Bi-Level Contextual Bandits for Individualized Resource Allocation Under Delayed Feedback

Mohammadsina Almasi, Hadis Anahideh*

University of Illinois Chicago
Chicago, IL, 60607 USA
{malmas6,hadis}@uic.edu

Abstract

Equitably allocating limited resources in high-stakes domains—such as education, employment, and healthcare—requires balancing short-term utility with long-term impact, while accounting for delayed outcomes, hidden heterogeneity, and ethical constraints. However, most learning-based allocation frameworks either assume immediate feedback or ignore the complex interplay between individual characteristics and intervention dynamics. We propose a novel bi-level contextual bandit framework for individualized resource allocation under delayed feedback, designed to operate in real-world settings with dynamic populations, capacity constraints, and time-sensitive impact. At the meta level, the model optimizes subgroup-level budget allocations to satisfy fairness and operational constraints. At the base level, it identifies the most responsive individuals within each group using a neural network trained on observational data, while respecting cooldown windows and delayed treatment effects modeled via resource-specific delay kernels. By explicitly modeling temporal dynamics and feedback delays, the algorithm continually refines its policy as new data arrive, enabling more responsive and adaptive decision-making. We validate our approach on two real-world datasets from education and workforce development, showing that it achieves higher cumulative outcomes, better adapts to delay structures, and ensures equitable distribution across subgroups. Our results highlight the potential of delay-aware, data-driven decision-making systems to improve institutional policy and social welfare.

Code — https://github.com/sinatorrr/MAB

Introduction

Resource allocation is a central challenge in high-stake domains such as healthcare, where practitioners determine treatment priorities (Lane et al. 2017; Aktaş, Ülengin, and Şahin 2007; Daniels et al. 2016); telecommunications, where capacity must be distributed across competing channels (Su et al. 2019; Hui 2002; Gibney and Jennings 1998); education, where instructional or financial resources are allocated to students (Monk 1981; Liefner 2003; Massy 1996); and social welfare, where job training and support programs are delivered (Nguyen et al. 2014; Roos and Rothe 2010). In these settings, decision-makers often observe contextual information at the individual level and must sequentially allocate scarce interventions to optimize long-term, population-wide outcomes (Hegazy 1999; Gong et al. 2012).

Classical resource allocation models provide useful abstractions but rely on idealized assumptions, often overlooking temporal, ethical, and institutional constraints that affect fairness, feasibility, and policy relevance in real-world settings (Wang et al. 2022; Zou et al. 2019; Obermeyer et al. 2019; Chouldechova 2017). To address these limitations, recent research has leveraged algorithmic frameworks such as multi-armed bandits (MABs), particularly their contextual variants, to make personalized, adaptive decisions based on observed features (Grover et al. 2018; Gyorgy and Joulani 2021; Joulani, Gyorgy, and Szepesvári 2013). However, several critical challenges remain unaddressed in the literature.

First, most existing MAB approaches assume that outcomes are observed immediately following an allocation. In reality, the effects of interventions unfold gradually: medical treatments manifest their efficacy over days or weeks (Hanna et al. 2020; Yanovski and Yanovski 2014), educational interventions accrue impact over semesters (Barnett 1995; Almalki and Mohammed 2022), and workforce programs influence long-term employment trajectories (Edmondson, Kern, and Rogge 2019). These delayed and temporally structured effects introduce feedback dynamics that are rarely modeled in full. While recent methods incorporate delay via fixed or stochastic lags (Lancewicki et al. 2021; Shi, Wang, and Wu 2023), most treat delay as a nuisance rather than learning the temporal impact profile of each intervention. To overcome these limitations, studies have introduced delay-aware allocation methods, including post hoc reward adjustments to capture deferred effects (Tang, Ho, and Liu 2021). Another line of work uses an episodic framework, modeling feedback delays as discrete random variables representing decision rounds between action and outcome. However, these models often assume fixed or context-independent lags, limiting their ability to capture heterogeneous, intervention-specific temporal dynamics (Kuang et al. 2023; Yin et al. 2023).

Second, traditional models assume a static population and ignore real-world deployment constraints. In practice, participants join and leave in cohort cycles (e.g., semesters or

*Corresponding Author.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

enrollment periods), creating time-varying populations that challenge fixed decision-pool assumptions. They also overlook ethical and institutional rules like cooldown periods restricting repeated allocations of the same resource to the same individual—constraints essential for fairness and feasibility (Li and Varakantham 2022; Patil et al. 2021). While some studies have begun to incorporate these elements, they typically do so under highly stylized conditions—assuming zero delay, a single homogeneous resource type, or i.i.d. reward structures (Wang et al. 2019; Zuo and Joe-Wong 2021; Burnetas, Kanavetas, and Katehakis 2025). Such simplifications fail to capture the structural dependencies and heterogeneity that characterize real-world allocation environments. Most existing algorithms target either individual personalization or group fairness, rarely both. Some optimize personalized rewards but ignore group equity; others enforce fairness while overlooking individual heterogeneity (Li and Varakantham 2022; Patil et al. 2021). Bridging both is essential for equitable, effective policy in settings requiring individual adaptation and group awareness.

In light of these challenges, we propose **Meta-level Contextual Upper Confidence Bandit (MetaCUB)** a novel bi-level contextual bandit framework for individualized resource allocation under delayed feedback and real-world constraints. At its core is a neural network that maps individual-level features to subgroup-level treatment effects, capturing latent heterogeneity in responsiveness. The meta level allocates sub-budgets across groups to ensure equity, while the base level selects the most responsive individuals within each group under resource-specific constraints such as cooldowns and budgets. Our framework models cohort-driven population dynamics, multiple resource types with distinct budgets, and heterogeneous delay-feedback profiles, where each resource has a delay kernel describing its temporal impact. Evaluations on real-world datasets show that **MetaCUB** learns effective, constraint-aware allocation policies that outperform strong baselines.

In summary, our contributions are fourfold. First, we propose a neural network–based learning framework that maps individual contexts to subgroup effects, capturing latent heterogeneity and improving regret over linear models. Second, we design a bi-level contextual bandit that jointly optimizes group- and individual-level allocations under fairness, cooldown, and capacity constraints. Third, we build a deployment-ready architecture modeling cohort dynamics, resource-specific budgets, delay kernels, and stochastic cooldowns. Finally, we validate our framework through extensive experiments on real-world datasets, showing superior cumulative reward, fairness, and delay adaptation over state-of-the-art baselines.

Related Works

The multi-armed bandit (MAB) framework is a foundational model for sequential decision-making under uncertainty, widely used for efficient resource allocation in dynamic settings (Kuleshov and Precup 2014; Agrawal and Goyal 2012). It has shown effectiveness in domains like telecommunications, finance, online platforms, and healthcare, where adaptive learning is critical (Huo and Fu 2017; Biswas et al. 2021; Bouneffouf, Rish, and Aggarwal 2020).

