Algorithmic Collective Action with Multiple Collectives

Claudio Battiloro[*] Harvard University
Pietro Greiner[*] Mila, LawZero
Bret Nestor Harvard University
Oumaima Amezgar University of Padova
Francesca Dominici Harvard University

Abstract

As learning systems increasingly influence everyday decisions, user-side steering via Algorithmic Collective Action (ACA)—coordinated changes to shared data—offers a complement to regulator-side policy and firm-side model design. Although real-world actions have been traditionally decentralized and fragmented into multiple collectives despite sharing overarching objectives-with each collective differing in size, strategy, and actionable goals, most of the ACA literature focused on single collective settings. In this work, we present the first theoretical framework for ACA with multiple collectives acting on the same system. In particular, we focus on collective action in classification, studying how multiple collectives can plant signals, i.e., bias a classifier to learn an association between an altered version of the features and a chosen, possibly overlapping, set of target classes. We provide quantitative results about the role and the interplay of collectives' sizes and their alignment of goals. Our framework, by also complementing previous empirical results, opens a path for a holistic treatment of ACA with multiple collectives.

1 Introduction

AI's rapid spread rides on massive training datasets, yielding stronger predictors, and wider applications. However, as data-driven systems shape and guide more aspects of life they raise acute risks: privacy violations, leakage of sensitive information, and biased decisions that entrench inequality.

Responses span firms, regulators, and users. At the firm's level, "Trustworthy AI" programs embed fairness checks, bias mitigation, privacy audits, and red teaming across the lifecycle [1], but often clash with performance and engagement goals. At the regulators' level, privacy laws—e.g., GDPR [7], PIPEDA [9], and CPRA [19]—set floors, yet compliance alone rarely ensures socially responsible outcomes [16, 20]. Finally, at the user's level, Algorithmic Collective Action (ACA) has recently emerged as an appealing paradigm to empower grassroots efforts by organizing users to coordinate data contributions or refusals to steer models [11, 6].

This work is motivated by the fact that the ACA literature has almost entirely focused on scenarios where only a single collective acts. However, real-world actions have traditionally been carried out by multiple collectives, resulting in decentralized and fragmented practices. Different collectives often vary widely in size, strategy, and even actionable goals, yet they can still align on common overarching objectives (e.g. climate justice or gender equality). Bruno Latour's actor-network theory [13] conceptualized why such movements rarely behave as a single unit: what is regarded as a global campaign is really a web of local interactions with no singular center. These loose networks manage

[*]Corresponding authors. Emails: cbattiloro@hsph.harvard.edu,pietro.greiner@mila.quebec

39th Conference on Neural Information Processing Systems (NeurIPS 2025)
Workshop on Algorithmic Collective Action (ACA@NeurIPS 2025).

to act together only through continual coordination. Indeed, they are inherently dynamic and not intrinsically unified. An interesting example is the cyberfeminist struggle, which has never been one monolithic front; instead, it consists of distributed efforts ranging from feminist hacker-art collectives to global hashtag campaigns. Donna Haraway's Cyborg Manifesto [10] famously proposed the cyborg as a metaphor for coalition across difference, emphasizing affinity over rigid identity categories. In that spirit, cyberfeminism has taken many decentralized forms around the world while still sharing core aims of challenging patriarchal tech culture and gender inequality. Environmental activism provides another strong example: the climate justice movement is a patchwork of groups employing diverse tactics (from youth-led social media strikes to militant direct-action cells) to achieve various actionable goals (from permit denials to pipeline stoppages and enforcement actions), but they all strive toward the same overarching objective of halting climate breakdown. Here, we aim to reflect and model the need for considering multiple collectives in ACA.

Related Works. ACA as users steering ML outputs toward a group goal was recently rigorously formalized in [11], building on the Data Leverage framework's view of data as an active instrument rather than a passive input [21]. In [17], the authors proposed a combinatorial model for ACA to study the strategic interaction between drivers and delivery platforms, inspired by the #DeclineNow DoorDash campaign. The work in [3] studied how the firm's chosen learning algorithms affect the success of ACA. Motivated by the growing regulatory focus on privacy and data protection, [18] studied the effectiveness of ACA if AI firms train differentially private models. The work in [2] focused on ACA in recommender systems. The authors of [8] proposed an alternative theoretical framework to [11] that empowers collectives via statistical inference, enabling them to learn better ACA strategies and infer the parameters that determine their success. The closest work to ours is [12], in which a conceptual framework for ACA with multiple collectives is introduced, and experiments are performed. However, no theoretical treatment is provided, and most of the focus is on ACA with two collectives. Our work is inspired by the empirical findings of [12] and, partly, by its conceptual framing. We move further by designing the first rigorous treatment for ACA with multiple collectives, generalizing the original theoretical framework from [11].

