare, or inequality. While they capture important values, they impose fairness as an external criterion rather than letting it arise from the agents themselves. What is missing in the literature is a framework that gives agents themselves an equal share of decision-making power. Inspired by Rawls' notion of *pure procedural justice* [18], we formalize procedural fairness in MA-MABs, a framework where each action (pulling an arm) produces potentially different rewards for each agent, sampled from potentially different distributions. This framework naturally captures both the allocation of benefits and the distribution of decision-making power in a simple and easy-to-understand way.

To situate procedural fairness, we compare it, both theoretically and empirically, with two other notions of fairness in multi-agent systems: *equality fairness*, where outcomes are distributed so that agents receive as equal outcomes as possible, and *utilitarian fairness*: decisions maximize aggregate welfare, prioritizing total benefit.

Our central claim is that procedural fairness deserves recognition alongside traditional notions of fairness like Nash welfare, inequality, and utilitarianism, not as an alternative, but as a principle of legitimacy. We now outline our main contributions:

- We define procedural fairness formally in MA-MABs, and compare it to utilitarian, equality fairness, and Nash welfare.
- We prove impossibility results: fairness notions are fundamentally incompatible, showing that fairness requires normative choices.
- We design algorithms for learning fair policies with sublinear regret guarantees.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus.

- We show that procedurally fair policies lie in the core, ensuring stability against coalitional deviation.
- We empirically evaluate our methods across a variety of settings, and show that procedural fairness balances efficiency and equality while preserving legitimacy.

2 MULTI-AGENT MULTI-ARMED BANDITS

Let N represent the number of agents in a multi-arm multi-agent bandit setting, and let $\mathcal{K} = \{1, ..., K\}$ be the set of arms where $K = |\mathcal{K}|$ represents the number of arms. Further, let $P = (p_1, \ldots, p_K)$ represent a *policy* where each element, p_k, represents the probability that arm $k \in \mathcal{K}$ is pulled in any given turn. Note that $0 \leq p_k \leq 1, \forall p_k \in P$, and that $\sum_{i=1}^{K} p_i = 1$. We, at times, abuse notation and refer to K as the set of arms.

When an arm is pulled, all agents receive some reward drawn from a distribution. Agents will not necessarily receive the same reward, and distributions may vary from agent to agent and from arm to arm. We let $\mu^* \in \mathbb{R}^{N \times K}$ represent the agents' true reward means, where $\mu^*_{i,k}$ represents the mean reward agent i receives when arm k is pulled. Additionally, we let $\hat{\mu}^t$ denote the agents' reward estimates at time t, where $\hat{\mu}^t_{i,k}$ is the estimate at time t of the reward agent i receives when arm k is pulled. We let $\sigma_{i,k}$ represent the standard deviation of rewards agent i receives when arm k is pulled. Finally, let F_i be the set of agent i's favourite arms, $F_i = \{j \in \mathcal{K} \mid \mu^*_{i,j} = \max_{k \in \mathcal{K}} \mu^*_{i,k}\}$ where $\mu^*_{i,j}$ represents the reward agent i receives when arm j is pulled.

In all instances, we assume that each true reward mean is strictly bounded between 0 and 1, i.e., $0 < \mu^*_{i,k} < 1, \quad \forall i \in \{1, \ldots, N\}, \forall k \in \{1, \ldots, K\}$, and that drawn rewards are in $[0, 1]$.

We next define the following concepts since they are foundational for our framework.

Definition 1 (Utility). *The utility of an agent i under a policy $P = (p_1, p_2, .., p_k)$ is defined as the agent's expected utility $U_i(P) = \sum_{k=1}^{K} p_k \mu^*_{i,k}$.*

Definition 2 (Decision Share). *The decision share of agent i under a policy $P = (p_1, \ldots, p_K)$ is defined as the total probability assigned to the agent's favourite arm(s): $DS_i(P) = \sum_{k \in F_i} p_k$, where F_i is the set of agent i's favorite arms, defined as $F_i = \{j \in \mathcal{K} \mid \mu^*_{i,j} = \max_{k \in \mathcal{K}} \mu^*_{i,k}\}$.*

Definition 3 (Utility-Based Nash Welfare). *The utility-based Nash welfare is defined similarly to [9]: $\prod_{i=1}^{N} \sum_{k=1}^{K} p_k \mu^*_{i,k}$ where $p_k \in (p_1, ..., p_K)$, which is some policy. This follows the traditional notion of Nash welfare found in the literature.*

Definition 4 (Decision-Share-Based Nash Welfare). *Nash welfare is defined as $\prod_{i=1}^{N} \sum_{k \in F_i} p_k$ where $p_k \in (p_1, ..., p_K)$, which is some policy and F_i is the set of agent i's favourite arms: $F_i = \{j \in \mathcal{K} \mid \mu^*_{i,j} = \max_{k \in \mathcal{K}} \mu^*_{i,k}\}$. In other words, its the Nash welfare of the agents' decision shares.*

3 RELATED WORK

Our work draws on a long-standing debate and evidence from psychology, economics, and political philosophy that one cannot simply consider the fairness of outcomes, but must consider the fairness of process. In psychology, Tyler and Lind's work on procedural justice [12, 24] shows that individuals value having a voice in decisions and weigh the legitimacy of decisions when deciding whether or not they follow these decisions. Further, empirical evidence suggests that individuals prefer fair process over optimal outcomes [1]. In political philosophy, Scanlon's contractualism [19] and Rawls' notion of pure procedural justice [18] articulate legitimacy as whether decision-making rules themselves are justifiable. While these ideas are separate from multi-agent learning, we incorporate these insights into our work.

When it comes to bandits, prior research focuses on arm-centric fairness, where arms are guaranteed minimum pull rates [16] or selected based on merit to avoid favoring worse arms [11]. Another distinct line is online fair division with private rewards, where sequentially arriving items are allocated to individual agents. Works like Procaccia et al. [17] and Schiffer and Zhang [20] maximize social welfare under constraints such as envy-freeness or proportionality in expectation, learning agent preferences via bandit feedback. This contrasts with our problem, which is more aligned with a public good setting since an arm pull may generate rewards for all.

The most analogous setting to ours is MA-MABs with public rewards, which captures the result of an arm-pull as an N-dimensional reward vector for each agent. However, here too, the focus has largely been on outcome-based fairness or efficient coordination. Outcome-focused fairness includes maximizing utility-based Nash Social Welfare (NSW), either as a product of utilities (NSW_{prod}) [9, 10] or its geometric mean [26]. Other aggregate outcome metrics include the Generalized Gini Index (GGI) [2] or achieving Pareto optimality in the reward vector space [25]. These methods evaluate fairness based on the properties of the resulting reward distributions. Separately, research on MA-MABs with public rewards also addresses efficient coordination and communication for utilitarian goals, such as maximizing collective team utility [3].

4 FAIRNESS IN MULTI-AGENT MULTI-ARMED BANDITS

We formally define three notions of fairness, namely, procedural, utilitarian, and equality fairness. We couple each of these definitions with a fairness score that, given some policy, P, measures how fair the policy is under the given fairness criteria. For each notion of fairness, you can find an illustrative example in Appendix A.1.

4.1 Procedural Fairness

Procedural fairness is the principle that each agent should have equal influence over how probabilities are distributed across the arms. Inspired by broader theories of justice (e.g., Rawls), this paper provides the first formalization of procedural fairness in the multi-agent bandit setting. Procedural fairness is particularly relevant in contexts where the legitimacy of the decision process is as important as the outcomes it produces. We formalize this notion of procedural fairness as follows:

Definition 5 (Procedural Fairness). *Let $P = (p_1, p_2, .., p_k)$ be a policy and let $p_i = \sum_k p_{i,k}$ be the probability mass allocated by agent i across all of the arms. A policy P is procedurally fair if it satisfies the following conditions:*

1. ***Equal decision-making influence.*** *Each agent $i \in \{1, \ldots, N\}$ is allocated an equal share of the total probability mass, $\sum_{k=1}^{K} p_{i,k} = \frac{1}{N}, \forall i \in \{1, \ldots, N\}$, where $p_{i,k}$ represents the probability mass that agent i contributes to selecting arm k.*
2. ***Preference-based allocation.*** *Each agent assigns their probability mass to their most preferred arm(s), defined as the set of arms with the highest mean reward for that agent: $F_i = \{j \in \mathcal{K} \mid \mu^*_{i,j} = \max_{k \in \mathcal{K}} \mu^*_{i,k}\}$. If multiple arms achieve the same maximum expected reward, an agent may distribute their probability mass arbitrarily among them.*

To score a given policy's procedural fairness, we formulate an optimization problem. The intuition is: given some probability distribution, can we allocate $\frac{1}{N}$ of probability on behalf of each agent on their favourite arms, subject to the given policy? The extent to which we can allocate these decision shares is the procedural fairness score. We define y_{ij} as a decision share variable, representing how much of agent i's decision share is allocated to arm j, provided that arm $j \in F_i$. The optimization, which we will call $PF(\mu, P)$, is as follows:

$$\begin{aligned} \max \quad & \sum_{i=1}^{N} \sum_{j \in F_i} y_{ij} \\ \text{subject to} \quad & \sum_{j \in F_i} y_{ij} \leq \frac{1}{N} \quad \forall i \in N \\ & \sum_{i : j \in F_i} y_{ij} \leq p_j \quad \forall j \in \mathcal{K} \\ & y_{ij} = 0 \quad \forall i, j \notin F_i \end{aligned}$$

For an illustrative example, please refer to Appendix A.2.

4.2 Equality Fairness

Equality fairness represents equal outcomes, where the policy aims to give each agent as close to equal expected rewards as possible. This principle is most useful in contexts where balance across agents matters more than efficiency or giving each agent equal voice. We define equality fairness as follows:

Definition 6 (Equality Fairness). *A policy $P = (p_1, \ldots, p_k)$ is equally fair if it minimizes inequality in expected rewards among agents. Formally, the policy P is given by:*

$$P = \arg\min_{p' \in \mathcal{P}} \frac{2}{N(N-1)} \sum_{i>j} \left(\sum_{k=1}^{K} p'_k \mu^*_{i,k} - \sum_{k=1}^{K} p'_k \mu^*_{j,k} \right)^2$$

The following formula serves as a measure for equality fairness:

$$EF(\mu^*, P) = 1 - |D(P) - D(P^*)|,$$

$$D(P) = \frac{2}{N(N-1)} \sum_{i>j} \left(\sum_{k} p_k \mu^*_{i,k} - \sum_{k} p_k \mu^*_{j,k} \right)^2,$$

and P^* is an optimal fairness policy. This objective captures equality by penalizing pairwise reward disparities, ensuring that agents achieve as similar expected utility as possible. Please note that $D(\cdot) \in [0, 1]$, so $EF(\mu^*, P) \in [0, 1]$.

4.3 Utilitarian Fairness

Utilitarian fairness is the notion of maximizing the overall utility of the group. This principle is most appropriate in efficiency-driven domains where aggregate outcomes matter most.