Contextual bandits extend classical MABs by leveraging individual-level covariates to enable personalized decision policies (Lu, Pál, and Pál 2010; Zhou 2015). Methods like LinUCB and contextual Thompson Sampling provide theoretical regret guarantees when rewards depend linearly or probabilistically on covariates (Agrawal and Goyal 2013; Kaufmann, Cappé, and Garivier 2012; Chouldechova 2017). These models have advanced personalization in areas such as online recommendations, clinical decisions, and educational interventions. Recent extensions to multi-agent and group-based learning enable collaborative and distributed allocation across multiple learners or subpopulations (Cui, Liu, and Nallanathan 2019; Xu, Tao, and Shen 2020).

Despite this progress, most contextual bandit approaches remain difficult to deploy effectively in socially impactful domains. First, they often assume immediate feedback, overlooking that real-world interventions—such as tutoring, job training, or healthcare treatments—produce delayed effects over time. Although recent studies address stochastic or bounded delays (Joulani, Gyorgy, and Szepesvári 2013; Gael et al. 2020), they often treat delay as a nuisance variable, neglecting its temporal dynamics. Several works have proposed tracking reward queues (Tang, Ho, and Liu 2021; Vernade, Cappé, and Perchet 2017), bounding adversarial regret (Erez, Levy, and Mansour 2024; Steiger, Li, and Lu 2022), coupling delay with payoff magnitude (Schlisselberg et al. 2025) but few frameworks model how reward signals are distributed across time in a resource-specific and learnable manner. Second, while constrained MABs, such as bandits with knapsacks (Badanidiyuru, Kleinberg, and Slivkins 2018; Tran-Thanh et al. 2012) or fairness-aware variants (Chen et al. 2020; Claure et al. 2020), address limitations on budgets or equity, they often assume static populations and homogeneous reward structures. In practice, many allocation settings involve dynamic cohorts, where individuals enter and exit over time (e.g., educational semesters or batched workforce programs). Some studies have examined the concept of a dynamic population, which captures the changing availability of arms, through frameworks such as contextual combinatorial bandits with volatile arms and submodular rewards, or interest-drift models with immediate feedback; however, these approaches overlook the partial observability of feedback inherent in educational and workforce outcomes (Chen, Xu, and Lu 2018; Xu et al. 2020). Additionally, real-world deployments must satisfy cooldown constraints, such as ethical limits on repeated treatment (Liu, Liu, and Zhao 2012; Chen, Liew, and Shao 2022; Mate et al. 2022), yet such pacing mechanisms are seldom integrated into existing MAB formulations; for example, classical blocking methods address availability pacing but ignore delayed impact (Basu et al. 2021).

Third, prior works often separate fairness from personalization. Some optimize individual outcomes without ensuring group equity, while others enforce group fairness with no within-group heterogeneity. Few models integrate both, enabling individualized treatment within group-level budget constraints under delayed feedback and dynamic population.

Problem Setup

We consider a sequential decision-making problem where a central planner (e.g., policymaker or service provider) allocates limited resources over time to individuals grouped by demographic or socioeconomic traits. The goal is to optimize long-term outcomes—like academic or employment success—via adaptive, context-aware decisions accounting for delayed effects and real-world constraints. This setting appears in high-stakes domains such as education (e.g., financial aid, tutoring) and workforce programs (e.g., training support). Our goal is to design an allocation framework for evolving, constrained, and fairness-sensitive environments.

Consider a set of N individuals of K demographic subgroups of size n^k each and $N = \sum_{k=1}^{K} n^k$. Each individual is associated with a context vector $\mathbf{x}^i \in \mathcal{X} \subseteq \mathbb{R}^M$, capturing M demographic and domain-specific attributes, including their recent resource assignments. The decision-maker manages R distinct resource types, each with an integer-valued budget $b^r \in \mathbb{Z}_+$ for $r \in R$. The allocation process unfolds over T discrete time steps. Let $\mathcal{I} \subseteq N$ denote the subset of individuals who receive at least one allocation during the decision horizon. At each decision round $t \in T$, one individual $i_t \in \mathcal{I}$ is selected and assigned a unit of resource $r_t \in R$, subject to the budget constraint:

$$\sum_{t=1}^{T} \mathbb{I}\{r_t = r\} \;\le\; b^r \quad \forall\, r \in R \tag{1}$$

At each round t, the action taken is the pair $a_t = (i_t, r_t)$, where $i_t \in \mathcal{I}$ is the selected individual and $r_t \in R$ is the assigned resource. The full allocation sequence over the horizon is denoted by $\{a_t\}_{t=1}^{T}$. Allocating a resource to an individual yields an instantaneous reward $y(t) = f\big(\mathbf{x}^i(t)\big)$, where $f : \mathcal{X} \to \mathbb{R}$ is a reward function mapping the individual's context, possibly including the allocated resource, to an expected outcome[1]. The decision-maker's objective is to select a sequence of actions $\{a_t = (i_t, r_t)\}_{t=1}^{T}$ that maximizes the expected cumulative reward over the time horizon:

$$\max_{\{a_t\}_{t=1}^{T}} \; \mathbb{E}\Big[\sum_{t=1}^{T} y(t)\Big] \tag{2}$$

Equivalently, the goal can be framed as minimizing the cumulative regret relative to the best feasible allocation policy in hindsight, i.e., the optimal policy that would have been chosen with full knowledge of individual responses.

A natural and flexible framework for modeling sequential resource allocation under uncertainty is the contextual multi-armed bandit (MAB) (Lu, Pál, and Pál 2010) In this setting, each feasible action—defined as an individual–resource pair $a_t = (i_t, r_t)$ is treated as an arm. The associated context $\mathbf{x}^i(t)$ provides side information about the individual and their history, while a pretrained reward function $f(\cdot)$ serves as the feedback model, predicting expected outcomes for each action. Resource budgets impose constraints on the number of allowable arm pulls, and the objective becomes minimizing regret relative to the best allocation policy in hindsight A common way to balance this trade-off is via Upper Confidence Bound (UCB) methods, which choose actions maximizing predicted reward plus an uncertainty bonus, balancing exploration and exploitation. A baseline constrained contextual MAB procedure is outlined in Algorithm 3 in Appendix. We extend it with additional components to better capture real-world complexities.

Population Change. In many deployment settings, participants enroll and exit in fixed-duration cycles, forming successive distinct cohorts. To capture this, we partition the decision horizon of T rounds into $H = \lceil T/L \rceil$ contiguous blocks of length L. Let $\mathcal{I}_h \subseteq N$ denote the set of individuals in cohort h, who are eligible to receive allocations only during rounds $t \in [(h-1)L + 1,\ hL],\ h = 1, \ldots, H$. At the beginning of each block h, the decision-maker observes the context vectors $\{\mathbf{x}^i(t) : i \in \mathcal{I}_h\}$ and allocates resources exclusively among individuals in cohort $\mathcal{I}_h$ for the next L rounds. At the end of this period, cohort $\mathcal{I}_h$ exits the program and is replaced by the incoming cohort $\mathcal{I}_{h+1}$. This structure introduces non-stationarity into the decision process, as the available pool of individuals varies across time. The algorithm must therefore learn not only whom to allocate resources to, but also adapt its policy to the evolving population across cohorts.