Contribution. We introduce the first theoretical framework for ACA with multiple collectives. We focus on collective action in classification, studying how M collectives can plant signals, i.e., bias a classifier to learn an association between an altered version of the features x and a chosen, possibly overlapping, set of target classes $\{y_c\}_{c=1}^{S \leq M}$. We analyze two regimes: when collectives can act on both features and labels, and when collectives are limited to acting on features only. For each regime, we derive lower bounds on the per-collective and global success, revealing interesting trade-offs driven by the interplay of the sizes of the collectives and by how closely their actionable goals align. We illustrate the framework with a use case in climate adaptation. Finally, we discuss what future methodological and empirical directions are needed to enrich our approach and make it a holistic framework for ACA with multiple collectives.

2 Algorithmic Collective Action with Multiple Collectives

We start from the definition of ACA with a single collective as given in [11, 8]. A firm deploys a learning algorithm on data coming from a certain user population. Within the overall user population, a collective forms that represents a given share of users (a positive fraction, but not the whole). This collective wants to steer the platform's outcome toward a certain objective. To do so, the collective agrees on one strategy for editing their contributions—potentially changing features, labels, or both—before those contributions reach the platform. The platform then receives a training set that blends unmodified data from the baseline process with the collective's edited data, and it deploys its algorithm on this blended data. The collective's impact is evaluated by a tailored success measure. Given its size, the collective's problem is to choose the editing rule that maximizes this success once the platform has trained on the blended data. Here, we rigorously generalize ACA [11] to a setting with multiple collectives, each acting with a possibly different strategy and actionable goal, but attempting to maximize both per-collective and global measures of success.

Data space. We define the data space $Z = X \times Y$ as the product of two finite spaces X and Y, being the feature and label spaces, respectively. All probability measures on Z will be defined on the discrete σ-algebra 2^Z of Z.

Population. We define the the population as a probability space $(\Omega, 2^{\Omega}, \pi)$, with $|\Omega| = N$ finite. Let $f_0 : \Omega \to Z$ be a random variable. We define the population distribution as the push-forward measure $P_0 := (f_0)_{\#}\pi$. Intuitively, Ω is the user population, with each element of Ω being a user, and π is the sampling distribution. Then, f_0 assigns a feature-label pair to each user, and P_0 is the induced distribution over feature–label pairs, i.e., it records, for each $(x, y) \in Z$, the π-fraction of users whose assigned pair is (x, y).

Collectives, Strategies, and Masses. We define an ensemble of M collectives as M disjoint subsets $\Omega_1, \ldots, \Omega_M$ of Ω such that $\pi(\Omega_c) = \alpha_c > 0$ for all c. We refer to α_c as the mass of the c-th collective. If π is the normalizing counting measure and $|\Omega_c| = N_c$, then α_c is exactly the fraction of the population N_c/N belonging to the collective Ω_c. Each collective Ω_c agrees on a potentially randomized strategy $h_c : Z \to Z$ from a space of available strategies, i.e., feasible changes to the data. We define the total mass as $\bar{\alpha} := \sum_{c=1}^{M} \alpha_c$. For every collective Ω_c and every $z \in Z$, we assume $P_0(f(\omega) = z) = P_0(f(\omega) = z \mid \omega \in \Omega_c)$, i.e., before applying the strategy h_c, every collective Ω_c is just a random slice of the same underlying population.

Mixture Distribution. In this setting, the action of a collective Ω_c of applying its strategy h_c steers P_0 towards $P_c := (h_c)_{\#}P_0$ with strength α_c. As such, a firm aiming to train a model $m : X \to Y$ on data coming from Ω will observe the mixture distribution:

$$P(\{\alpha_c\}_c, \{h_c\}_c) \;=\; \left(1 - \bar{\alpha}\right) P_0 \;+\; \sum_{c=1}^{M} \alpha_c P_c. \tag{1}$$

Per-collective and Global Success. Given a strategy h_c for each collective, we denote by $S_c(\alpha_c)$ and $S(\{S_c\}_c)$ the measure of per-collective and global success, respectively. In this work, we study how per-collective success $S_c(\alpha_c)$ is maximized as a function of the mass α_c, and how global success $S(\{S_c\}_c)$ is maximized as a function of the per-collective successes $\{S_c\}_c$.