Definition 7 (Utilitarian Fairness). *A policy $P = (p_1, \ldots, p_k)$ is utilitarian if it maximizes the expected utility among all agents. Formally, the policy P is given by: $P = \arg\max_{P' \in \mathcal{P}} \sum_{i=1}^{N} \sum_{k=1}^{K} p'_k \mu^*_{i,k}$*

The fairness score for utilitarian fairness is calculated using the following equation: $UF(\mu^*, P) = \frac{\sum_{i \in N} \sum_{k \in \mathcal{K}} p_k \mu^*_{i,k}}{\sum_{i \in N} \sum_{k \in \mathcal{K}} p^*_k \mu^*_{i,k}}$. This provides a percentage share of what the policy is achieving with respect to what can be achieved.

5 ALGORITHMS

We present learning algorithms for the MA-MAB setting, each optimizing for a specific fairness objective. While each definition of fairness calls for the optimization of a different objective, Algorithm 1 presents the general learning procedure. In Algorithm 1, each arm is sampled once, and then it calls a function, `OptimizationStep`, which optimizes for the specific objective, for a total of $T - K$ steps. In the end, it returns the learned policy. In the following sections, we define `OptimizationStep` for each of our three fairness objectives.

For each fairness type, we also prove regret bounds. Because each fairness notion optimizes a fundamentally different objective, the appropriate notion of regret must also be defined relative to that objective. These cannot be directly compared because the underlying ideals themselves are different for each notion.
Procedural Fairness Regret. $R^{PF}(T) = \sum_{t=1}^{T} \mathbf{1}\{\exists i : F_i(t) \neq F_i\}$, counting mismatches between estimated and true favourite-arm sets. Unlike EF and UF, which are outcome-based and naturally admit score-gap regrets, PF is process-based. Its correctness depends on identifying each agent's favourite set and enforcing equal influence in the resulting policy. For this reason, we define PF regret in terms of the number of mismatch rounds in favourite-set recovery, which is analogous to a mistake-bound criterion.

Equality Fairness Regret. $R^{EF}(T) = \sum_{t=1}^{T} [D(P_t) - D(P^*)]$, measuring deviation in inequality from the equally fair optimum.

Utilitarian Fairness Regret. $R^{UF}(T) = \sum_{t=1}^{T} [U(P^*) - U(P_t)]$, a more standard version of regret, where $U(P) = \sum_{i}^{N} \sum_{k}^{K} p_k \mu^*_{i,k}$.

Please note that all algorithms use K arms, N agents, rewards $\mu \in [0, 1]^{N \times K}$, policy $p \in \Delta^K$.

5.1 Procedural Fairness

To learn a procedurally fair policy, we formulate a constrained optimization problem that ensures each agent allocates an equal decision share to their most preferred arms. This formulation is then used in Algorithm 2, which underpins the `OptimizationStep` in Algorithm 1. Additionally, recall that $\hat{\mu}^t_{i,k}$ is the estimated mean of arm k for agent i at timestep t.

$$
\begin{aligned}
\max_{p,y} \quad & \prod_{i=1}^{N} \sum_{j \in F_i} p_j \\
\text{subject to} \quad & \sum_{j \in F_i} y_{ij} = \frac{1}{N} \quad \forall i \in N \qquad (1)\\
& \sum_{i: j \in F_i} y_{ij} = p_j \quad \forall j \in \mathcal{K} \qquad (2)\\
& 0 \leq p_j \leq 1 \quad \forall j \in \mathcal{K}, \quad y_{ij} = 0 \quad \forall i \in N, j \notin F_i
\end{aligned}
$$

Constraint (1) ensures that each agent's total decision share is $\frac{1}{N}$, while constraint (2) guarantees that each arm's total probability equals the sum of decision shares it receives.

When agents have multiple arms, procedural fairness permits many valid allocations. To resolve this ambiguity, we break ties by maximizing decision-share-based Nash welfare. The reason for using this tie-breaking method is that it has some nice theoretical properties, which we will discuss later. In simple cases (e.g., when each agent has a single favoured arm), closed-form solutions exist. However, in the general case with multiple favoured arms, an LP formulation is required to preserve the theoretical properties we study. However, any tie-breaking method will satisfy procedural fairness on its own. We adopt the LP approach throughout.

When learning the optimal policy, the favourite set is derived using UCB-style concentration bounds. For each agent, we compare every arm's upper confidence bound with the lower confidence bound of the empirically best arm. Any arm whose UCB overlaps this lower bound remains in the favourite set. To guarantee convergence, we must ensure that these intervals shrink over time, as we need the intervals to converge to 0 to recover the true favourite set. To solve this problem, we select an arm at random with probability $t^{-(1-\gamma)}$, where $\gamma \in (0, 1)$ is a decay parameter. This guarantees that every arm is pulled sufficiently often so that the confidence radius vanishes as $t \to \infty$. This is proven in Appendix B.4, Lemma 3.

Theorem 1. *With high probability, the regret bound for the Procedural Fairness algorithm, $R^{PF}(T)$, is $O(T^{\gamma} + [\frac{(1+\alpha)^2 \gamma K \ln(NKT)}{\Delta^2_{min}}]^{\frac{1}{\gamma}})$, where $\Delta_{\min} := \min_{i \in [N]} \min_{j \in F_i} \min_{k \notin F_i} \left(\mu^*_{i,j} - \mu^*_{i,k}\right) > 0$, and F_i is the set of agent i's favourite arms based on the true means.*

Refer to Appendix B.6 for the full proof.

Procedurally fair policies also have additional guarantees with respect to the total amount of decision share that each agent will receive:

Observation 1. *With any tie-breaking method, the procedural fairness policy gives agents at least $1/N$ of their maximum decision share and at least $1/N$ of their maximum achievable utility, in expectation.*

Please refer to Appendix B.7 for the full proof.

The complexity of the optimization problem is also relevant. Fortunately, we can easily determine that this problem is convex, and thus can be solved in polynomial time:

Observation 2. *The Procedural Fairness optimization problem is convex and solvable in polynomial time.*

Algorithm 1 LearnPolicy

Require: Fairness method $method \in \{\text{PF}, \text{EF}, \text{UF}\}$, Number of timesteps T, Exploration decay parameter γ if $method = PF$, tolerance α

Initialize policy $P_0 = \frac{1}{K}\mathbf{1}_K$
Initialize arm counts $n = \mathbf{0}_K$, and estimates $\hat{\mu} = \mathbf{0}_{N \times K}$
for each arm $k \in \mathcal{K}$ **do**
  Pull arm k and observe rewards $r_i \forall i \in N$
  Update estimates $\hat{\mu}_{i,k} \leftarrow r_i \forall i \in N$
  Update arm counts $n_k \leftarrow n_k + 1$
end for
for $t = K$ to T **do**
  $P_t \leftarrow$ OptimizationStep$(\hat{\mu}, n, N, K, t, \gamma, \alpha)$
  Sample an arm k according to P_t
  Pull arm k and observe rewards $r_i \forall i \in N$
  Update estimates $\hat{\mu}_{i,k}$ using incremental mean update
  Update arm counts $n_k \leftarrow n_k + 1$
end for
return final policy P_T

Proof. Note that the constraints are linear, so the feasible set is convex. The objective (equivalent to $\sum_{i=1}^{N} \log \left(\sum_{j \in F_i} p_j\right)$) is concave (a sum of log-affine terms). Thus we are maximizing a concave objective over a convex set, which is a convex optimization problem. Such problems are solvable in polynomial time, and in practice, we use a standard solver.

□

5.2 Equality Fairness

To learn an equally fair policy in a multi-arm multi-agent bandit, we balance exploration and exploitation in a UCB-like fashion, also using an optimization step:

$$
\begin{aligned}
\min_{p} \quad & \frac{2}{N(N-1)} \sum_{i>j} \left(\sum_k p_k \hat{\mu}_{i,k} - \sum_k p_k \hat{\mu}_{j,k} \right)^2 - \alpha \sum_k p_k \sqrt{\frac{2\ln(NKt)}{n_k^t}} \\
\text{subject to} \quad & \sum_{k \in \mathcal{K}} p_k = 1, \quad 0 \leq p_k \leq 1, \quad \forall k \in \mathcal{K}.
\end{aligned}
$$

The resulting P will be the policy that we return to the main algorithm.

Theorem 2. *The regret bound for the Equality Fairness algorithm, $R^{EF}(T)$, is $O(\sqrt{KT \ln(NKT)})$ when $\alpha = 4$.*

Please refer to Appendix B.9 for the full proof.

5.3 Utilitarian Fairness

For utilitarian fairness, we use a very similar algorithm to UCB. The reason for this is that UCB optimizes utilitarian fairness by trying to find the arm that maximizes overall utility. Thus,

$$
\begin{aligned}
\max_{p} \quad & \sum_{i \in N} \sum_{k \in \mathcal{K}} p_k \hat{\mu}_{i,k} + \alpha \sum_k p_k \sqrt{\frac{2\ln(NKt)}{n_k^t}} \\
\text{subject to} \quad & \sum_{k \in \mathcal{K}} p_k = 1, \quad 0 \leq p_k \leq 1, \quad \forall k \in \mathcal{K}.
\end{aligned}
$$

Algorithm 2 $\text{OptimizationStep}_{\text{PF}}$: Procedural Fairness Policy Update

Require: Estimates $\hat{\mu}$, pull counts $n^t = (n^t_k)_{k\in[K]}$, number of agents N, number of arms K, time step t, exploration decay γ, tolerance α

With probability $t^{-(1-\gamma)}$, select an arm uniformly at random and **return**. Otherwise:
for each arm $k \in [K]$ **do**
  $z^t_k \leftarrow \begin{cases} \sqrt{\frac{2\ln(NKt)}{n^t_k}}, & \text{if } n^t_k > 0 \\ +\infty, & \text{if } n^t_k = 0 \end{cases}$
end for
$\hat{F} \leftarrow \{\}$ ▷ container holding $\hat{F}_i$ for all i
for each agent $i \in [N]$ **do**
  $j \leftarrow \arg\max_{j'\in[K]} \hat{\mu}^t_{i,j'}$ ▷ break ties arbitrarily
  $\hat{F}_i \leftarrow \{\, k \in [K] : \hat{\mu}^t_{i,k} + \alpha z^t_k \ \geq\ \hat{\mu}^t_{i,j} - \alpha z^t_j \,\}$
  $\hat{F} \leftarrow \hat{F} \cup \{\hat{F}_i\}$
end for
Solve the procedural fairness optimization using $\hat{F}$ (and $\hat{\mu}$) to obtain P
return updated policy P

The resulting P is returned to the main algorithm.

Theorem 3. *The regret bound for the Utilitarian Fairness algorithm, $R^{UF}(T)$, is $O((N+\alpha)\sqrt{KT\ln(NKT)})$, with $\alpha \geq N$.*

Please refer to Appendix B.10 for the full proof.

6 THEORETICAL RESULTS

6.1 Impossibility Results

We show that these different notions of fairness provably conflict, and that there are instances where one must be prioritized over the others. These results make explicit that no single policy can be guaranteed to satisfy all fairness ideals simultaneously, underscoring that procedural fairness is not just another outcome-based criterion but a distinct axis of fairness. These results serve as a motivation for procedural fairness as a unique, but necessary, definition of fairness.