Delayed Feedback. Resource allocations in real-world scenarios often exhibit delayed effects. We model resource-specific feedback delays over a horizon of T rounds. For each resource $r \in R$, we define a delay kernel K^r, a non-negative function over the time horizon that distributes the realized reward across future rounds. Formally, let $K^r : \{0, \ldots, T-1\} \to [0, 1], \quad \sum_{\tau=0}^{T-1} K^r(\tau) = 1$, where $K^r(\tau)$ denotes the proportion of the reward from allocating resource r that is observed τ rounds after the allocation. By definition, $K^r(\tau) = 0$ for $\tau < 0$ or $\tau > T-1$, ensuring bounded support. To construct these kernels, we discretize a Beta distribution over $[0, 1]$ into T equal-width bins. In particular,

$$K^r(\tau) = \int_{\frac{\tau}{T}}^{\frac{\tau+1}{T}} \mathrm{Beta}(z; \alpha^r, \beta^r)\, dz \tag{3}$$

for $\tau = 0, 1, \ldots, T-1$, where $\mathrm{Beta}(z; \alpha, \beta) = \frac{z^{\alpha-1}(1-z)^{\beta-1}}{B(\alpha,\beta)}$ for $z \in (0, 1)$, and $B(\alpha, \beta) = \int_0^1 z^{\alpha-1}(1 - z)^{\beta-1}\, dz$. This formulation flexibly models feedback latency: $\alpha^r < 1$ gives immediate feedback, $\beta^r < 1$ produces long-tail delays, and $\alpha^r, \beta^r > 1$ yield unimodal kernels. Mixtures of Beta densities can represent more complex or multimodal delays. Discretized Beta kernels assign unit mass to $\{0, \cdots, T-1\}$ and flexibly capture early, late, or long-tailed delays. They attribute outcomes to service rounds without leakage beyond the operational window, aligning with program accounting. For each resource, (α, β) parameters are chosen from plausible timing profiles and fixed during learning. The framework remains distribution-agnostic, any normalized delay kernel is admissible, and supports adaptive or meta-learned kernel estimation when greater flexibility is needed (Kassraie, Rothfuss, and Krause 2022). Overall, this kernel-based framework extends be-

[1]The reward function can be adapted to the application domain; for binary outcomes, for example, one may use $f : \mathcal{X} \to [0, 1]$.

yond fixed delays or exponential decay, capturing heterogeneous, resource-dependent feedback dynamics that mirror real-world interventions.

Allocation Cooldown. In real-world interventions, individuals are rarely allowed, or advised, to receive the same resource repeatedly in short intervals (Weiner et al. 2012; Légaré et al. 2018). Treatment effects take time to manifest, capacity is limited, and regulations often restrict repeated support. To model this, we introduce *cooldown* constraints that prevent reallocation of the same resource to an individual for several rounds after use.

Let $c^r \in \mathbb{Z}_+$ with $c^r < T$ denote the cooldown length, i.e., the number of rounds during which an individual is ineligible to receive the same resource $r \in R$ again. When $c^r = T$, the cooldown spans the full horizon, limiting each individual to at most one allocation of r over the T rounds. After individual $i \in \mathcal{I}_h$ receives resource $r \in R$ at round $t \in T$, they become temporarily ineligible to receive the same resource again for the next c^r consecutive rounds. Formally, let $z_{i,r}(t) \in \{0,1\}$ whether resource r is allocated to individual i at round t. That is we impose the following constraint:

$$\sum_{s=t}^{t+c^r} z_{i,r}(s) \leq 1 \; \forall i \in \mathcal{I}_h, \; \forall r \in R, \; t = 1, \ldots, T - c^r \quad (4)$$

This condition limits each individual to one unit of resource r within any $c^r + 1$ consecutive rounds. While the delay kernel K^r models reward evolution, the cooldown enforces allocation spacing, jointly forming a temporally aware framework that balances impact and pacing.

Proposed Approach

We extend the base model in Equation (2) to include *population change*, *delayed feedback*, and *allocation cooldowns*. Keeping binary decisions $z_{i,r}(t) \in 0, 1$ and budget limits (Equation (1)), we add time-varying eligibility for cohort dynamics, cooldowns (Equation (4)) to control repeated allocations, and cumulative rewards reflecting temporally distributed effects through resource-specific delay kernels. At each decision round t, the observed reward $y(t)$ aggregates the delayed impacts of all past allocations whose effects materialize at time t. Formally, the reward is computed as:

$$y(t) = \sum_{u=1}^{t} \sum_{i \in \mathcal{I}_h(u)} \sum_{r \in R} K^r(t-u) \, f\left(\mathbf{x}^i(u)\right) z_{i,r}(u). \quad (5)$$

where $K^r(\cdot)$ is the delay kernel associated with resource r, and $f(\cdot)$ is a learned model that maps context vectors to predicted outcomes. The full problem is then formulated as the following constrained optimization program. We use the following shorthand: "1/round" for one resource per individual per round, "B" for total resource budget constraints, and "CD" for cooldown restrictions on repeated allocations.

$$\max_{z_{i,r}(t)} \quad \mathbb{E}\left[\sum_{t=1}^{T} y(t)\right] \quad (6a)$$

s.t.

$$\sum_{r \in R} z_{i,r}(t) \leq 1 \qquad \forall i, t \quad \text{(1/round)} \quad (6b)$$

$$\sum_{t=1}^{T} \sum_{i \in \mathcal{I}_h(t)} z_{i,r}(t) \leq b^r \quad \forall r \quad \text{(B)} \quad (6c)$$

$$\sum_{s=t}^{t+c^r} z_{i,r}(s) \leq 1 \qquad \forall i, r, \; t = 1, \ldots, T - c^r \quad \text{(CD)} \quad (6d)$$

$$z_{i,r}(t) \in \{0,1\} \qquad \forall i, r, t$$

To solve the extended problem formulation in (6)a-d, we propose **MetaCUB** (Meta level Contextual Upper Confidence Bandit), a bi-level contextual bandit optimization framework. At the upper level, a meta-bandit allocates fractional resource budgets across demographic groups to maximize population-wide impact under equity constraints. At the lower level, an individual bandit selects individuals in each group using contextual features and a learned mapping from profiles to expected outcomes (e.g., mean GPA).