3 ACA with Multiple Collectives in Classification

Following [11, 8], we start by studying the case where the firm's learning algorithm m is a classifier. We assume that the firm chooses an approximately optimal classifier on the distribution P.

ε-suboptimal Classifier.[11] A classifier $m : X \to Y$ is ε-suboptimal on a set $X' \subseteq X$ under the distribution P if there exists a P' with $\mathrm{TV}\left(P_{Y|X=x}, P'_{Y|X=x}\right) \leq \varepsilon$ such that for all $x \in X'$

$$m(x) = \operatorname*{argmax}_{y \in Y} P'(y \mid x). \tag{2}$$

Since any classifier is at worst 0.5-suboptimal, we assume $\varepsilon < 0.5$ without loss of generality.

Planting Signals. We focus on the setting in which each collective Ω_c wants the classifier m to learn an association between an altered version of the features $g_c(x)$ and a chosen target class y_c^*. Formally, given a transformation $g_c : X \to X$ induced by the strategy h_c, each collective Ω_c measures per-collective success as

$$S_c(\alpha_c) = \Pr_{x \sim P_0}\left[m(g_c(x)) = y_c^*\right]. \tag{3}$$

As explained in the Sec.2 , the ensemble of M collective then adopts and aims to maximize a measure of global success $S(\{S_c\}_c)$. In this work, we propose and study two different choices for $S(\{S_c\}_c)$:

(No One is Left Behind) We first consider an egalitarian metric, which takes the per-collective success of the least successful collective as the global measure os success (for example, worst-group accuracy, group DRO [15]), i.e.,

$$S_{\min}(S_c) := \min_{c \in [M]} \; S_c(\alpha_c). \tag{4}$$

where $[M] = \{1, \ldots, M\}$.

(The Bigger the Better) We then consider a metric proportional to each collective's mass, which averages the per-collective successes of all the collectives, weighting them with their masses, i.e.,

$$S_{\mathrm{w}}(S_c) := \frac{1}{\bar{\alpha}} \sum_{c=1}^{M} \alpha_c \, S_c(\alpha_c). \tag{5}$$

Generalizing the single collective case [11], we investigate simple strategies for planting signals. We then study their success as a function of four key sets of parameters: the collectives' masses $\{\alpha_c\}_c$, the per-collective uniqueness $\{\xi_c\}_c$ of each signal they aim to plant, the global uniqueness ξ of the signals they aim to plant, and their target alignments $\{\beta_c\}_c$. Let us define the marginal distributions P_0^X and P_0^Y of P_0 as $P_0^X := (\gamma_X)_\# P_0$ and $P_0^Y := (\gamma_Y)_\# P_0$, where $\gamma_X : X \times Y \to X$ and $\gamma_Y : X \times Y \to X$ are the projections $\gamma_X(x, y) = x$ and $\gamma_Y(x, y) = y$. If there is no risk of ambiguity, we denote $P_0^X(x)$ and $P_0^Y(y)$ simply with $P_0(x)$ and $P_0(y)$, respectively. Similarly, we will employ the same shorthand notation for any other involved marginal distribution when possible. Furthermore, let us denote with $X^* \subseteq X$ the subset of the feature space representing the image of the elegible set under h_c, i.e., the image of the features alterable by the adoption of h_c.

Per-collective ξ_c-Unique Signals.[11] For each collective Ω_c, we say that a signal of the collective Ω_c is individually ξ_c-unique if it satisfies $P_0\left(X_c^*\right) \leq \xi_c$. Intuitively, ξ_c is a measure of the rarity of the signal Ω_c is trying to plant.

Global ξ-Unique Signals. For each collective Ω_c, the marginal P_c^X models where Ω_c deposits feature mass after its editing action g_c. We say signals are globally ξ-unique if, for all c and all $j \neq c$,

$$P_j^X\left(X_c^*\right) \leq \xi. \tag{6}$$

Intuitively, the signals are globally ξ-unique if it is rare that collectives emit the same feature point under their "actioned" distribution P_c^X.

Target Alignment. For each collective Ω_c, we define the target alignment as

$$\beta_c := \sum_{j \neq c : y_j^* = y_c^*} \alpha_j. \tag{7}$$

Therefore, β_c measures the total mass of all the collectives that have the same target as Ω_c.

Suboptimality Gap.[11] For each collective Ω_c, we define the suboptimality gap of y_c^* as

$$\Delta_c = \max_{x \in X_c^*} \left(\max_{y \in Y} P_0(y \mid x) - P_0\left(y_c^* \mid x\right) \right) \tag{8}$$

Intuitively, Δ_c measures how suboptimal the target label y_c^* is on the signal set under the base distribution P_0.