Observation 3. *For multi-agent multi-armed bandit settings with $N \geq 2$ agents and $K \geq 2$ arms, there exist reward structures for which no policy can simultaneously achieve perfect procedural fairness and perfect equality fairness.*

Proof by counterexample. Please refer to Appendix B.11 for the full proof.

Furthermore, it is not possible to guarantee perfect procedural fairness and utilitarian fairness.

Observation 4. *For multi-agent multi-armed bandit settings with $N \geq 2$ agents and $K \geq 2$ arms, there exist reward structures for which no policy can simultaneously achieve perfect procedural fairness and perfect utilitarian fairness.*

Proof by counterexample. Please refer to Appendix B.12 for the full proof.

6.2 Procedural Fairness and the Core

The core is a stability notion originating in cooperative game theory [21]. In the context of public decision-making [6], it represents a distribution over alternatives (arms) that no coalition of agents $A \subseteq \{1, 2, \ldots, N\}$ of size $|A|$ would have an incentive to deviate from, given their proportional share of probability ($|A|/N$). We define the core in our setting in two ways, considering both utility (outcome core) and decision share (procedural core).

Definition 8 (Outcome Core). *Recall that given a distribution over arms, P, an agent i's expected utility is $u_i(P) = P \cdot \mu^*_i$ where $\mu^*_i \in \mathbb{R}^k$ is the reward vector for agent i. We say a distribution $P \in \Delta^k$ is in the outcome core if there is no coalition of agents $A \subseteq \{1, 2, ..., N\}$ and distribution $P' \in \Delta^k$ such that $\frac{|A|}{N} u_i(P') \geq u_i(P), \forall i \in A$ with at least one strict inequality.*

The procedural core adapts the classic notion of the core in cooperative game theory to the setting of procedural fairness. Rather than considering the agents' expected utility, we consider their decision share–the total probability mass assigned to the agents' most preferred arms.

Definition 9 (Procedural Core). *Recall that μ is the reward matrix. Let $F_i = \{k \in \mathcal{K} | \mu^*_{i,k} = \max_{j\in\mathcal{K}} \mu^*_{i,j}\}$ denote agent i's favourite arms. Define a binary vector $X_i \in \{0,1\}^K$ for each agent i, where $X_i[k]$ is 1 if $k \in F_i$ and 0 otherwise. Thus, given a policy, P, the decision share of agent i is defined as $\beta_i(P) = \sum_{k=1}^{K} X_i[k] p_k = PX_i$. Same as the outcome core, a policy P is in the core if there is no coalition of agents $A \subseteq \{1, 2, ..., N\}$ and distribution $P' \in \Delta^k$ such that $\frac{|A|}{N}\beta_i(P') \geq \beta_i(P) \quad \forall i \in A$ with at least one strict inequality.*

The procedural core carries interesting implications for our setting, namely, that maximizing utility-based Nash welfare, as presented by Hossain et al ([9]) does not necessarily lie in the procedural core.

Theorem 4. *A utility-based Nash Welfare-maximizing distribution need not lie in the procedural core.*

Proof by counterexample. Please see Appendix B.13 for the full proof.

Interestingly, probability distributions that are procedurally fair *do*, in fact, lie within the procedural core, if we tie break between an agent's favourite arms by maximizing the decision-share-based Nash welfare, as we do when we aim to learn procedurally fair policies on a multi-arm multi-agent bandit.

Theorem 5. *With decision-share-based Nash-welfare maximizing tie-breaking, the procedural fairness policy is in the procedural core.*

Proof Sketch. *The intuition behind this proof is that if any coalition could block the chosen decision-share-based Nash welfare-maximizing policy, then forming the convex combination of the original and deviating policy would strictly increase its decision-share-based Nash welfare, which contradicts our assumptions, so no such coalition exists. Please see Appendix B.14 for the full proof.*

Additionally, we can see that the procedural core requires procedural fairness.

Theorem 6. *Procedural core implies procedural fairness.*

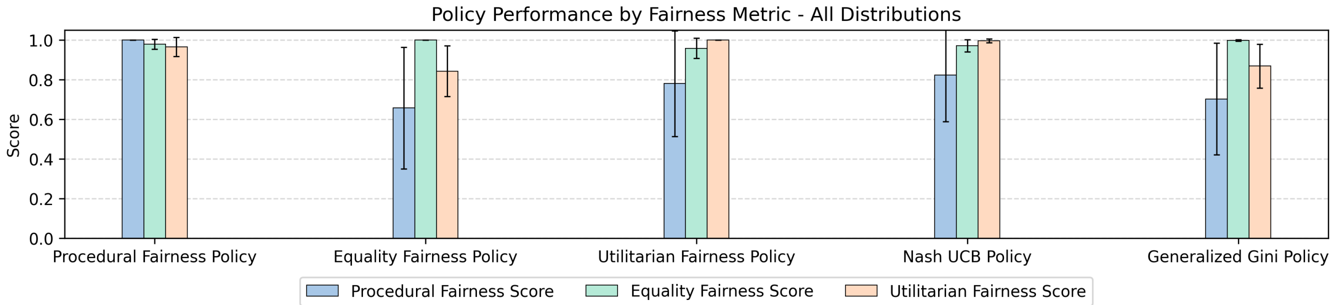


Figure 1: Average fairness metrics per policy type. Each column refers to a specific policy maximizing a certain fairness notion, such as procedural, equality, or utilitarian fairness, maximizing utility-based Nash social welfare, or the Generalized Gini Index. Each bar represents a score for a fairness type, such as procedural, equality, or utilitarian fairness, and the error bars represent one standard deviation from the mean.

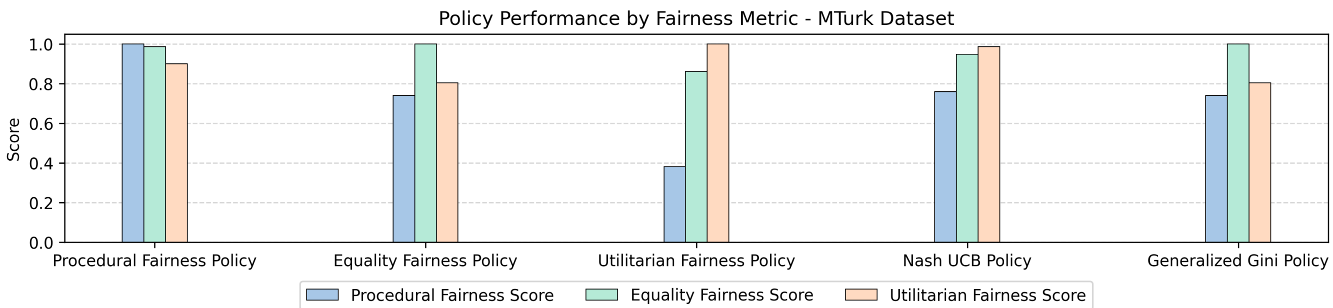


Figure 2: Shows each fairness notion's optimal policy scored on the three metrics. Each bar indicates each fairness metric, and the columns indicate each fairness notion or algorithm's optimal policy. This graph has no error bars at it is the result from a single bandit instance (the dataset).

Proof. Consider any coalition consisting of a single agent. This agent would allocate all probability to their favourite arm. This would give them a decision share of $1/N$. Thus any distribution in the procedural core must give any agent at least $1/N$ probability on their favourite arm. Thus it satisfies our definition of procedural fairness above. □

6.3 Lack of Pareto Dominance Between Fairness Concepts

We briefly discuss the subject of Pareto optimality. We frame Pareto optimality in the case of fairness metrics. Specifically, a policy Pareto dominates another policy if it scores at least as high as another policy on all three fairness metrics, with at least one strictly greater. Formally:

Definition 10. ***Pareto dominance.*** *A policy $P' \in \Delta^K$* Pareto dominates *another policy $P \in \Delta^K$ with respect to the fairness metrics (PF, EF, UF) if $PF(\mu^*, P') \geq PF(\mu^*, P)$, $EF(\mu^*, P') \geq EF(\mu^*, P)$, and $UF(\mu^*, P') \geq UF(\mu^*, P)$, and at least one of these inequalities is strict.*

We consider a benchmark policy which maximizes the utility-based Nash social welfare. This objective, optimized by the NashUCB algorithm [9], balances fairness and efficiency by favouring allocations where all agents receive non-trivial utility. While it doesn't account for procedural fairness, it serves as a strong comparator due to its prominence in the literature.

Here, we show that optimizing for utility-based Nash welfare does not guarantee Pareto dominance of procedurally fair policies, and a procedurally fair policy does not guarantee Pareto dominance of utility-based Nash welfare-optimized policies.

Theorem 7. *Utility-based Nash welfare-optimal policies and procedurally fair policies are not guaranteed to Pareto dominate one another.*

Proof by counterexample. Please see Appendix B.15 for the full proof.

This result underscores the critical limitation of relying on utility-based Nash welfare as a default fairness benchmark. While Nash welfare is widely used because it balances efficiency and inequality, it imposes an outcome-centric criterion that does not preserve equal representation in the decision-making process. The implication here is clear: there is no universal one-size-fits-all fairness metric. Treating Nash welfare as sufficient effectively prioritizes efficiency over equal voice; a normative choice that is often left implicit. On the other hand, procedural fairness highlights this tradeoff and sacrifices outcome-centric utilities to center the decision-making *process* itself. Our findings show that Nash welfare is not a neutral baseline. It is a value-laden standard that sacrifices representation for outcome efficiency, and its legitimacy is contestable unless we defend why balancing utilities—rather than ensuring equal voice—should define fairness.

7 EXPERIMENTS

To understand how our methods work in practice, we conduct experiments on different scenarios and evaluate their performance. We conduct a full factorial sweep across all described parameters: $N \in \{2, 3, 4, 5, 6, 7, 8, 9, 10\}, K \in \{2, 3, 4, 5, 6, 7, 8, 9, 10\}, |F_i| \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. We also have 2 settings for favourite arms, the first where each agent has the same number of favourite arms, and the second where arms are drawn uniformly at random for each agent between 1 and $|F_i|$.

Moreover, we generate preference orderings using PrefVoting [8]. We use the uniform, single-peaked [4], impartial culture [7], and Mallows [13] distributions with $\phi \in \{0.01, 0.25, 0.5, 0.75, 0.99\}$. Then, for each agent, we generate K reward means using a $\mathcal{N}(0.5, 0.25)$ distribution, and assign the highest reward value to the top-ranked arm, the second-highest to the second-ranked arm, and so on, in descending order of rank. Each agent has its own mean for each arm, and each arm for each agent may not have the same reward mean as the others. We use a seed of 42 for this experiment. We filter out combinations that are not possible (like where $|F_i| > K$).

This results in 7,776 different experiment settings. This experiment takes just under 10 minutes on an M2 Pro chip. Note that this experiment setup does not "explore" the bandits, it's simply finding the optimal policy given each notion of fairness or algorithm. We consider the following notions of fairness/algorithms: procedural fairness, equality fairness, utilitarian fairness, NashUCB [9], and Generalized Gini Index (GGI) [2].