Meta-level Framework. We assume a fixed computational budget of T_m iterations at the meta-level. At each iteration $t_m \in \{1, \ldots, T_m\}$, the meta-level algorithm selects a candidate meta-allocation policy $\bar{\boldsymbol{z}}(t_m) = \left\{\bar{z}_r^k(t_m)\right\}_{k \in K, r \in R} \in \widetilde{\Delta}^{|K| \cdot |R|}$, where $\Delta^{|K| \cdot |R|}$ denotes the $(|K| \cdot |R|)$-dimensional probability simplex (i.e.,)the non-negative vectors in $\mathbb{R}^{|K| \cdot |R|}$ that sum to 1.) ensuring that the total resource allocation across all subgroup–resource pairs remains normalized. Each entry $\bar{\boldsymbol{z}}_r^k(t_m)$ specifies the proportion of the total resource budget allocated to subgroup $k \in K$ for resource type $r \in R$ at iteration t_m. The policy must satisfy the following simplex constraint:

$$\sum_{k \in K} \sum_{r \in R} \bar{\boldsymbol{z}}_r^k(t_m) \leq 1 \quad \forall t_m \in \{1, ..., T_m\} \quad (7)$$

To initiate the optimization, we generate an initial set of n_0 candidate meta-allocation policies $\{\bar{\boldsymbol{z}}^{(j)}\}_{j=1}^{n_0}$ sampled from the interior of the simplex. For each candidate policy $\bar{\boldsymbol{z}}^{(j)}$, we simulate the subgroup-level outcomes by randomly assigning individuals within each group to resource types according to the respective sub-budgets $\bar{\boldsymbol{z}}_r^{k(j)}$, using the shared learned outcome model f to compute predicted individual outcomes. We then compute the mean predicted outcome $\mu_r^k(\bar{\boldsymbol{z}})$ for each group-resource pair and aggregate them into a global utility score:

$$\hat{y}(\bar{\boldsymbol{z}}) = \sum_{k \in K} \sum_{r \in R} \bar{\boldsymbol{z}}_r^k \cdot \mu_r^k(\bar{\boldsymbol{z}}). \quad (8)$$

Instead of fitting and retraining a separate Gaussian Process (GP) over the high-dimensional meta-policy space, we reuse the learned outcome model f-which maps individual context to expected reward-as a simulation-based surrogate. This captures the functional relationship between allocation decisions and observed outcomes, enabling fast and scalable evaluation of candidate meta-policies $\bar{\boldsymbol{z}}$.

For any candidate $\bar{\boldsymbol{z}}$, we simulate assignments by probabilistically distributing resources within each subgroup in proportion to $\bar{z}_r^k$ and then use f to predict individual outcomes. The resulting subgroup-level predictions are aggregated to estimate the overall utility $\hat{y}(\bar{\boldsymbol{z}})$. This surrogate approach avoids the computational burden of GP inference in high dimensions while leveraging the contextual expressivity of f, which captures latent structure across individuals and resources. At each round t_m, to select the next candidate meta-policy, we adopt an UCB acquisition rule adapted to this simulation setting. Specifically, we estimate the posterior mean $\mu(\bar{\boldsymbol{z}})$ and empirical standard deviation $\sigma(\bar{\boldsymbol{z}})$ over multiple stochastic rollout simulations. The next meta-policy is then selected as the one maximizing the acquisition score:

$$\bar{\boldsymbol{z}}(t_m) = \underset{\bar{\boldsymbol{z}} \in \widetilde{\Delta}^{|K|\cdot|R|}}{\arg\max} \left(\mu(\bar{\boldsymbol{z}}) + \beta_{t_m}\, \sigma(\bar{\boldsymbol{z}}) \right) \tag{9}$$

where β_{t_m} is a time-dependent exploration parameter that balances exploitation of high-utility policies with exploration of uncertain regions of the meta-policy space. This acquisition strategy allows us to exploit the expressive power of f while efficiently navigating the meta-policy space, eliminating the need to fit an explicit Gaussian Process. This approach scales effectively with dimensionality and adapts to contextual heterogeneity encoded in the population. As summarized in Algorithm (1), this meta-level optimization yields an optimal subgroup-level resource allocation policy $\bar{\boldsymbol{z}}^*$, which is then passed to the base-level assignment phase for individual-level decision making.

Base-level Framework. Given the subgroup-level meta-allocation policy $\bar{\boldsymbol{z}}^* = \{\bar{z}_r^k\}_{k\in K, r\in R}$ produced by the meta-level optimization, the base-level framework performs an individual-level contextual bandit search within each (k, r) cell to identify the most promising recipients. This step refines the coarse-grained allocation $\bar{z}_r^k$ by using a local UCB rule over individual contexts to balance exploitation of high-expected responders with exploration under uncertainty. Let $\mathcal{I}_k$ denote the set of eligible individuals in subgroup k, and let f be the shared predictive model mapping individual context $\mathbf{x}^i$ to expected outcome $\hat{y}_{i,r} = f(\mathbf{x}^i)$. For each (k, r) pair with $\bar{z}_r^k > 0$, we define the target number of allocations as $n_{k,r} = \left\lfloor \bar{z}_r^k \cdot |\mathcal{I}_k| \right\rfloor$. To allocate resource r to the top $n_{k,r}$ individuals in $\mathcal{I}_k$, we compute UCB scores $G_{i,r} = \hat{y}_{i,r} + \beta u_{i,r}$, where $u_{i,r}$ denotes uncertainty (e.g., prediction variance), and β balances exploration and exploitation. The top $n_{k,r}$ individuals by $G_{i,r}$ receive resource r. This yields an individualized policy that respects meta-level group constraints while exploiting within-group variation (Algorithm 2).

Algorithm 1: MetaCUB: Phase 1: Meta-level

Input: Subgroups K, Resources R, Computational Budget T_m, Initial Policies $\{\bar{\boldsymbol{z}}^{(j)}\}_{j=1}^{n_0}$, Outcome model f
Output: $\bar{\boldsymbol{z}}^* = \underset{(\bar{\boldsymbol{z}},\hat{y})\in\mathcal{D}_{T_m}}{\arg\max}\ \hat{y}$

Initialize: Dataset $\mathcal{D}_0 \leftarrow \emptyset$
for $j = 1$ to n_0 **do**
  Simulate individual-level assignments using $\bar{\boldsymbol{z}}^{(j)}$
  $\hat{y}^{(j)} \leftarrow$ Evaluate$(f, \bar{\boldsymbol{z}}^{(j)})$ {Estimate outcome via f}
  $\mathcal{D}_0 \leftarrow \mathcal{D}_0 \cup \{(\bar{\boldsymbol{z}}^{(j)}, \hat{y}^{(j)})\}$
end for
for $t_m = n_0 + 1$ to T_m **do**
  Sample candidate set $\mathcal{S}_{t_m} \subset \widetilde{\Delta}^{|K|\cdot|R|}$
  for each $\bar{\boldsymbol{z}} \in \mathcal{S}_{t_m}$ **do**
    Perform B simulations under $\bar{\boldsymbol{z}}$ using f
    Estimate $\mu(\bar{\boldsymbol{z}})$ and $\sigma(\bar{\boldsymbol{z}})$
    Compute UCB score: $a(\bar{\boldsymbol{z}}) = \mu(\bar{\boldsymbol{z}}) + \beta_{t_m} \cdot \sigma(\bar{\boldsymbol{z}})$
  end for
  Select best candidate: $\bar{\boldsymbol{z}}(t_m) \leftarrow \underset{\bar{\boldsymbol{z}}\in\mathcal{S}_{t_m}}{\arg\max}\, a(\bar{\boldsymbol{z}})$
  Simulate assignments under $\bar{\boldsymbol{z}}(t_m)$
  $\hat{y}(t_m) \leftarrow$ Evaluate$(f, \bar{\boldsymbol{z}}(t_m))$
  $\mathcal{D}_{t_m} \leftarrow \mathcal{D}_{t_m-1} \cup \{(\bar{\boldsymbol{z}}(t_m), \hat{y}(t_m))\}$
end for