As usually done [11, 8, 18], we consider (i) the case in which the users can modify both features and labels, referring to the resulting strategies as feature-label strategies, and (ii) the case in which the users can access both features and labels, but modify only the features, referring to the resulting strategies as feature-only strategies.

Feature-label Signal Strategy. For each collective Ω_c, the feature-label signal strategy is given by

$$h_c(x, y) = \left(g_c(x), y_c^*\right). \tag{9}$$

In this case, $X^* = g_c(X)$.

We derive the following lower bound on the per-collective success of each collective as a function of the key parameters described above.

Theorem 1. *For each collective Ω_c, under the feature-label strategy in* (9)*, ε-suboptimality of the classifier m, and (ξ_c, ξ)-uniqueness, it holds:*

$$S_c\left(\alpha_c\right) \geq 1 \; - \; \xi_c \cdot \frac{\Delta_c + 2\varepsilon}{1 - 2\varepsilon} \cdot \frac{1 - \bar{\alpha}}{\alpha_c} \; - \; \xi \cdot \frac{1 + 2\varepsilon}{1 - 2\varepsilon} \cdot \frac{\bar{\alpha} - \alpha_c - \beta_c}{\alpha_c}. \tag{10}$$

Proof. See Appendix A.

Feature-only Signal Strategy. For each collective Ω_c, the feature-only signal strategy is given by

$$h_c(x, y) = \begin{cases} \left(g_c(x), y_c^*\right), & \text{if } y = y_c^* \\ (x, y), & \text{otherwise} \end{cases}. \tag{11}$$

In this case, $X_c^* = g_c(\operatorname{supp} P_0(\cdot \mid y = y_c^*))$.

Again, we derive a lower bound on the per-collective success of each collective as a function of the key parameters. For a collective Ω_c, feature-only strategies can fail when the base distribution strongly favors some label $y \neq y_c^*$, leaving little label uncertainty. To avoid this degenerate case, we make the positivity assumption $P_0(y_c^* \mid x) \geq p_c > 0$ for all $x \in X$ [11].

Theorem 2. *For each collective Ω_c, under the feature-only strategy in* (11)*, ε-suboptimality of the classifier m, (ξ_c, ξ)-uniqueness, and $p_c \leq (1+2\varepsilon)/2$, it holds:*

$$S_c(\alpha_c) \;\geq\; 1 \;-\; \xi_c \,\frac{(1-\bar{\alpha})(1+2\varepsilon-2p_c) \;+\; 2\varepsilon\,\bar{\alpha}}{(1-2\varepsilon)\,p_c} \cdot \frac{1}{\alpha_c} \;-\; \xi\,\frac{1+2\varepsilon}{1-2\varepsilon}\,\frac{\bar{\alpha}-\alpha_c-\beta_c}{\alpha_c\,p_c}. \tag{12}$$

Proof. See Appendix A.

Discussion. The bounds in (10) and (12) are the complement of the sum of two products. The first product can be interpreted as a "rarity × difficulty × mass". In particular, is scales linearly with:

(Rarity) How rare the signal of the collective is under the baseline, ξ_c;

(Difficulty) How hard it is to flip the baseline on that signal, $\frac{\Delta_c+2\varepsilon}{1-2\varepsilon}$ and $\frac{(1-\bar{\alpha})(1+2\varepsilon-2p_c)\,+\,2\varepsilon\,\bar{\alpha}}{(1-2\varepsilon)}$;

(Mass) How much effective mass the collective brings, $\frac{1-\bar{\alpha}}{\alpha_c}$ and $\frac{1}{\alpha_c\,p_c}$.

The second product can be interpreted as a "competition" penalty given by "signal overlap × misalignment". In particular, it scales linearly with:

(Signal Overlap) How often others land on the collective's signal, ξ;

(Misalignment) The fraction of collectives not aligned with the target y_c^*, $\frac{\bar{\alpha}-\alpha_c-\beta_c}{\alpha_c}$ and $\frac{\bar{\alpha}-\alpha_c-\beta_c}{\alpha_c\,p_c}$.

We now derive lower bounds for the global success measures in (4)-(5) under both feature-label and feature-only strategies, as direct consequences of Theorems 1-2, respectively.