Table 1 shows the numerical results of the different algorithms in our experiment. As expected, procedural fairness yields perfect procedural fairness scores (as the algorithm is inherently designed to do). More importantly, however, is that it achieves the best balance across these three fairness metrics. On the other hand, optimizing for fairness notions that are not procedural fairness leads to significant drops in procedural fairness, indicating that it is difficult to incidentally satisfy without explicitly optimizing for it. Figure 1 illustrates this point. Each coloured dot shows the average fairness score for an algorithm across all settings (with the blue dots marking procedural fairness scores). Another interesting finding is that while procedural fairness does not perfectly satisfy the other two fairness notions, it achieves high fairness scores with low standard deviation. On the other hand, algorithms other than procedural fairness perform poorly on the procedural fairness metric and have significantly larger standard deviations, indicating that optimizing for equality or utility maximization does not inadvertently optimize for equal voice and is quite unstable in outcomes.

7.1 Real World Example

To illustrate a real-world example, we pull a dataset from PrefLib [15], notably the Mechanical Turk Dots dataset [14], specifically the variant with 800 voters and 4 candidates. In this dataset, Mechanical Turk workers were shown images of dots and were asked to rank the images from fewest to most dots, producing elections with around 800 voters over 4 candidates. We chose this data because it provides complete preference orderings over a small set of alternatives ($K = 4$), making the experiment tractable and easy to interpret. This dataset then allowed us to create a non-trivial bandit instance and demonstrate the procedural fairness algorithm running on it to learn the optimal policy. For simplicity and speed, we sample 50 votes from the 800 uniformly at random. We generate the bandit by setting an agent's first choice as 0.9, their second choice as 0.63, their third choice as 0.37, and their last choice as 0.1. We then calculate each algorithm's optimal policy and score them using our three fairness metrics, as can be seen in Figure 2.

	PF Score	EF Score	UF Score
PF Policy	1.00 ± 0.00	0.98 ± 0.02	0.97 ± 0.05
EF Policy	0.66 ± 0.31	1.00 ± 0.00	0.84 ± 0.13
UF Policy	0.78 ± 0.27	0.96 ± 0.05	1.00 ± 0.00
NSW Policy	0.82 ± 0.23	0.97 ± 0.03	1.00 ± 0.01
GG Policy	0.70 ± 0.28	1.00 ± 0.00	0.87 ± 0.11

Table 1: Performance metrics for each algorithm. Reported as mean ± one standard deviation. Rows denote algorithms' optimal policy, columns denote fairness scores.

We also run our learning algorithm for procedural fairness on this dataset. Since each reward must be within 0 and 1, we use a Beta distribution. We fix the standard deviation at 0.1 for all agents, and solve for alpha and beta given the standard deviation and mean. We set $\gamma = 0.7$, $\alpha = 1$, and run for $100,000$ steps. Figure 3 shows the progression of the fairness scores as the algorithm learns a procedurally fair policy, and the evolution of the policy over time.

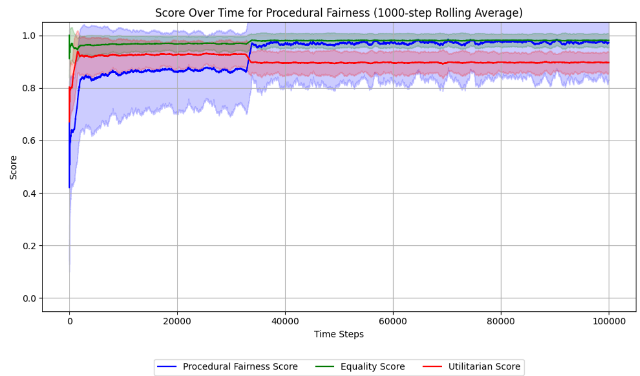
8 DISCUSSION & CONCLUSIONS

8.1 Fairness as Legitimacy

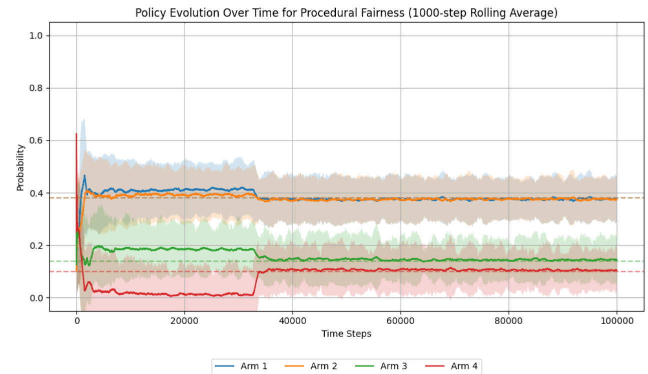
As established, fairness in multi-agent systems has too often been reduced to outcomes over procedure. It is easy to understand why: these are elegant and simple to compute measures. Nonetheless, these are values imposed from the outside. When fairness is defined by an external hand, it risks being viewed as illegitimate.

Procedural fairness offers us another path: a path of respecting the dignity of equal participation and influence in the decision-making process. It does not ask which balance of outcomes is most ideal, but how the decisions are made and who influenced them. This principle is not novel. It is the same principle that gave us the Magna Carta, that sustains constitutions and democracies centuries later, and is backed up by evidence as being preferable [1, 12, 24]. Such systems endure not because they guarantee optimal outcomes, but because the participants recognize the legitimacy of the decisions made. This is what allows them to last.

For multi-agent decision-making, this means changing the design of systems from asking "what distribution of rewards is best?" to "whose preferences shaped the decision?" When we focus on centering the voice of the agents and ensuring fair process, we



(a) The procedural fairness algorithm's fairness scores over time. The shading indicates 1 standard deviation with a rolling average of 1000 steps.



(b) The procedural fairness algorithm's policy over time. The shading indicates 1 standard deviation with a rolling average of 1000 steps.

Figure 3: Comparison of procedural fairness algorithm's performance over time: (a) fairness scores, (b) policy evolution.

move away from normative judgements that impose values from the outside, towards systems that agents regard as legitimate and more accurately reflect what humans view as preferable.

8.2 Tradeoffs as Design Choices

What it means to be fair has never been a unifying principle. It is a contest of rival claims, philosophies, and moral visions, each irreconcilable with the others. This paper makes this blunt claim: no system can satisfy all fairness criteria at once, and every fairness choice declares a normative judgment. To choose one is always to forsake another.

This is not a defect in our framework, in fact, anything but. This is the human condition. Fairness is not discovered in equations, but declared in values. To make decisions on behalf of a group therefore requires us to choose: whose values will govern? Whose interest will reign supreme? The designer's, or the agents' expressed through equal voice?

Historically, we have relied on external metrics to account for fairness: efficiency, equality, or some balance of these two. But elegance and simplicity is not legitimacy. Procedural fairness does not erase tradeoffs; it exposes them, and insists that no agent is denied representation in the decisions that shape their fate.

8.3 Robustness of Procedural Fairness

While procedural fairness is not only normatively appealing, it is also game-theoretically robust. We showed that procedurally fair solutions lie in the procedural core, meaning that no set of agents has an incentive to deviate. In other words, these policies are stable against defection precisely because each agent is guaranteed equal representation in the decision-making process. Moreover, our empirical results demonstrate that procedurally fair policies consistently achieve a strong balance between different fairness metrics, more so than other notions of fairness. Procedurally fair policies do not always maximize utility or equality; they perform well across contexts, avoiding extreme inefficiencies. In fact, while procedural fairness seems to inherently lead to relatively efficient and equal outcomes on its own, no other notion of fairness preserves equal voice in any meaningful way.

Taken together, these properties mark procedural fairness as more than a normative principle. It is a design principle that yields legitimacy, stability, and resilience. In a field where agents must not only cooperate but endure, procedural fairness stands as the strongest baseline for multi-agent systems.

8.4 Broader Applications of Procedural Fairness

One of the most important points of this paper is that procedural fairness is not confined to the abstract setting of multi-agent bandits. It appeals to a greater idea that legitimacy turns less on the outcome achieved than on the process by which it was reached.

Participatory budgeting offers one of the clearest illustrations. Communities routinely accept allocations that are neither maximally efficient nor perfectly equal. Yet, these processes endure, and do so because they give every participant a voice. Legitimacy is preserved, not because the allocation is ideal, but because the decision is shared.

Such a takeaway must not be forgotten for artificial systems. Whether it be allocating computational resources, governing platforms, or coordinating autonomous agents, the critical question is not *what* was decided, but *how*. To embed procedural fairness into multi-agent systems is therefore not to borrow a human tradition for symbolic value, but to ensure that the systems we build today and tomorrow reflect not only intelligence, but humanity's enduring commitment to fairness through equal voice.

8.5 Future Work

There are many directions to which this work may lead. This paper has focused on bandits, but the framework and idea behind procedural fairness is general and can be extended further. One clear direction is to adapt these ideas into richer, sequential settings such as multi-agent reinforcement learning. Another important direction is experimental validation with human experiments. While past work and our experimental results demonstrate that democratizing multi-agent bandits achieves balanced performance on different notions of fairness, the ultimate test here lies in perception. Investigating how people react to procedurally fair outcomes compared to outcome-based policies could provide valuable evidence of its practical relevance, particularly in settings where trust is an essential requirement for human adoption.

REFERENCES

[1] Paul Anand. 2001. Procedural fairness in economic and social choice: Evidence from a survey of voters. *Journal of Economic Psychology* 22, 2 (2001), 247–270. https://doi.org/10.1016/S0167-4870(01)00031-9

[2] Róbert Busa-Fekete, Balázs Szörényi, Paul Weng, and Shie Mannor. 2017. Multi-objective bandits: Optimizing the generalized Gini index. In *International Conference on Machine Learning*. PMLR, 625–634.

[3] Mithun Chakraborty, Kai Yee Phoebe Chua, Sanmay Das, and Brendan Juba. 2017. Coordinated Versus Decentralized Exploration In Multi-Agent Multi-Armed Bandits. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. 164–170. https://doi.org/10.24963/ijcai.2017/24

[4] V. Conitzer. 2009. Eliciting Single-Peaked Preferences Using Comparison Queries. *Journal of Artificial Intelligence Research* 35 (June 2009), 161–191. https://doi.org/10.1613/jair.2606

[5] Vincent Conitzer, Rupert Freeman, and Nisarg Shah. 2017. Fair public decision making. In *Proceedings of the 2017 ACM Conference on Economics and Computation*. 629–646.

[6] Brandon Fain, Kamesh Munagala, and Nisarg Shah. 2018. Fair allocation of indivisible public goods. In *Proceedings of the 2018 ACM Conference on Economics and Computation*. 575–592.

[7] Georges Théodule Guilbaud. 1952. Les Théories de l'intérêt général et le Problème Logique de l'agrégation. *Economie appliquée* 5, 4 (1952), 501–584.

[8] Wesley H. Holliday and Eric Pacuit. 2025. pref_voting: The Preferential Voting Tools package for Python. *Journal of Open Source Software* 10, 105 (2025), 7020. https://doi.org/10.21105/joss.07020

[9] Safwan Hossain, Evi Micha, and Nisarg Shah. 2021. Fair algorithms for multi-agent multi-armed bandits. *Advances in Neural Information Processing Systems* 34 (2021), 24005–24017.