Algorithm 2: MetaCUB: Phase 2: Base-level

Inputs: Meta-policy $\bar{\boldsymbol{z}}^* = \{\bar{z}_r^k\}_{k\in K, r\in R}$; Outcome model f; Individual sets $\{\mathcal{I}_k\}_{k\in K}$
Output: Individual-level allocation $\mathcal{A}$

Initialize allocation set $\mathcal{A} \leftarrow \emptyset$
for each subgroup $k \in K$ **do**
  for each resource $r \in R$ **do**
    if $\bar{z}_r^k > 0$ **then**
      Set allocation count: $n_{k,r} \leftarrow \left\lfloor \bar{z}_r^k \cdot |\mathcal{I}_k| \right\rfloor$
      for each individual $i \in \mathcal{I}_k$ **do**
        Compute predicted reward: $\hat{y}_{i,r} \leftarrow f(\mathbf{x}^i)$
        Compute UCB score: $G_{i,r} \leftarrow \hat{y}_{i,r} + \beta \cdot u_{i,r}$
      end for
      Select top $n_{k,r}$ individuals by $G_{i,r}$: $\mathcal{S}_{k,r}$
      $\mathcal{A} \leftarrow \mathcal{A} \cup \{(i, r) \mid i \in \mathcal{S}_{k,r}\}$
    end if
  end for
end for
return $\mathcal{A}$

Fairness Properties of MetaCUB

MetaCUB's bi-level design promotes equitable outcomes by decoupling global resource allocation (meta-level) from individual-level targeting (base-level). This structure mitigates group-level allocation disparities often amplified in flat contextual.

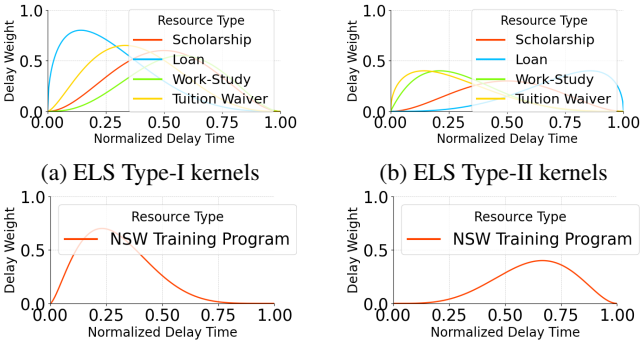
Lemma 1 (Disparity Reduction) *Let $\mathcal{A}_{\texttt{MetaCUB}}$ denote the bi-level allocation under MetaCUB, and $\mathcal{A}_{\texttt{Flat}}$ denote the allocation from a one-level contextual bandit (e.g., LinUCB) that selects individual-resource pairs without subgroup constraints. Define disparity as the difference between the maximum and minimum average outcome across subgroups, Disparity$(\mathcal{A}) = \max_{k\in K} \bar{y}_k(\mathcal{A}) - \min_{k\in K} \bar{y}_k(\mathcal{A})$, where $\bar{y}_k(\mathcal{A})$ is the mean outcome for subgroup k under allocation $\mathcal{A}$. Then, under mild assumptions on exploration and model accuracy, Disparity$(\mathcal{A}_{\texttt{MetaCUB}}) \leq$ Disparity$(\mathcal{A}_{\texttt{Flat}}) - \delta(T_m, f)$, for some $\delta(T_m, f) > 0$ that increases with meta-rounds T_m and predictor fidelity f.*

Sketch. Flat contextual bandits optimize reward across individuals but can disproportionately favor dominant subgroups with higher estimated outcomes, leading to allocation imbalance. In contrast, MetaCUB first distributes re-

sources across subgroups via meta-level optimization, ensuring broader coverage. Then, within each group, the base-level bandit targets high-benefit individuals. This structure bounds the inter-group disparity by ensuring minimum subgroup coverage and reducing waste through outcome-aware targeting. The fairness gap δ arises from this structure.
The full formal proof is provided in Appendix. □

Experiments

We evaluate the proposed bi-level delayed-feedback framework on two real-world datasets: the Educational Longitudinal Study (ELS) (for Education Statistics 2025), where resources represent financial aid packaging, and the JOBS randomized field experiment (Dehejia and Wahba 2025), where the resource is job training. Detailed dataset specifications appear in Appendix. Our simulations cover varied experimental conditions, including delayed vs. immediate feedback, linear vs. nonlinear outcome mappings, different delay kernels, task types (regression for ELS, classification for JOBS), and resource dimensionality (multi-type in ELS, single-type in JOBS).



(a) ELS Type-I kernels (b) ELS Type-II kernels

(c) JOBS Type-I kernel (d) JOBS Type-II kernel

Figure 1: Delay kernel distributions.

To systematically benchmark our method *MetaCUB*, we compare its performance against a suite of baseline algorithms encompassing classical bandits, linear contextual approaches, combinatorial models, and adversarial formulations. *UCB* (Auer, Cesa-Bianchi, and Fischer 2002) treats each resource–recipient pair as an independent arm, ignoring context and group structure; *LinUCB* (Li et al. 2010) incorporates individual contexts via linear regression but omits subgroup budgets; *CUCB* (Chen et al. 2016) selects multiple arms per round yet suffers from limited feedback and scalability; *EXP3* (Auer et al. 2002) is robust to adversarial or delayed rewards but disregards stochastic structure and real-world constraints; *mEXP3* (Tang, Ho, and Liu 2021) explores over full allocation policies but incurs high variance in evaluation; and *DUCB* (Garivier and Moulines 2011) and *SWUCB* (Garivier and Moulines 2011) adapt to non-stationarity via decay or sliding windows yet lack subgroup-aware allocation mechanisms. A more detailed description of these baselines is provided in Table 2 in Appendix.

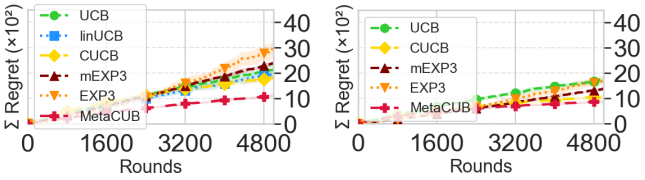
In both datasets, the number of base arms K corresponds to racial subgroups. For ELS (GPA regression), we use ridge regression and a neural network for linear and nonlinear mappings, respectively. For JOBS (binary employment classification), we apply logistic regression and the same neural network to ensure consistent subgroup performance (e.g., 86% overall accuracy on ELS: Asian=87%, Black=85%, Hispanic=86%, White=86%). To simulate dynamic populations, individuals are grouped into fixed-length cohorts: 8 semesters for ELS and 12 months for JOBS. One cohort is active per round and replaced upon completion. Shaded bands in plots denote active cohort periods. Feedback delay is modeled via two kernel types per dataset (Figure 1): four resource-specific kernels in ELS and two variants for the single JOBS resource. These test our method's robustness to heterogeneous, delayed rewards. To reflect real-world constraints, we impose stochastic cooldowns: after receiving a resource, individuals enter a cooldown sampled uniformly from $1, 2, 3$ rounds—adaptable to other settings.