Corollary 1. *For each collective Ω_c, under the feature-label strategy in* (9)*, ε-suboptimality of m, and (ξ_c, ξ)-uniqueness, it holds:*

$$S_{\min}(S_c) \geq 1 - \max_{c\in[M]}\left[\xi_c \cdot \frac{\Delta_c+2\varepsilon}{1-2\varepsilon}\cdot\frac{1-\bar{\alpha}}{\alpha_c} \;+\; \xi\cdot\frac{1+2\varepsilon}{1-2\varepsilon}\cdot\frac{\bar{\alpha}-\alpha_c-\beta_c}{\alpha_c}\right], \tag{13}$$

$$S_{\mathrm{w}}(S_c) \geq 1 - \frac{1-\bar{\alpha}}{\bar{\alpha}}\sum_{c=1}^{M}\xi_c\cdot\frac{\Delta_c+2\varepsilon}{\alpha_c(1-2\varepsilon)} - \xi\cdot\frac{1+2\varepsilon}{1-2\varepsilon}\cdot\frac{1}{\bar{\alpha}}\sum_{c=1}^{M}\frac{\bar{\alpha}-\alpha_c-\beta_c}{\alpha_c}. \tag{14}$$

Corollary 2. *For each collective Ω_c, under the feature-only strategy in* (11)*, ε-suboptimality of m, (ξ_c, ξ)-uniqueness, and the positivity assumption, it holds:*

$$S_{\min}(S_c) \;\geq\; 1 - \max_{c\in[M]}\left[\xi_c\,\frac{(1-\bar{\alpha})(1+2\varepsilon-2p_c) \;+\; 2\varepsilon\,\bar{\alpha}}{(1-2\varepsilon)\,p_c}\cdot\frac{1}{\alpha_c} \;+\xi\,\frac{1+2\varepsilon}{1-2\varepsilon}\,\frac{\bar{\alpha}-\alpha_c-\beta_c}{\alpha_c\,p_c}\right], \tag{15}$$

$$S_{\mathrm{w}}(S_c) \;\geq\; 1 - \frac{1}{\bar{\alpha}}\sum_{c=1}^{M}\left[\xi_c\,\frac{(1-\bar{\alpha})(1+2\varepsilon-2p_c) \;+\; 2\varepsilon\,\bar{\alpha}}{(1-2\varepsilon)\,p_c}\cdot\frac{1}{\alpha_c} \;+\xi\,\frac{1+2\varepsilon}{1-2\varepsilon}\,\frac{\bar{\alpha}-\alpha_c-\beta_c}{\alpha_c\,p_c}\right]. \tag{16}$$

On Critical Masses and Alignment. A core quantity of ACA with a single collective is the critical mass, i.e., the smallest mass required to ensure a target level S^* of success for the collective. However, as one could intuitively expect and as our theoretical results confirm, if multiple collectives are present, then not only their masses but also their alignment on target classes are especially important to determine the success of an action. As such, in ACA with multiple collectives, two key questions can be investigated to derive conditions about when a target global success threshold S^* is met: (i) "Given a certain alignment, how big should the masses be to achieve S^*?, and (ii) "Given certain masses, how big should the alignment be to achieve S^*?. These questions naturally suggest leveraging the bounds in (13)-(14)-(15)-(16) to derive lower bounds on collectives' masses (given a certain alignment) and alignment (given certain masses), for a fixed S^*. We will explore this key aspect in an extended version of this work.

4 A Real-World Use Case in Climate Adaptation

In this section, we provide a key real-world use case for ACA with multiple collectives in the space of interventions for climate adaptation. Consider a city that uses a text classifier trained on neighborhood-level records to identify necessary interventions. The neighborhood-level records encompass both official statistics and community-driven forum discussions. In this case, if V is the vocabulary of the text classifier, then $X = 2^V$. The model recommends one intervention per neighborhood from a finite menu Y (e.g., street trees and shade, cool roofs or pavement, expanded cooling centers, or targeted heat-health outreach). Assume to have M collectives, being grassroots neighborhood associations. Each collective is run by neighborhood residents and therefore has a nuanced, ground-level understanding of local conditions—including qualitative, hard-to-measure factors—and seeks to coordinately change its related shared data to steer the classifier toward the intervention $y_c^* \in Y$ it has judged to be most necessary. Each collective thus deploys an editing strategy h_c to achieve its actionable goal. For example, in a feature–label regime, a collective of a waterfront neighborhood documents chronic flooding with household infiltration tests and evidence collected by interviewing community members. The collective then edits its features and labels using that documentation to bias the forum discussions in a coordinated way and explicitly request rain gardens/bioswales. Similarly, collectives would likely need to resort to feature-only strategies if the city's intake pipeline does not allow community label edits, i.e., if the intervention field is locked and labeled by staff/administrative codes. Citywide (a.k.a. global) success would then emphasize equity (lifting the worst-served neighborhood with $S_{\min}$) or scale (a mass-weighted average with S_{w}). Even if this example is presented in a didactic way, there is precedence for community-reported concerns to be ingested into decision-making tools. NOAA– and CAPA-led "Heat Watch" campaigns have produced neighborhood-scale heat maps in $120+$ communities and feed them into planning and public-health practice [14, 4]. European cities are deploying city-scale digital twins: Rotterdam's Open Urban Platform with a Digital Twin underpins climate-resilience analysis and participatory planning, with citywide availability announced from January 2025 [5].