[10] Matthew Jones, Huy Nguyen, and Thy Nguyen. 2023. An efficient algorithm for fair multi-agent multi-armed bandit with low regret. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 8159–8167.

[11] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. 2016. Fairness in learning: Classic and contextual bandits. *Advances in Neural Information Processing Systems* 29 (2016).

[12] E.A. Lind and T.R. Tyler. 2013. *The Social Psychology of Procedural Justice*. Springer US. https://books.google.ca/books?id=j97VBQAAQBAJ

[13] Colin L Mallows. 1957. Non-Null Ranking Models. I. *Biometrika* 44, 1/2 (1957), 114–130.

[14] Andrew Mao, Ariel Procaccia, and Yiling Chen. 2013. Better Human Computation Through Principled Voting. *Proceedings of the AAAI Conference on Artificial Intelligence* 27, 1 (Jun. 2013), 1142–1148. https://doi.org/10.1609/aaai.v27i1.8460

[15] Nicholas Mattei and Toby Walsh. 2013. PrefLib: A Library of Preference Data HTTP://PREFLIB.ORG. In *Proceedings of the 3rd International Conference on Algorithmic Decision Theory (ADT 2013) (Lecture Notes in Artificial Intelligence)*. Springer.

[16] Vishakha Patil, Ganesh Ghalme, Vineet Nair, and Yadati Narahari. 2021. Achieving fairness in the stochastic multi-armed bandit problem. *Journal of Machine Learning Research* 22, 174 (2021), 1–31.

[17] Ariel D Procaccia, Ben Schiffer, and Shirley Zhang. 2024. Honor among bandits: No-regret learning for online fair division. *Advances in Neural Information Processing Systems* 37 (2024), 13183–13227.

[18] John Rawls. 1971. *A Theory of Justice: Original Edition*. Harvard University Press. http://www.jstor.org/stable/j.ctvjf9z6v

[19] T.M. Scanlon. 1998. *What We Owe to Each Other*. Harvard University Press. https://books.google.ca/books?id=FwuZcwMdtzwC

[20] Benjamin Schiffer and Shirley Zhang. 2025. Improved Regret Bounds for Online Fair Division with Bandit Learning. *arXiv preprint arXiv:2501.07022* (2025).

[21] Lloyd S Shapley. 1971. Cores of convex games. *International journal of game theory* 1 (1971), 11–26.

[22] Walter Sinnott-Armstrong. 2023. Consequentialism. In *The Stanford Encyclopedia of Philosophy* (Winter 2023 ed.), Edward N. Zalta and Uri Nodelman (Eds.). Metaphysics Research Lab, Stanford University.

[23] J.J.C. Smart and B. Williams. 1973. *Utilitarianism: For and Against*. Cambridge University Press. https://books.google.ca/books?id=J0w3ER2fWv4C

[24] T.R. Tyler. 1990. *Why People Obey the Law*. Yale University Press. https://books.google.ca/books?id=ZstoQgAACAAJ

[25] Mengfan Xu and Diego Klabjan. 2023. Pareto regret analyses in multi-objective multi-armed bandit. In *International Conference on Machine Learning*. PMLR, 38499–38517.

[26] Mengxiao Zhang, Ramiro Deo-Campo Vuong, and Haipeng Luo. 2024. No-Regret Learning for Fair Multi-Agent Social Welfare Optimization. *Advances in Neural Information Processing Systems* 37 (2024), 57671–57700.

A EXAMPLES

A.1 Fair Policies

Consider a multi-arm multi-agent bandit setting with $N = 3$ and $K = 2$, and the following reward structure:

$$\mu = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

In this scenario, for procedural fairness, agents 1 and 2 would place their $\frac{1}{N}$ probability mass on the first arm, and agent 3 would place their probability mass on the second arm. This results in a policy of $(\frac{2}{3}, \frac{1}{3})$. An *equally fair* policy would be $P = (\frac{1}{2}, \frac{1}{2})$, as each agent's expected utility from such a policy would be 0.5. A *utilitarian* policy would be $P = (1, 0)$, as the leftmost arm has an overall utility among all agents of 2, where the rightmost arm has an overall utility of 1 among all agents.

A.2 Procedural Score

To illustrate, consider the utility matrix

$$\mu = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

and policy $P = (\frac{1}{2}, \frac{1}{2})$. We aim to find y_{ij} for each agent i and arm j, distributing each agent's $\frac{1}{3}$ share of the total probability mass to their favourite arm without exceeding any arm's total availability in the policy $P = (\frac{1}{2}, \frac{1}{2})$.

Agent 1 prefers arm 1. Allocate their full share: $y_{11} = \frac{1}{3}$, $y_{12} = 0$. Then, agent 2 also prefers arm 1. Since $\frac{1}{3}$ is already used by agent 1 with arm 1, only $\frac{1}{6}$ remains (recall that $y_{11} + y_{21} \leq \frac{1}{2}$ must hold, and we already have $\frac{1}{3} + y_{21} \leq \frac{1}{2}$ from agent 1 allocating their decision share to arm 1). So: $y_{21} = \frac{1}{6}$, $y_{22} = 0$. Finally, Agent 3 prefers arm 2. We can fully allocate their decision share to arm 2, so: $y_{32} = \frac{1}{3}$, $y_{31} = 0$. Summing all y_{ij} gives $\frac{5}{6}$, the procedural fairness score. This reflects agent-level allocation ratios of 1, 0.5, and 1, averaging to $\frac{5}{6}$.

B THEORETICAL RESULTS

B.1 Hoeffding Mean Concentration

Lemma 1 (Hoeffding Mean Concentration). *Let* $z_k^\ell = \sqrt{\frac{2\ln(NKt)}{n_k^\ell}}$, *where* n_k^ℓ *represents the number of times arm* k *has been pulled by timestep* ℓ. *Then, with probability at least* $1 - \frac{2}{(NKt)^3}$, *we have that* $\forall i \in [N], k \in [K], \ell \in [t] : |\hat{\mu}_{i,k}^\ell - \mu_{i,k}^*| \leq z_k^\ell$

Proof. This proof is nearly identical to an existing proof [9]. Consider a fixed timestep t. Let $X_{i,k}^\ell$ be the set of observations of arm k at timestep ℓ, so $\hat{\mu}^\ell$ is equal to the mean of all $X_{i,k}^\ell$ sets (separately, as a matrix of means). So $|X_{i,k}^\ell| = n_k^\ell$. Further, let $z_k^\ell = \sqrt{\frac{2\ln(NKt)}{n_k^\ell}}$. Recall each reward $x \in [0, 1]$. Then, by Hoeffding's inequality followed by a union bound, we have:

$$\begin{aligned}
\forall i \in [N],\, k \in [K],\, \ell \in [t] : P\left(\left|\sum_{x \in X_{i,k}^\ell} x - n_k^\ell \mu_{i,k}^*\right| > n_k^\ell z_k^\ell\right) &= P\left(\left|n_k^\ell \hat{\mu}_{i,k}^\ell - n_k^\ell \mu_{i,k}^*\right| > n_k^\ell z_k^\ell\right) \\
&= P\left(|\hat{\mu}_{i,k}^\ell - \mu_{i,k}^*| > z_k^\ell\right) \\
&\leq 2\exp\left(\frac{-2(n_k^\ell z_k^\ell)^2}{n_k^\ell}\right) \\
&= 2\exp\left(-2n_k^\ell (z_k^\ell)^2\right) \\
&= 2\exp\left(-2n_k^\ell \left(\sqrt{\frac{2\ln(NKt)}{n_k^\ell}}\right)^2\right) \\
&= 2\exp\left(-4\ln(NKt)\right) \\
&= \frac{2}{(NKt)^4}
\end{aligned}$$

By the union bound, we get

$$\Pr\Big[\exists\, i \in [N],\, k \in [K],\, \ell \in [t] :\ \big|\hat{\mu}_{i,k}^{\ell} - \mu_{i,k}^{*}\big| > z_k^{\ell}\Big] \le \sum_{i\in[N]} \sum_{k\in[K]} \sum_{\ell\in[t]} \frac{2}{(NKt)^4}$$
$$= \frac{2\,NKt}{(NKt)^4} = \frac{2}{(NKt)^3}\,.$$

Taking the complement, we get

$$\Pr\Big[\forall\, i \in [N],\, k \in [K],\, \ell \in [t] :\ \big|\hat{\mu}_{i,k}^{\ell} - \mu_{i,k}^{*}\big| \le z_k^{\ell}\Big] \;\ge\; 1 - \frac{2}{(NKt)^3}\,.$$

□

B.2 Uniform-in-Time Hoeffding Mean Concentration

Corollary 8 (Uniform-in-time Hoeffding Mean Concentration). *Fix a horizon T. Then*

$$\Pr\Big(\forall t \le T,\ \forall i \in [N],\ k \in [K],\ \ell \le t :\ |\hat{\mu}_{i,k}^{\ell} - \mu_{i,k}^{*}| \le z_k^{\ell}(t)\Big) \;\ge\; 1 - \frac{3}{(NK)^3}.$$

Proof. By the preceding lemma and union-bounding over $t = 1, \ldots, T$,

$$\Pr\Big(\forall t \le T,\ \forall i \in [N],\ k \in [K],\ \ell \le t :\ |\hat{\mu}_{i,k}^{\ell} - \mu_{i,k}^{*}| \le z_k^{\ell}(t)\Big) \;\ge\; 1 - \frac{2}{(NK)^3} \sum_{t=1}^{T} \frac{1}{t^3}.$$

Applying the integral bound for decreasing functions,

$$\sum_{t=1}^{T} \frac{1}{t^3} \;\le\; 1 + \int_1^T x^{-3}\,dx = \tfrac{3}{2} - \tfrac{1}{2T^2},$$

$$\Pr\Big(\forall t \le T,\ \forall i \in [N],\ k \in [K],\ \ell \le t :\ |\hat{\mu}_{i,k}^{\ell} - \mu_{i,k}^{*}| \le z_k^{\ell}(t)\Big) \;\ge\; 1 - \frac{2}{(NK)^3}\Big(\tfrac{3}{2} - \tfrac{1}{2T^2}\Big) \;\ge\; 1 - \frac{3}{(NK)^3}.$$

□

B.3 Randomized Exploration Bound

Lemma 2. *If at round t the procedural fairness algorithm explores with probability $p_{rand}^t = t^{-(1-\gamma)}$, then the total number of exploration rounds up to time T is given by:*

$$\sum_{t=1}^{T} p_{rand}^t = \sum_{t=1}^{T} t^{-(1-\gamma)} = \Theta(\frac{1}{\gamma} T^{\gamma})$$

And since each exploration round can incur at most one mismatch, then the total mismatches coming from forced exploration is bounded by $O(\frac{1}{\gamma}T^{\gamma})$.