All experiments are conducted with 20 independent random seeds to ensure robustness. Simulations are implemented in Python 3.11.5 using NumPy, scikit-learn, and BoTorch, and executed on an Apple M4 Pro (14-core CPU, 20-core GPU, 16-core Neural Engine, 24 GB RAM) running macOS 15.5. Reported performance metrics are averaged across all runs.

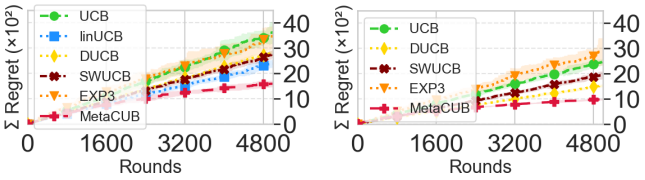
Results

The plots in Figure 2, 3, 4, and 5 illustrate the cumulative regret trajectories of all baseline algorithms compared to our proposed method *MetaCUB* across four experimental settings derived from both the ELS dataset and JOBS datasets with two distinct delay kernel configurations; lines show mean regret over 20 runs, and shaded regions indicate standard deviations. These settings vary along two dimensions: the nature of the outcome function (linear vs. nonlinear) and the presence or absence of delayed feedback (delayed vs. immediate). In all scenarios, *MetaCUB* consistently achieves the lowest cumulative regret, demonstrating its superior ability to adaptively balance exploration and exploitation under both immediate and delayed reward settings. The performance gap is especially pronounced in the delayed-feedback environments, where conventional bandits like *UCB* and *EXP3* exhibit substantially higher regret due to their lack of temporal sensitivity. Algorithms such as *DUCB* and *SWUCB* show improved robustness under delay but still underperform relative to *MetaCUB*, which leverages subgroup-level structure and kernelized delay modeling.

Moreover, the performance differences observed between both datasets experiments primarily arise from the distinct resource feedback delay kernels used in each scenario. Type-I kernels (Figure 1a, and Figure 1c) exhibit peaked, unimodal shapes with concentrated delay weights at early-to-mid normalized times, meaning that the majority of reward signals arrive relatively soon after allocation. This structure allows learning algorithms to receive informative feedback more rapidly, enabling faster adaptation and significantly lower cumulative regret—most notably for *MetaCUB*, which leverages structured delay-awareness. In contrast, Type-II kernels (Figure 1b, and Figure 1d) are more flattened and dispersed, with broader support across the time

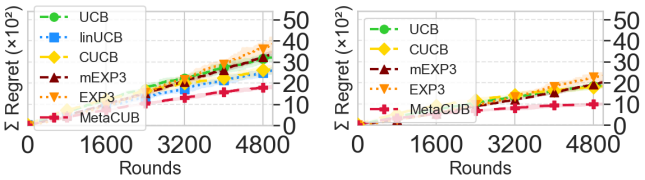


(a) Immediate * linear (b) Immediate * non-linear

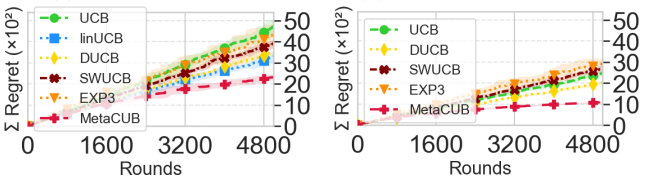


(c) Delayed * linear (d) Delayed * non-linear

Figure 2: ELS, Delay Kernel Type-I: Cumulative regret.



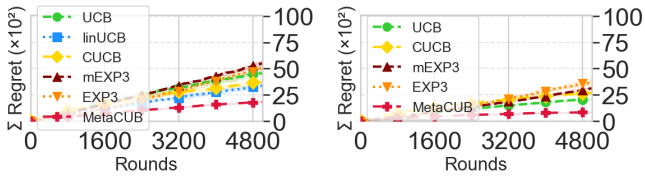
(a) Immediate * linear (b) Immediate * non-linear



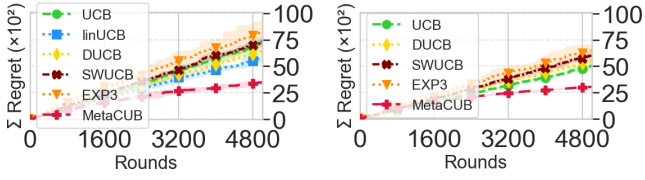
(c) Delayed * linear (d) Delayed * non-linear

Figure 3: ELS, Delay Kernel Type-II: Cumulative regret.

axis. This results in temporally diluted reward signals and increased uncertainty, impeding the learning efficiency of delay-agnostic methods. Under these broader delays, baseline algorithms such as *UCB*, *EXP3*, and their variants accumulate regret more rapidly, particularly in the delayed linear case, while *MetaCUB* remains consistently more resilient. These findings emphasize the practical importance of modeling heterogeneous and temporally diffuse feedback mechanisms when deploying learning-based allocation policies. In both datasets, algorithms evaluated under nonlinear reward functions incur substantially lower regret than their linear counterparts, indicating that relaxing linearity improves learning and allocation quality, even for classical methods such as *UCB* and *CUCB*. Building on this, we evaluate the fairness of allocation decisions by analyzing representational balance across subgroups. While Lemma 1 supports fairness convergence, we complement it with empirical fairness ratios, defined as the proportion of selected individuals in each subgroup. As shown in Tables 1 and Appendix Tables 4–5, *MetaCUB* consistently achieves the most balanced subgroup coverage across datasets and feedback regimes, supporting its fairness-aware design.

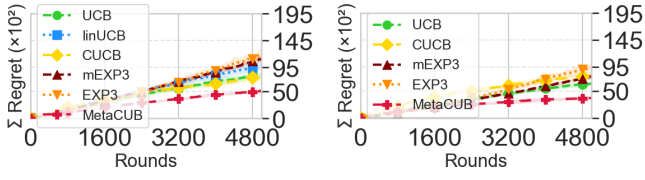


(a) Immediate * linear (b) Immediate * non-linear

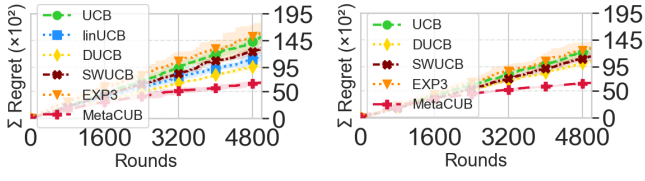


(c) Delayed * linear (d) Delayed * non-linear

Figure 4: JOBS, Delay Kernel Type-I: Cumulative regret.



(a) Immediate * linear (b) Immediate * non-linear



(c) Delayed * linear (d) Delayed * non-linear

Figure 5: JOBS, Delay Kernel Type-II: Cumulative regret.

Algs.	Asian		White		Black		Hispanic	
	Imm	Del	Imm	Del	Imm	Del	Imm	Del
UCB	0.62	0.41	1.29	1.42	0.48	0.33	0.57	0.51
CUCB	0.59	*	0.82	*	0.51	*	0.63	*
EXP3	0.32	0.27	0.91	1.16	0.34	0.22	0.41	0.36
mEXP3	0.45	*	1.27	*	0.36	*	0.28	*
DUCB	*	0.22	*	1.05	*	0.57	*	0.52
SWUCB	*	0.52	*	1.27	*	0.39	*	0.29
MetaCUB	**0.84**	**1.02**	**1.03**	**0.96**	**1.02**	**1.00**	**0.98**	**0.97**

Table 1: ELS Allocation Fairness (full table in Appendix)

Conclusion

We propose *MetaCUB*, a bi-level contextual bandit framework for fair and adaptive resource allocation under delayed feedback. *MetaCUB* outperforms baselines in cumulative regret and achieves more balanced subgroup coverage, supported by both empirical results and theoretical guarantees.