5 Future Directions

The theoretical framework we designed is just the first step to comprehensively characterize ACA with multiple collectives. Several research directions can be pursued.

On the theoretical and methodological side, as we mention in Sec. 2, it is needed to properly characterize how the notion of critical mass from [11] evolves in the multiple collectives setting. Moreover, it would be interesting to generalize the statistical framework of [8] to a multi-collective setting. Another possibility is studying scenarios where collectives seek to erase [11, 8] or unplant [8] signals, as well as mixed-objective settings in which some collectives plant signals while others erase or unplant them. Similarly, a relevant case is also when collectives have heterogeneous capabilities, thus some collectives can use feature–label strategies while others are restricted to feature-only strategies.

On the applied side, it is crucial to design other meaningful use-cases, beyond the one proposed in Sec. 4. Moreover, it is necessary to perform exhaustive experiments to validate the proposed bounds and study the impact of the proposed strategies in real-world scenarios. Finally, it would be interesting to formally reconcile our theoretical framework with the conceptual characterization presented in [12], fostering mutual enrichment.

6 Conclusion

In this work, we introduced the first theoretical framework for Algorithmic Collective Action with multiple collectives. We focused on collective action in classification, studying how multiple collectives can bias a classifier to learn an association between an altered version of their features and a chosen, possibly overlapping, set of target classes. We provided quantitative results showcasing the interesting interplay of collectives' sizes and their alignment of values. Finally, we discussed a potential use case for our framework in the space of interventions for climate adaptation, validating the importance of taking into account the actions of multiple collectives and their interaction.

7 Acknowledgements

C.B acknowledges support from the National Institutes of Health Grants U24ES035309, R01ES34021, and R01ES037156, and the AWS/HDSI via the grant "A Climate and Health AI Foundation Model and Decision-Making Framework for More Effective and Equitable Climate Adaptation".

References

[1] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.

[2] Joachim Baumann and Celestine Mendler-Dünner. Algorithmic collective action in recommender systems: promoting songs by reordering playlists. *Advances in Neural Information Processing Systems*, 37:119123–119149, 2024.

[3] Omri Ben-Dov, Jake Fawkes, Samira Samadi, and Amartya Sanyal. The role of learning algorithms in collective action. In *International Conference on Machine Learning*, pages 3443–3461. PMLR, 2024.

[4] CAPA Strategies, LLC. Heat mapping – heat watch, 2025. URL: `https://www.capastrategies.com/heat-watch`.

[5] City of Rotterdam, CIO Office. Rotterdam in transformation: Vision on the digital city 1.0, 2024. URL: `https://oascities.org/wp-content/uploads/2024/11/Rotterdam-in-tranformation-vision-on-the-digital-city-1.0.pdf`.

[6] Alicia DeVrio, Motahhare Eslami, and Kenneth Holstein. Building, shifting, & employing power: A taxonomy of responses from below to algorithmic harm. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1093–1106, 2024.

[7] European Parliament and Council of the European Union. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). *Official Journal of the European Union*, (L 119):1–88, May 2016. URL: `https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679`.

[8] Etienne Gauthier, Francis Bach, and Michael I. Jordan. Statistical collusion by collectives on learning platforms. In *Forty-second International Conference on Machine Learning*, 2025. URL: `https://openreview.net/forum?id=46yLEXtav4`.

[9] Government of Canada. Personal information protection and electronic documents act, 2000. S.C. 2000, c. 5. URL: `https://laws-lois.justice.gc.ca/eng/acts/P-8.6/`.

[10] Donna J. Haraway. A cyborg manifesto: Science, technology, and socialist-feminism in the late twentieth century. In *Simians, Cyborgs, and Women: The Reinvention of Nature*. Routledge, New York, 1991. First edition.

[11] Moritz Hardt, Eric Mazumdar, Celestine Mendler-Dünner, and Tijana Zrnic. Algorithmic collective action in machine learning. In *International Conference on Machine Learning*, pages 12570–12586. PMLR, 2023.