Proof. This can be easily shown via an integral: $\sum_{t=1}^{T} p_{rand}^t = \sum_{t=1}^{T} t^{-(1-\gamma)} \le 1 + \int_1^T x^{\gamma-1} = 1 + \frac{T^{\gamma}-1}{\gamma} = \Theta(\frac{1}{\gamma}T^{\gamma})$ □

B.4 Arm Counts and Adjustment Bounds

Lemma 3. *If at round t the procedural fairness algorithm explores with probability $p_{rand}^t = t^{-(1-\gamma)}$, and selects an arm uniformly at random, then we know the following:*

(1) $\mathbb{E}[n_k^t] \ge \frac{1}{K}\sum_{s=1}^{t} p_{rand}^t = \Omega(\frac{t^{\gamma}}{\gamma K})$
(2) $n_k^t = \Omega(\frac{t^{\gamma}}{\gamma K})$
(3) $z_k^t = \sqrt{\frac{2\ln(NKt)}{n_k^t}} = O(\sqrt{\frac{\gamma K \ln(NKt)}{t^{\gamma}}}) \longrightarrow 0$

Proof. We can see by construction that $\mathbb{E}[n_k^t] \ge \frac{1}{K}\sum_{s=1}^{t} p_{rand}^t$. From Lemma 2, we know that $\sum_{s=1}^{t} p_{rand}^t = \Theta(\frac{1}{\gamma}T^{\gamma})$, so it follows that $\mathbb{E}[n_k^t] \ge \frac{1}{K}\sum_{s=1}^{t} p_{rand}^t = \Omega(\frac{t^{\gamma}}{\gamma K})$ This result also provides, using a Chernoff bound:

$$\Pr\Big[n_k^t < \tfrac{1}{2}\,\mathbb{E}[n_k^t]\Big] \le \exp\!\big(-\tfrac{1}{8}\,\mathbb{E}[n_k^t]\big) = \exp(-\Omega(\frac{t^{\gamma}}{\gamma K})).$$

With high probability, we have that:

$$n_k^t \;\geq\; \tfrac{1}{2}\,\mathbb{E}[n_k^t] = \Omega\big(\frac{t^\gamma}{\gamma K}\big).$$

Therefore, we have that $n_k^t = \Omega\big(\frac{t^\gamma}{\gamma K}\big)$, which dominates $\ln t$ with $0 < \gamma < 1$. Because of this, we then have $z_k^t = \sqrt{\frac{2\ln(NKt)}{n_k^t}} = O(\sqrt{\frac{\gamma K \ln(NKt)}{t^\gamma}})$. Since t^γ dominates $\gamma K \ln(NKt)$, we can see that as $t \to \infty$, the latter two bounds approach 0, completing our proof. □

B.5 Exploitation Phase Bound

Lemma 4. *Define a mismatch to refer to a step where the actual favourite arm set does not equal the estimated favourite arm set for at least one agent. In the exploitation phase, the number of mismatches is bounded by* $O([\frac{4(1+\alpha)^2\gamma K \ln(NKT)}{\Delta_{min}^2}]^{\frac{1}{\gamma}})$

Proof. Let F_i^* represent agent i's actual favourite arms set, and let $j^* \in F_i^*$ and $k \notin F_i^*$, and $\Delta = \Delta_{i,j^*,k} > 0$, where $\Delta_{i,j^*,k} = \mu_{i,j^*}^* - \mu_{i,k}^*$ (in the case where every arm is a favourite arm, the problem is trivial). Let $j \in \arg\max_j \hat{\mu}_{i,j}^t$ be an estimated favourite arm for agent i. In order for an arm k to remain in the favourite arm set, we must satisfy the following condition: $\hat{\mu}_{i,k}^t + \alpha z_k^t \geq \hat{\mu}_{i,j}^t - \alpha z_j^t$ (note that j^* may or may not be the same arm as j). Thus, we want to find at what timestep t, the following will be true: $\hat{\mu}_{i,k}^t + \alpha z_k^t < \hat{\mu}_{i,j}^t - \alpha z_j^t$.

We know, by definition, that $\hat{\mu}_{i,j}^t \geq \hat{\mu}_{i,j^*}^t$, as j represents the arm that has the empirically highest mean at time t, so the estimate is at least as high as j^*'s estimate. This provides a lower bound to the right-hand side of our equation. We can then use the more conservative exclusion inequality $\hat{\mu}_{i,k}^t + \alpha z_k^t < \hat{\mu}_{i,j^*}^t - \alpha \max\{z_j^t, z_{j^*}^t\}$. We can replace our other inequality with more conservative inequalities, as we are simply looking for an upper bound. If we replace our inequality with a more conservative inequality, then we know that if the conservative inequality is satisfied, then the original inequality is satisfied.

From Lemma 1, we know with high probability that $\hat{\mu}_{i,k}^t \leq \mu_{i,k}^* + z_k^t$, providing an upper bound to the left-hand side of our inequality. We can then use the new, more conservative exclusion inequality $\mu_{i,k}^* + z_k^t + \alpha z_k^t < \hat{\mu}_{i,j^*}^t - \alpha \max\{z_j^t, z_{j^*}^t\}$. By the same Lemma, we also know that $\hat{\mu}_{i,j^*}^t \geq \mu_{i,j^*}^* - z_{j^*}^t \geq \mu_{i,j^*}^* - \max\{z_j^t, z_{j^*}^t\}$. Similarly, we have a new lower bound for our right-hand side, so we can replace our exclusion inequality with $\mu_{i,k}^* + z_k^t + \alpha z_k^t < \mu_{i,j^*}^* - \max\{z_j^t, z_{j^*}^t\} - \alpha \max\{z_j^t, z_{j^*}^t\}$. Rearranging and factoring gives: $\Delta_{i,j^*,k} > (1+\alpha)(z_k^t + \max\{z_j^t, z_{j^*}^t\})$.

For our bound, we are primarily concerned with the smallest Δ, so we use $\Delta_{\min} := \min_{i\in[N]} \min_{j\in F_i} \min_{k\notin F_i} \left(\mu_{i,j}^* - \mu_{i,k}^*\right) > 0$, and our inequality becomes $\Delta_{min} > (1+\alpha)(z_k^t + \max\{z_j^t, z_{j^*}^t\})$.

From Lemma 3 (with high probability via a union bound), we know that $z_k^t = O(\sqrt{\frac{\gamma K \ln(NKt)}{t^\gamma}})$. So for any k and j^*, there exists some constant C such that $z_k^t + \max\{z_j^t, z_{j^*}^t\} \leq 2C\sqrt{\frac{\gamma K \ln(NKt)}{t^\gamma}}$. We can then find the threshold time by solving $\Delta_{min} > (1+\alpha)2C\sqrt{\frac{\gamma K \ln(NKt)}{t^\gamma}}$. Then, since $t \leq T$ and $\ln(NKt) \leq \ln(NKT)$, we can use T in our inequality, making it $\Delta_{min} > (1+\alpha)2C\sqrt{\frac{\gamma K \ln(NKT)}{t^\gamma}}$.

Solving, we get:

$$t > [\frac{4C^2(1+\alpha)^2\gamma K \ln(NKT)}{\Delta_{min}^2}]^{\frac{1}{\gamma}}$$

Thus, we can conclude that the exclusion inequality will be true when the above inequality is true, thus proving our bound. For simplicity, we exclude the constant C from our bound, providing the final regret bound for the exploitation phase:

$$O([\frac{(1+\alpha)^2\gamma K \ln(NKT)}{\Delta_{min}^2}]^{\frac{1}{\gamma}})$$

By Corollary 8, Lemma 1 holds simultaneously for all $t \leq T$ with probability at least $1 - \frac{3}{(NK)^3}$. Therefore, the above regret bound holds uniformly across the entire horizon with high probability. □

B.6 Full Procedural Fairness Regret Bound

Proof. Let F_i represent agent i's true favourite-arm set under the true means μ^*. At round t, the algorithm uses the estimates, $\hat{\mu}^t$, and $z_k^t = \sqrt{\frac{2\ln(NKt)}{n_k^t}}$ for arm k at some timestep t, and n_k^t represents the number of times arm k has been pulled by time t.

Additionally, let $\Delta_{i,j,k} = \mu_{i,j}^* - \mu_{i,k}^* > 0$ iff $k \notin F_i$ for any $j \in F_i$ and $k \notin F_i$. Let $\Delta_{min} = min_{i,j\in F_i,k\notin F_i}\Delta_{i,j,k} \geq 0$. Note Δ_{min} is only 0 if all agents are indifferent between all arms, that is, $|F_i| = K$ for all agents i, in which case the problem is trivial for that agent. We also have $\hat{F}_i = \{k : \hat{\mu}_{i,k}^t + \alpha z_k^t \geq \hat{\mu}_{i,j}^t - \alpha z_j^t\}$, where $j = \arg\max_j \hat{\mu}_{i,j}^t$ for some agent i.

In our setting, we define *regret* to refer to the number of mismatches. Specifically, we count a *mismatch* at time t if $\exists i : \hat{F}_i(t) \neq F_i$. Thus:

$$R(T) = \sum_{t=1}^{T} \mathbf{1}\{\exists i : \hat{F}_i(t) \neq F_i\} \leq \underbrace{\sum_{t=1}^{T} p^t_{rand}}_{(a)} + \underbrace{\sum_{t=1}^{T} (1 - p^t_{rand}) \mathbf{1}\{\hat{F}_i(t) \neq F_i \text{ during non-random steps}\}}_{(b)}$$

From Lemma 2, we have that part (a) has the bound $O(\frac{T\gamma}{\gamma})$, and from Lemma 4, we have that part (b) has bound $O([\frac{(1+\alpha)^2 \gamma K \ln(NKT)}{\Delta^2_{min}}]^{\frac{1}{\gamma}})$. Using part (a) and (b) together, we have our final regret bound:

$$R(T) = O(T^\gamma) + O([\frac{(1+\alpha)^2 \gamma K \ln(NKT)}{\Delta^2_{min}}]^{\frac{1}{\gamma}})$$

□

B.7 Procedural Fairness Guarantees

Proof. The highest amount of utility an agent could achieve is if all probability were placed on their favourite arm. Since the procedural fairness policy guarantees each agent at least $1/N$ probability on their favourite arm, each agent receives at least $1/N$ of their maximum achievable utility in expectation, thus giving them a proportional share. This is similar to the concept of *proportionality* [5]. □

B.8 Fractional Pull Bound

Lemma 5 (Fractional Pulls). *Suppose* $z^t_k = \sqrt{\frac{2\ln(NKt)}{n^t_k}}$, *then for either the Equality Fairness or Utilitarian Fairness algorithm:*

$$\sum_{t=1}^{T} \sum_{k=1}^{K} p^t_k z^t_k = O(\sqrt{KT \ln(NKT)}).$$

Proof. Let I^t_k be 1 if arm k is pulled at timestep t, otherwise 0. We want to find a bound for $\mathbb{E}[\sum_{t=1}^{T} \sum_{k=1}^{K} p^t_k \sqrt{\frac{2\ln(NKt)}{n^t_k}}]$.

By definition, we have:

$$\sum_{t=1}^{T} \sum_{k=1}^{K} p^t_k \sqrt{\frac{2\ln(NKt)}{n^t_k}} \leq \sqrt{2\ln(NKT)} \sum_{t=1}^{T} \sum_{k=1}^{K} \frac{p^t_k}{\sqrt{n^t_k}}$$

We want to bound $\mathbb{E}[\sum_{t=1}^{T} \sum_{k=1}^{K} \frac{p^t_k}{\sqrt{n^t_k}}]$.