References

Agrawal, S.; and Goyal, N. 2012. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, 39–1. JMLR Workshop and Conference Proceedings.

Agrawal, S.; and Goyal, N. 2013. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, 127–135. PMLR.

Aktaş, E.; Ülengin, F.; and Şahin, Ş. Ö. 2007. A decision support system to improve the efficiency of resource allocation in healthcare management. *Socio-Economic Planning Sciences*, 41(2): 130–146.

Almalki, A. D. A.; and Mohammed, A. I. 2022. The Effect of Immediate and Delayed Feedback in Virtual Classes on Mathematics Students' Higher Order Thinking Skills. *Journal of Positive School Psychology*, 6(6).

Auer, P.; Cesa-Bianchi, N.; and Fischer, P. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2): 235–256.

Auer, P.; Cesa-Bianchi, N.; Freund, Y.; and Schapire, R. E. 2002. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1): 48–77.

Badanidiyuru, A.; Kleinberg, R.; and Slivkins, A. 2018. Bandits with knapsacks. *Journal of the ACM (JACM)*, 65(3): 1–55.

Barnett, W. S. 1995. Long-term effects of early childhood programs on cognitive and school outcomes. *The future of children*, 25–50.

Basu, S.; Papadigenopoulos, O.; Caramanis, C.; and Shakkottai, S. 2021. Contextual blocking bandits. In *International Conference on Artificial Intelligence and Statistics*, 271–279. PMLR.

Biswas, A.; Aggarwal, G.; Varakantham, P.; and Tambe, M. 2021. Learn to intervene: An adaptive learning policy for restless bandits in application to preventive healthcare. *arXiv preprint arXiv:2105.07965*.

Bouneffouf, D.; Rish, I.; and Aggarwal, C. 2020. Survey on applications of multi-armed and contextual bandits. In *2020 IEEE congress on evolutionary computation (CEC)*, 1–8. IEEE.

Burnetas, A. N.; Kanavetas, O.; and Katehakis, M. N. 2025. Optimal data driven resource allocation under multi-armed bandit observations. *Annals of Operations Research*, 1–28.

Chen, G.; Liew, S. C.; and Shao, Y. 2022. Uncertainty-of-information scheduling: A restless multiarmed bandit framework. *IEEE Transactions on Information Theory*, 68(9): 6151–6173.

Chen, L.; Xu, J.; and Lu, Z. 2018. Contextual combinatorial multi-armed bandits with volatile arms and submodular reward. *Advances in Neural Information Processing Systems*, 31.

Chen, W.; Wang, Y.; Yuan, Y.; and Wang, Q. 2016. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *Journal of Machine Learning Research*, 17(50): 1–33.

Chen, Y.; Cuellar, A.; Luo, H.; Modi, J.; Nemlekar, H.; and Nikolaidis, S. 2020. Fair contextual multi-armed bandits: Theory and experiments. In *Conference on Uncertainty in Artificial Intelligence*, 181–190. PMLR.

Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2): 153–163.

Claure, H.; Chen, Y.; Modi, J.; Jung, M.; and Nikolaidis, S. 2020. Multi-armed bandits with fairness constraints for distributing resources to human teammates. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 299–308.

Cui, J.; Liu, Y.; and Nallanathan, A. 2019. Multi-agent reinforcement learning-based resource allocation for UAV networks. *IEEE Transactions on Wireless Communications*, 19(2): 729–743.

Daniels, N.; del Pilar Guzmán Urrea, M.; Rentmeester, C. A.; Kotchian, S. A.; Fontaine, S.; Hernández-Aguado, I.; Lumbreras, B.; Blacksher, E.; Goold, S. D.; Gómez, M. I.; et al. 2016. Resource allocation and priority setting. *Public health ethics: Cases spanning the globe*, 61–94.

Dehejia, R.; and Wahba, S. 2025. NSW/PSID Job Training Data (Jobs dataset). Accessed: July 15, 2025.

Edmondson, D. L.; Kern, F.; and Rogge, K. S. 2019. The co-evolution of policy mixes and socio-technical systems: Towards a conceptual framework of policy mix feedback in sustainability transitions. *Research Policy*, 48(10): 103555.

Erez, L.; Levy, O.; and Mansour, Y. 2024. Regret Guarantees for Adversarial Contextual Bandits with Delayed Feedback. In *Seventeenth European Workshop on Reinforcement Learning*.

for Education Statistics, N. C. 2025. Education Longitudinal Study of 2002 (ELS:2002). Accessed: July 15, 2025.

Gael, M. A.; Vernade, C.; Carpentier, A.; and Valko, M. 2020. Stochastic bandits with arm-dependent delays. In *International Conference on Machine Learning*, 3348–3356. PMLR.

Garivier, A.; and Moulines, E. 2011. On upper-confidence bound policies for switching bandit problems. In *International conference on algorithmic learning theory*, 174–188. Springer.

Gibney, M.; and Jennings, N. R. 1998. Dynamic resource allocation by market-based routing in telecommunications networks. In *International Workshop on Intelligent Agents for Telecommunication Applications*, 102–117. Springer.

Gong, Y.-J.; Zhang, J.; Chung, H. S.-H.; Chen, W.-N.; Zhan, Z.-H.; Li, Y.; and Shi, Y.-H. 2012. An efficient resource allocation scheme using particle swarm optimization. *IEEE Transactions on Evolutionary Computation*, 16(6): 801–816.

Grover, A.; Markov, T.; Attia, P.; Jin, N.; Perkins, N.; Cheong, B.; Chen, M.; Yang, Z.; Harris, S.; Chueh, W.; et al. 2018. Best arm identification in multi-armed bandits with delayed feedback. In *International conference on artificial intelligence and statistics*, 833–842. PMLR.

Gyorgy, A.; and Joulani, P. 2021. Adapting to delays and data in adversarial multi-armed bandits. In *International Conference on Machine Learning*, 3988–3997. PMLR.

Hanna, T. P.; King, W. D.; Thibodeau, S.; Jalink, M.; Paulin, G. A.; Harvey-Jones, E.; O'Sullivan, D. E.; Booth, C. M.; Sullivan, R.; and Aggarwal, A. 2020. Mortality due to cancer treatment delay: systematic review and meta-analysis. *bmj*, 371.

Hegazy, T. 1999. Optimization of resource allocation and leveling using genetic algorithms. *Journal of construction engineering and management*, 125(3): 167–175.

Hui, J. Y. 2002. Resource allocation for broadband networks. *IEEE Journal on selected areas in communications*, 6(9): 1598–1608.

Huo, X.; and Fu, F. 2017. Risk-aware multi-armed bandit problem with application to portfolio selection. *Royal Society open science*, 4(11): 171377.