[12] Aditya Karan, Nicholas Vincent, Karrie Karahalios, and Hari Sundaram. Algorithmic collective action with two collectives. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 1468–1483, 2025.

[13] Bruno Latour. *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford University Press, Oxford, 2005.

[14] National Oceanic and Atmospheric Administration (NOAA). Mapping campaigns. HEAT.gov, 2025. URL: `https://www.heat.gov/pages/mapping-campaigns`.

[15] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. URL: `https://openreview.net/forum?id=ryxGuJrFvS`.

[16] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 59–68, 2019.

[17] Dorothee Sigg, Moritz Hardt, and Celestine Mendler-Dünner. Decline now: A combinatorial model for algorithmic collective action. *arXiv*, abs/2410.12633, 2024.

[18] Rushabh Solanki, Meghana Bhange, Ulrich Aïvodji, and Elliot Creager. Crowding out the noise: Algorithmic collective action under differential privacy. *arXiv preprint arXiv:2505.05707*, 2025.

[19] State of California. California privacy rights act of 2020 (cpra), 2020. Proposition 24, approved November 3, 2020. URL: `https://oag.ca.gov/privacy/ccpa`.

[20] Christine Utz, Martin Degeling, Sascha Fahl, Florian Schaub, and Thorsten Holz. (un) informed consent: Studying gdpr consent notices in the field. In *Proceedings of the 2019 acm sigsac conference on computer and communications security*, pages 973–990, 2019.

[21] Nicholas Vincent, Hanlin Li, Nicole Tilly, Stevie Chancellor, and Brent J. Hecht. Data leverage: A framework for empowering the public in its relationship with technology companies. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2020.

A Proofs

Lemma 1. *If m is ε-suboptimal on a set X, then for any $x \in X$:*

$$P(y^* \mid x) \geq \max_{y \neq y^*} P(y \mid x) + 2\varepsilon \quad \Longrightarrow \quad m(x) = y^*.$$

Proof. By ε-suboptimality on X, for every $x \in X$ there exists a conditional distribution $P'(\cdot \mid x)$ on Y such that

$$m(x) \in \arg\max_{y \in Y} P'(y \mid x) \qquad \text{and} \qquad \frac{1}{2} \sum_{y \in Y} \left| P'(y \mid x) - P(y \mid x) \right| \leq \varepsilon.$$

For every $y \in Y$, it then holds

$$P'(y^* \mid x) \;\geq\; P(y^* \mid x) - \varepsilon \qquad \text{and} \qquad P'(y \mid x) \;\leq\; P(y \mid x) + \varepsilon.$$

Fix $x \in X$ and assume $P(y^* \mid x) \geq \max_{y \neq y^*} P(y \mid x) + 2\varepsilon$. Then, for every $y \neq y^*$,

$$P'(y^* \mid x) \;\geq\; P(y^* \mid x) - \varepsilon \;\geq\; \max_{z \neq y^*} P(z \mid x) + \varepsilon \;\geq\; P(y \mid x) + \varepsilon \;\geq\; P'(y \mid x).$$

Hence y^* is an (in fact, the) maximizer of $P'(\cdot \mid x)$, so $m(x) = y^*$. □

Lemma 2. *For $a, b > 0$, it holds that*

$$1_{\{a<b\}} \;\leq\; \frac{b}{a}.$$

Proof. If $a < b$, then $1_{\{a<b\}} = 1$ and $b/a > 1$, so $1 \leq b/a$. If $a \geq b$, then $1_{\{a<b\}} = 0$ and $0 \leq b/a$ since $a > 0$. In both cases, the inequality holds. □

Theorem 1. *For each collective Ω_c, under the feature-label strategy in* (9)*, ε-suboptimality of the classifier m, and (ξ_c, ξ)-uniqueness, it holds:*

$$S_c(\alpha_c) \geq 1 \;-\; \xi_c \cdot \frac{\Delta_c + 2\varepsilon}{1 - 2\varepsilon} \cdot \frac{1 - \bar{\alpha}}{\alpha_c} \;-\; \xi \cdot \frac{1 + 2\varepsilon}{1 - 2\varepsilon} \cdot \frac{\bar{\alpha} - \alpha_c - \beta_c}{\alpha_c}. \tag{10}$$

Proof. We are going to prove the theorem in two steps: first we will provide a point wise estimate for $x \in X$ and then we will proceed to bound the expectation over X.