Because $n^t_k = n^{t-1}_k + I^t_k$, we know that $\frac{1}{\sqrt{n^t_k}} \leq \frac{1}{\sqrt{n^{t-1}_k}}$. Therefore, $\sum_{t=1}^{T} \sum_{k=1}^{K} \frac{p^t_k}{\sqrt{n^t_k}} \leq \sum_{t=1}^{T} \sum_{k=1}^{K} \frac{p^t_k}{\sqrt{n^{t-1}_k}}$.

We know, by definition, that $\mathbb{E}[I^t_k | \text{arm pull history}] = p^t_k$. So $\mathbb{E}[\frac{p^t_k}{\sqrt{n^{t-1}_k}}] = \mathbb{E}[\frac{1}{\sqrt{n^{t-1}_k}} \mathbb{E}[I^t_k | \text{arm pull history}]] = \mathbb{E}[\frac{I^t_k}{\sqrt{n^{t-1}_k}}]$, making our new bound $\mathbb{E}[\sum_{t=1}^{T} \sum_{k=1}^{K} \frac{I^t_k}{\sqrt{n^{t-1}_k}}]$.

Consider the term $\sum_{t=1}^{T} \frac{I^t_k}{\sqrt{n^{t-1}_k}}$. We can reindex the summation and convert into an integral to get a bound:

$$\sum_{t=1}^{T} \frac{I^t_k}{\sqrt{n^{t-1}_k}} = \sum_{m=1}^{n^T_k} \frac{1}{\sqrt{m-1}} \leq 1 + \sum_{m=2}^{n^T_k} \frac{1}{\sqrt{m-1}} \leq 1 + \int_0^{n^T_k - 1} \frac{1}{\sqrt{x}} dx \leq 2\sqrt{n^T_k}$$

Summing over K, we get $\sum_{t=1}^{T} \sum_{k=1}^{K} \frac{I^t_k}{\sqrt{n^{t-1}_k}} \leq 2 \sum_{k=1}^{K} \sqrt{n^T_k}$.

Now we must bound $\sum_{k=1}^{K} \sqrt{n^T_k}$ and then multiply it by the original ln term, and we have our final bound.

Consider the Cauchy-Schwarz inequality: $(\sum_k a_k b_k)^2 \leq \sum_k a^2_k \sum_k b^2_k$. Applying this to our situation gives us: $\sum_{k=1}^{K} \sqrt{n^T_k} \leq \sqrt{KT}$. Combining our prior results, this gives: $\mathbb{E}[\sum_{t=1}^{T} \sum_{k=1}^{K} \frac{I^t_k}{\sqrt{n^{t-1}_k}}] \leq 2\sqrt{KT}$.

Putting it all together, this gives us:

$$\mathbb{E}[\sum_{t=1}^{T}\sum_{k=1}^{K} p_k^t \sqrt{\frac{2\ln{(NKt)}}{n_k^t}}] \leq \sqrt{2\ln{(NKT)}}\mathbb{E}[\sum_{t=1}^{T}\sum_{k=1}^{K}\frac{p_k^t}{\sqrt{n_k^t}}] \leq \sqrt{2\ln{(NKT)}}\cdot 2\sqrt{KT} = 2\sqrt{2KT\ln{(NKT)}} = O(\sqrt{KT\ln{(NKT)}})$$

□

B.9 Equality Fairness Bound

Proof. Recall that μ^* are the true reward means and $\hat{\mu}^t$ are the estimated reward means at timestep t, and that the regret at timestep t, is represented by $r_t = f(p_t, \mu^*) - min_p f(p, \mu^*)$. The total regret over T timesteps is represented by $R^T = \sum_{t=1}^{T} r_t$. From Lemma 1, we have $|\hat{\mu}_{i,k}^t - \mu_{i,k}^*| \leq z_k^t$. Assume $\alpha = 4$.

Now, we want to establish that $|f(p, \hat{\mu}^t) - f(p, \mu^*)| \leq 4\sum_{k=1}^{K} p_k z_k^t$. Let u_i represent the expected utility of agent i under the actual means μ^* and some policy p, and let u_i' represent the expected utility of agent i under the estimated means $\hat{\mu}^t$ and the same policy p. Then, by the Lipschitz bound and triangle inequality we have:

$$\begin{aligned}
|f(p, \hat{\mu}^t) - f(p, \mu^*)| &= \frac{2}{N(N-1)}\left|\sum_{i<j}(u_i - u_j)^2 - \sum_{i<j}(u_i' - u_j')^2\right| \\
&= \frac{2}{N(N-1)}\sum_{i<j}\left|(u_i - u_j)^2 - (u_i' - u_j')^2\right| \\
&\leq \frac{2}{N(N-1)}\sum_{i<j} 2\left|(u_i - u_j) - (u_i' - u_j')\right| \\
&\leq \frac{4}{N(N-1)}\sum_{i<j}\left|u_i - u_i'\right| + \left|u_j - u_j'\right|
\end{aligned}$$

Then, since every index appears exactly $N-1$ times, we have that

$$\sum_{i<j}\left|u_i - u_i'\right| + \left|u_j - u_j'\right| \leq (N-1)\sum_{i}\left|u_i - u_i'\right|$$

Which gives us that $|f(p, \hat{\mu}^t) - f(p, \mu^*)| \leq \frac{4}{N}\sum_i \left|u_i - u_i'\right|$.

Now, by definition of u_i, we know that $u_i - u_i' = \sum_{k=1}^{K} p_k(\mu_{i,k}^* - \hat{\mu}_{i,k}^t)$, and therefore $\left|u_i - u_i'\right| \leq \sum_{k=1}^{K} p_k \left|\mu_{i,k}^* - \hat{\mu}_{i,k}^t\right|$. From our Hoeffding inequality before, we can assume with high probability, that $|\hat{\mu}_{i,k}^t - \mu_{i,k}^*| \leq z_k^t$, so therefore:

$$\left|u_i - u_i'\right| \leq \sum_{k=1}^{K} p_k z_k^t$$

and

$$\sum_{i=1}^{N}\left|u_i - u_i'\right| \leq \sum_{i=1}^{N}\sum_{k=1}^{K} p_k z_k^t = N\sum_{k=1}^{K} p_k z_k^t$$

Therefore:

$$\begin{aligned}
|f(p, \hat{\mu}^t) - f(p, \mu^*)| &\leq \frac{4}{N}\sum_{i<j}\left|u_i - u_i'\right| \\
&\leq \frac{4}{N}N\sum_{k=1}^{K} p_k z_k^t \\
&= 4\sum_{k=1}^{K} p_k z_k^t
\end{aligned}$$

Now we show that the total objective is bounded. By definition of p_t, we know that $f(p_t, \hat{\mu}^t) - \alpha\sum_k p_k^t z_k^t \leq f(p^*, \hat{\mu}^t) - \alpha\sum_k p_k^* z_k^t$ where p^* is the optimal policy (recall that $\hat{\mu}$ is an estimate).

So, we have:

$$f(p_t, \hat{\mu}^t) - \alpha\sum_k p_k^t z_k^t \leq f(p^*, \hat{\mu}^t) - \alpha\sum_k p_k^* z_k^t$$

Adding $-f(p_t, \mu^*) + f(p_t, \mu^*)$ to the LHS and $-f(p^*, \mu^*) + f(p^*, \mu^*)$ to the RHS and rearranging gives us:

$$
\begin{aligned}
f(p_t, \mu^*) - f(p^*, \mu^*) &\le [f(p_t, \mu^*) - f(p_t, \hat{\mu}^t)] + [f(p^*, \hat{\mu}^t) - f(p^*, \mu^*)] + \alpha \sum_k p_k^t z_k^t - \alpha \sum_k p_k^* z_k^t \\
&\le 4 \sum_k p_k^t z_k^t + 4 \sum_k p_k^* z_k^t + \alpha \sum_k p_k^t z_k^t - \alpha \sum_k p_k^* z_k^t \quad \text{(By earlier Lipschitz bound)} \\
&\le (4+\alpha) \sum_k p_k^t z_k^t + (4-\alpha) \sum_k p_k^* z_k^t \quad \text{(Set } \alpha = 4 \text{ to kill the second term)} \\
&= 8 \sum_k p_k^t z_k^t
\end{aligned}
$$

Since $R^T \le 8 \sum_{t=1}^T \sum_{k=1}^K p_k^t z_k^t = 8 \sum_{k=1}^K \sum_{t=1}^T p_k^t z_k^t$, we now find a bound on $\sum_{k=1}^K \sum_{t=1}^T p_k^t z_k^t$, which from Lemma 5 we know is $O(\sqrt{KT \ln{(NKT)}})$, thus completing our proof. □

B.10 Utilitarian Fairness Bound

Proof. Let $M_k = \sum_{i=1}^N \mu_{i,k}^*$, $M^* = \max_k M_k$, $\Delta_k = M^* - M_k$. At round t, the algorithm chooses a distribution p^t by solving $\max_p \sum_k p_k(\widehat{M}_k^t + \alpha z_k^t)$, $\widehat{M}_k^t = \sum_{i=1}^N \widehat{\mu}_{i,k}$, $z_k^t = \sqrt{\frac{2\ln{(NKt)}}{n_k^t}}$, where the regret at time t is defined as $r_t = M^* - \sum_k p_k^t M_k$, and total regret is $R_T = \sum_{t=1}^T r_t$. Assume $\alpha \ge N$.

By Lemma 1, we have with high probability that $|\widehat{M}_k^t - M_k| \le N z_k^t$.

By definition, at round t, the algorithm selects a distribution such that $\sum_k p_k^t(\widehat{M}_k^t + \alpha z_k^t) \ge \widehat{M}_{k^*}^t + \alpha z_{k^*}^t$, where k^* is the actual best arm under the true means. Let $\widehat{M}_k^t = M_k + \varepsilon_k^t$, where $\varepsilon_k^t := \widehat{M}_k^t - M_k$. From earlier we have that $|\varepsilon_k^t| \le N z_k^t$. This gives us:

$$
\sum_k p_k^t(\widehat{M}_k^t + \alpha z_k^t) \ge \sum_k p_k^*(\widehat{M}_k^t + \alpha z_k^t)
$$

Where p^* is an optimal policy. Rearranging this equation gives:

$$
\begin{aligned}
\sum_k p_k^* \widehat{M}_k^t - \sum_k p_k^t \widehat{M}_k^t &= \sum_k \widehat{M}_k^t \,(p_k^* - p_k^t) && (1)\\
&\le \alpha \left(\sum_k p_k^t z_k^t - \sum_k p_k^* z_k^t \right) && (2)\\
&= \alpha(D_t - D_t^*) && (3)
\end{aligned}
$$

With $D_t = \sum_k p_k^t z_k^t$ and $D_t^* = \sum_k p_k^* z_k^t$. Recall that regret is defined as $r_t = M^* - \sum_k p_k^t M_k$, expanding gives us $r_t = \sum_k p_k^* M_k - \sum p_k^t M_k = \sum_k M_k(p_k^* - p_k^t)$.