Joulani, P.; Gyorgy, A.; and Szepesvári, C. 2013. Online learning under delayed feedback. In *International conference on machine learning*, 1453–1461. PMLR.

Kassraie, P.; Rothfuss, J.; and Krause, A. 2022. Meta-learning hypothesis spaces for sequential decision-making. In *International Conference on Machine Learning*, 10802–10824. PMLR.

Kaufmann, E.; Cappé, O.; and Garivier, A. 2012. On Bayesian upper confidence bounds for bandit problems. In *Artificial intelligence and statistics*, 592–600. PMLR.

Kuang, N. L.; Yin, M.; Wang, M.; Wang, Y.-X.; and Ma, Y. 2023. Posterior sampling with delayed feedback for reinforcement learning with linear function approximation. *Advances in neural information processing systems*, 36: 6782–6824.

Kuleshov, V.; and Precup, D. 2014. Algorithms for multi-armed bandit problems. *arXiv preprint arXiv:1402.6028*.

Lancewicki, T.; Segal, S.; Koren, T.; and Mansour, Y. 2021. Stochastic multi-armed bandits with unrestricted delay distributions. In *International Conference on Machine Learning*, 5969–5978. PMLR.

Lane, H.; Sarkies, M.; Martin, J.; and Haines, T. 2017. Equity in healthcare resource allocation decision making: a systematic review. *Social science & medicine*, 175: 11–27.

Légaré, F.; Adekpedjou, R.; Stacey, D.; Turcotte, S.; Kryworuchko, J.; Graham, I. D.; Lyddiatt, A.; Politi, M. C.; Thomson, R.; Elwyn, G.; et al. 2018. Interventions for increasing the use of shared decision making by healthcare professionals. *Cochrane database of systematic reviews*, 21(7).

Li, D.; and Varakantham, P. 2022. Efficient resource allocation with fairness constraints in restless multi-armed bandits. In *Uncertainty in Artificial Intelligence*, 1158–1167. PMLR.

Li, L.; Chu, W.; Langford, J.; and Schapire, R. E. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, 661–670.

Liefner, I. 2003. Funding, resource allocation, and performance in higher education systems. *Higher education*, 46(4): 469–489.

Liu, H.; Liu, K.; and Zhao, Q. 2012. Learning in a changing world: Restless multiarmed bandit with unknown dynamics. *IEEE Transactions on Information Theory*, 59(3): 1902–1916.

Lu, T.; Pál, D.; and Pál, M. 2010. Contextual multi-armed bandits. In *Proceedings of the Thirteenth international conference on Artificial Intelligence and Statistics*, 485–492. JMLR Workshop and Conference Proceedings.

Massy, W. F. 1996. *Resource allocation in higher education*. University of Michigan Press.

Mate, A.; Madaan, L.; Taneja, A.; Madhiwalla, N.; Verma, S.; Singh, G.; Hegde, A.; Varakantham, P.; and Tambe, M. 2022. Field study in deploying restless multi-armed bandits: Assisting non-profits in improving maternal and child health. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 11, 12017–12025.

Monk, D. H. 1981. Toward a multilevel perspective on the allocation of educational resources. *Review of Educational Research*, 51(2): 215–236.

Nguyen, N.-T.; Nguyen, T. T.; Roos, M.; and Rothe, J. 2014. Computational complexity and approximability of social welfare optimization in multiagent resource allocation. *Autonomous agents and multi-agent systems*, 28(2): 256–289.

Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453.

Patil, V.; Ghalme, G.; Nair, V.; and Narahari, Y. 2021. Achieving fairness in the stochastic multi-armed bandit problem. *Journal of Machine Learning Research*, 22(174): 1–31.

Roos, M.; and Rothe, J. 2010. Complexity of social welfare optimization in multiagent resource allocation. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*, 641–648.

Schlisselberg, O.; Cohen, I.; Lancewicki, T.; and Mansour, Y. 2025. Delay as Payoff in MAB. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 19, 20310–20317.

Shi, L.; Wang, J.; and Wu, T. 2023. Statistical inference on multi-armed bandits with delayed feedback. In *International Conference on Machine Learning*, 31328–31352. PMLR.

Steiger, J.; Li, B.; and Lu, N. 2022. Learning from delayed semi-bandit feedback under strong fairness guarantees. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, 1379–1388. IEEE.

Su, R.; Zhang, D.; Venkatesan, R.; Gong, Z.; Li, C.; Ding, F.; Jiang, F.; and Zhu, Z. 2019. Resource allocation for network slicing in 5G telecommunication networks: A survey of principles and models. *IEEE Network*, 33(6): 172–179.

Tang, W.; Ho, C.-J.; and Liu, Y. 2021. Bandit learning with delayed impact of actions. *Advances in Neural Information Processing Systems*, 34: 26804–26817.

Tran-Thanh, L.; Chapman, A.; Rogers, A.; and Jennings, N. 2012. Knapsack based optimal policies for budget–limited multi–armed bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, 1134–1140.

Vernade, C.; Cappé, O.; and Perchet, V. 2017. Stochastic bandit models for delayed conversions. *arXiv preprint arXiv:1706.09186*.

Wang, Y.; Hu, J.; Chen, X.; and Wang, L. 2019. Distributed bandit learning: Near-optimal regret with efficient communication. *arXiv preprint arXiv:1904.06309*.

Wang, Y.; Sharma, M.; Xu, C.; Badam, S.; Sun, Q.; Richardson, L.; Chung, L.; Chi, E. H.; and Chen, M. 2022. Surrogate for long-term user experience in recommender systems. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 4100–4109.

Weiner, B. J.; Lewis, M. A.; Clauser, S. B.; and Stitzenberg, K. B. 2012. In search of synergy: strategies for combining interventions at multiple levels. *Journal of the National Cancer Institute Monographs*, 2012(44): 34–41.

Xu, X.; Dong, F.; Li, Y.; He, S.; and Li, X. 2020. Contextual-bandit based personalized recommendation with time-varying user interests. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 6518–6525.

Xu, X.; Tao, M.; and Shen, C. 2020. Collaborative multi-agent multi-armed bandit learning for small-cell caching. *IEEE Transactions on Wireless Communications*, 19(4): 2570–2585.

Yanovski, S. Z.; and Yanovski, J. A. 2014. Long-term drug treatment for obesity: a systematic and clinical review. *Jama*, 311(1): 74–86.

Yin, T.; Raab, R.; Liu, M.; and Liu, Y. 2023. Long-term fairness with unknown dynamics. *Advances in Neural Information Processing Systems*, 36: 55110–55139.

Zhou, L. 2015. A survey on contextual multi-armed bandits. *arXiv preprint arXiv:1508.03326*.

Zou, L.; Xia, L.; Ding, Z.; Song, J.; Liu, W.; and Yin, D. 2019. Reinforcement learning to optimize long-term user engagement in recommender systems. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2810–2818.

Zuo, J.; and Joe-Wong, C. 2021. Combinatorial multi-armed bandits for resource allocation. In *2021 55th Annual Conference on Information Sciences and Systems (CISS)*, 1–4. IEEE.