By definition, we have that $M_k = \widehat{M}_k^t - \epsilon_k^t$, so plugging in we get:

$$
\begin{aligned}
r_t &= \sum_k M_k(p_k^* - p_k^t) && (4)\\
&= \sum_k (\widehat{M}_k^t - \epsilon_k^t)(p_k^* - p_k^t) && (5)\\
&= \sum_k \widehat{M}_k^t(p_k^* - p_k^t) - \sum_k \epsilon_k^t(p_k^* - p_k^t) && (6)\\
&\le \alpha(D_t - D_t^*) - \sum_k \epsilon_k^t(p_k^* - p_k^t) && (7)
\end{aligned}
$$

Recall that $|\epsilon_k^t| \le N z_k^t$ by Lemma 1, so we have that $|\sum_k \epsilon_k^t(p_k^* - p_k^t)| \le \sum_k |\epsilon_k^t||(p_k^* - p_k^t)| \le \sum_k N z_k^t |(p_k^* - p_k^t)|$.
Since p_k^* and p_k^t and always at least 0, we have that $|p_k^* - p_k^t| \le p_k^* + p_k^t$. This gives: $\sum_k N z_k^t |(p_k^* - p_k^t)| \le \sum_k N z_k^t (p_k^* + p_k^t) = N(D_t^* + D_t)$.
Putting it all together, we have:

$$r_t \leq \alpha(D_t - D_t^*) - \sum_k \epsilon_k^t (p_k^* - p_k^t) \tag{8}$$

$$\leq \alpha(D_t - D_t^*) + \left|\sum_k \epsilon_k^t (p_k^* - p_k^t)\right| \tag{9}$$

$$\leq \alpha(D_t - D_t^*) + N(D_t^* + D_t) \tag{10}$$

$$= (N + \alpha)D_t + (N - \alpha)D_t^* \tag{11}$$

Set $\alpha \geq N$ to drop the second term, leaving us with $r_t \leq (N + \alpha)D_t = (N + \alpha) \sum_k p_k^t z_k^t$.

Summing over all t, we get $R_T = \sum_{t=1}^T r_t \leq (N + \alpha) \sum_{t=1}^T \sum_k p_k^t z_k^t$. From lemma 5, we know that $\mathbb{E}[\sum_{t=1}^T \sum_k p_k^t z_k^t] = O(\sqrt{KT \ln(NKT)})$, therefore, $\mathbb{E}[R_T] = O((N + \alpha)\sqrt{KT \ln(NKT)})$, our desired bound.

□

B.11 Procedural Fairness and Equality Fairness Impossibility

PROOF. Consider a simple setting with two agents and two arms with rewards:

$$\mu = \begin{bmatrix} M & 0 \\ 0 & 1 \end{bmatrix}$$

where $M > 1$. Under procedural fairness, each agent must allocate $\frac{1}{2}$ probability to their preferred arm, yielding policy $P_1 = (\frac{1}{2}, \frac{1}{2})$ with expected rewards of $\frac{M}{2}$ and $\frac{1}{2}$ for the two agents. This creates a utility disparity that grows with M.

For equality fairness, we require equal expected rewards, which is achieved by $P_2 = (\frac{1}{M+1}, \frac{M}{M+1})$. However, this allocates only $\frac{1}{M+1}$ probability to arm 1, significantly less than the $\frac{1}{2}$ required by procedural fairness when M is large. As $M \to \infty$, the procedural fairness score approaches zero.

Therefore, no policy can simultaneously satisfy both fairness criteria for all reward structures. □

B.12 Procedural Fairness and Utilitarian Fairness Impossibility

PROOF. Consider a simple setting with two agents and two arms with rewards:

$$\mu = \begin{bmatrix} M & 0 \\ 0 & 1 \end{bmatrix}$$

where $M > 1$. Under procedural fairness, each agent will allocate $\frac{1}{2}$ probability to their preferred arm, which will yield $P_1 = (\frac{1}{2}, \frac{1}{2})$ with expected rewards of $\frac{M}{2}$ and $\frac{1}{2}$ for the two agents and an expected total utility of $\frac{M+1}{2}$. However, the policy P_2 which maximizes overall utility and satisfies perfect utilitarian fairness would be $P_2 = (1, 0)$ which would result in an expected total utility of M. However, $M > \frac{M+1}{2}$ for all values of $M > 1$, leading to a disparity.

□

B.13 Utility-Based Nash Welfare Not In Procedural Core

PROOF. We prove this through a counter-example. Consider a setting with two agents ($N = 2$) and two arms ($K = 2$). The reward matrix

$$\mu = \begin{pmatrix} 1 & 0.99 \\ 0 & 1 \end{pmatrix}$$

Agent 1's favourite arm is arm 1, and Agent 2's favourite arm is arm 2. We will now find the distribution over arms that maximizes the utility-based Nash welfare.

Let p_2 be the probability of pulling arm 2 and $p_1 = 1 - p_2$ be the probability of pulling arm 1. Then:

$$\text{Agent 1's expected reward:} \quad u_1(p_1, p_2) = 1 \cdot p_1 + 0.99 \cdot p_2 = p_1 + 0.99\, p_2,$$

$$\text{Agent 2's expected reward:} \quad u_2(p_1, p_2) = 0 \cdot p_1 + 1 \cdot p_2 = p_2.$$

Utility-based Nash Welfare (the product of expected rewards) is

$$\Big(p_1 + 0.99\, p_2\Big) \cdot p_2 = (p_1)\, p_2 + 0.99\, p_2^2.$$

Since $p_1 = 1 - p_2$, this equals

$$(1 - p_2)p_2 + 0.99\, p_2^2 = p_2 - p_2^2 + 0.99\, p_2^2 = p_2 - 0.01\, p_2^2.$$

Taking the derivative:

$$\frac{d}{dp_2}\left(p_2 - 0.01\,p_2^2\right) \;=\; 1 \;-\; 0.02\,p_2.$$

On the interval $p_2 \in [0, 1]$, this derivative never vanishes (it is $1 - 0.02\,p_2 > 0$ for all $p_2 \in [0, 1]$). Hence the function is strictly increasing over $[0, 1]$, with its maximum at the boundary $p_2 = 1$.

Thus the unique maximizer of utility-based Nash Welfare is

$$(p_1, p_2) \;=\; (0,\; 1).$$

We can now show that this distribution is not in the procedural core. Under $(0, 1)$, the *procedural utility* of an agent is the total probability on that agent's favourite arm(s). Thus:

$$u_1^{(\text{proc})}\big((0, 1)\big) = p_1 = 0$$

$$u_2^{(\text{proc})}\big((0, 1)\big) = p_2 = 1$$

We can look at the single-agent coalition $C = \{\text{Agent } 1\}$. By deviating to the distribution $(1, 0)$ (which puts probability 1 on arm 1), Agent 1's procedural utility becomes

$$u_1^{(\text{proc})}\big((1, 0)\big) = 1.$$

Since $\frac{|C|}{N} = \frac{1}{2}$, we scale this by $\frac{1}{2}$ to obtain

$$\tfrac{|C|}{N} \cdot u_1^{(\text{proc})}\big((1, 0)\big) = \tfrac{1}{2} \times 1 = 0.5 \quad > \quad 0 = u_1^{(\text{proc})}\big((0, 1)\big).$$

Hence Agent 1 alone can *strictly* increase their procedural utility when switching from $(0, 1)$ to $(1, 0)$. By definition of the procedural core, $(0, 1)$ is therefore *blocked* and cannot be in the procedural core. Thus, the unique distribution maximizing utility-based Nash Welfare in this instance, $(0, 1)$, is not in the procedural core because a single-agent coalition has a profitable deviation. This shows that the utility-based Nash welfare maximizing solution is not guaranteed to be in the procedural core. □

B.14 Procedural Fairness Implies Procedural Core

Proof. Let x be a procedurally fair policy that maximizes the decision share-based Nash product, namely:

$$\prod_{i=1}^{n} \sum_{j \in F_i} p_j$$

for some distribution $P = (p_1, ..., p_K) \in \Delta^k$ Further, let $u_i(x)$ represent the procedural utility obtained by agent i under the policy x. In other words, $u_i(x) = \sum_{j \in F_i} p_j$.

Suppose, for the sake of contradiction, that some coalition $C \subseteq N$ can block x by switching to some other procedurally fair policy y. In order to block, the following must be satisfied:

$$\frac{|C|}{N} u_i(y) \geq u_i(x) \forall i \in C$$

with at least one strict equality. Then, let $\alpha = \frac{|C|}{N}$ and

$$z = \alpha y + (1 - \alpha) x$$

Since both x and y fall within the set of policies that satisfy the procedural fairness constraints (each agent places $\frac{1}{N}$ on their favourite arms, and total mass sums to 1), so does their convex combination z. Since we have that $\frac{|C|}{N} u_i(y) \geq u_i(x)$, we also have that

$$u_i(z) = \alpha u_i(y) + (1 - \alpha) u_i(x) \geq u_i(x)$$

with at least one strict inequality. This gives us:

$$\prod_{i \in N} u_i(z) > \prod_{i \in N} u_i(x)$$

because at least one factor strictly increased while all others stayed the same, and we know, by definition, since we are using procedurally fair policies that no agent will have a utility of 0.

However, x was specifically chosen to maximize $\prod_{i \in N} u_i(x)$, which is a contradiction. Therefore, the procedural fairness policy with decision share-based Nash welfare tie-breaking is in the procedural core.

□

B.15 Pareto Incomparability

Proof. Proof by counterexample. Consider the following utility matrix:

$$\mu = \begin{bmatrix} A & 1 \\ 1 & B \end{bmatrix}$$

Where $0 \leq A < B < 1$. Thus, we know for certain that the procedurally fair policy here $P_{PF} = (\frac{1}{2}, \frac{1}{2})$. Further, we also know that the utilitarian fair policy would be $P_{UF} = (0, 1)$, as $A < B$. Thus, the utility-based Nash welfare solution would be Pareto incompatible with the procedurally fair solution if the probability it places on the rightmost arm is strictly greater than 0.5 but strictly less than 1.

Let p represent the probability we pull the leftmost arm, and $1 - p$ denote the probability we pull the rightmost arm. Let $P_{NW} = (p, 1 - p)$. Then, for each agent, we have utilities:

$$U_1(P_{NW}) = p(A - 1) + 1$$

$$U_2(P_{NW}) = B + p(1 - B)$$

Then, to find the optimal utility-based Nash welfare solution, we want to optimize:

$$f(P_{NW}) = U_1(P_{NW})U_2(P_{NW}) = (p(A - 1) + 1)(B + p(1 - B))$$

$$f'(P_{NW}) = B(A - 1) + 2p(1 - B)(A - 1) + 1 - B = 0$$

$$p = \frac{B(2 - A) - 1}{2(1 - B)(A - 1)}$$

Let $A = 0.4$ and $B = 0.6$, which satisfies our constraints from earlier. Then, $p = \frac{1}{12}$ and $1 - p = \frac{11}{12} > \frac{1}{2}$, but less than 1. We can therefore derive, that:

$$PF(P_{PF}, \mu) = 1 > PF(P_{NW}, \mu)$$

and

$$UF(P_{NW}, \mu) > UF(P_{PF}, \mu)$$

Therefore, neither policy can guarantee Pareto dominance over the other.

□