First step: Fix a collective Ω_c and $x \in X$. For every $y \in Y$ we define the unnormalized score $N(y \mid x)$:

$$N(y \mid x) := (1 - \bar{\alpha}) P_0(y \mid x) P_0(x) + \sum_{j=1}^{M} \alpha_j P_j(x, y).$$

Under the feature label strategy we have that $P_c(x, y) = 1_{\{Y = y_c^*\}} P_c(x)$. So for any $y \neq y_c^*$ we have that:

$$N(y_c^* \mid x) - N(y \mid x) = (1 - \bar{\alpha})(P_0(y_c^* \mid x) - P_0(y \mid x)) P_0(x) + \underbrace{\sum_{j : y_j^* = y_c^*} \alpha_j P_j(x)}_{S_{al}} - \underbrace{\sum_{j : y_j^* = y} \alpha_j P_j(x)}_{S_y}.$$

In this context we can apply lemma 1 as follows: if for every $y \neq y^*$ it holds that

$$N(y^* \mid x) - N(y \mid x) \geq 2\varepsilon\, P(x) \tag{A_1}$$

where

$$P(x) = (1-\bar{\alpha})P_0(x) + \sum_{j=1}^{M} \alpha_j P_j(x)$$

then $m(x) = y^*$.

Now in order to get A_1 we find a lower bound for the left-hand side using the following two observations:

- for $x \in X_c^*$, by definition of Δ_c

$$P_0(y_c^* \mid x) - P_0(y \mid x) \geq -\Delta_c$$

  hence:

$$(1-\bar{\alpha})(P_0(y_c^* \mid x) - P_0(y \mid x))P_0(x) \geq -(1-\bar{\alpha})\Delta_c P_0(x);$$

- for every $y \neq y_c^*$:

$$S_{\text{ctr}} := \sum_{j:y_j^* \neq y_c^*} \alpha_j P_j(x) \geq \sum_{j:y_j^* = y} \alpha_j P_j(x) = S_y.$$

With these two observations, we can lower-bound the left-hand term of A_1:

$$N(y_c^* \mid x) - N(y \mid x) \geq -(1-\bar{\alpha})\Delta_c\, P_0(x) + S_{\text{al}} - S_{\text{ctr}}$$

which is independent from y. We also use S_{ctr} to rewrite the right hand term as:

$$2\varepsilon\, P(x) = 2\varepsilon\big((1-\bar{\alpha})P_0(x) + S_{\text{al}} + S_{\text{ctr}}\big).$$

So we now have the following sufficient condition:

$$-(1-\bar{\alpha})\Delta_c P_0(x) + S_{\text{al}} - S_{\text{ctr}} \geq 2\varepsilon\big((1-\bar{\alpha})P_0(x) + S_{\text{al}} + S_{\text{ctr}}\big).$$

We can now rearrange to get:

$$(1-2\varepsilon)\, S_{\text{al}} \;-\; (1+2\varepsilon)\, S_{\text{ctr}} \;\geq\; (1-\bar{\alpha})\big(\Delta_c + 2\varepsilon\big)\, P_0(x),$$

and since $S_{\text{al}} \geq \alpha_c P_c(x)$, a stricter sufficient condition is

$$\alpha_c P_c^X(x) \geq (1-\bar{\alpha})\, \underbrace{\frac{\Delta_c + 2\varepsilon}{1-2\varepsilon}}_{:=A_c(\varepsilon,\Delta_c)}\, P_0(x) \;+\; \underbrace{\frac{1+2\varepsilon}{1-2\varepsilon}}_{=:C(\varepsilon)} S_{\text{ctr}}(x) \;=: R(x).$$

This sufficient condition can be expressed in terms of characteristic functions as:

$$1_{\{m(x)=y^*\}} \geq 1_{\{\alpha_c P_c^X(x) \geq R(x)\}}. \tag{A2}$$

Second step: First we apply lemma 2 to the right hand side of A2 we get:

$$1_{\{m(x)\neq y^*\}} \leq 1_{\{\alpha_c P_c^X(x) < R(x)\}} \leq \frac{R(x)}{\alpha_c P_c^X(x)}.$$

now we can take the expectation over $x \sim P_c(x)$:

$$1 - S_c = \mathbb{E}_{P_c(x)}\big[1_{\{m(x)\neq y^*\}}\big] \leq \frac{1}{\alpha_c}\mathbb{E}_{P_c(x)}\left[(1-\bar{\alpha})A_c\frac{P_0(x)}{P_c(x)} + C(\varepsilon)\sum_{j:y_j^*\neq y_c^*}\alpha_j\frac{P_j(x)}{P_c(x)}\right].$$

Now we use the following facts:

- Rarity identity (FL): $\mathbb{E}_{P_c^X}\left[\frac{P_0(x)}{P_c^X(x)}\right] = P_0(X_c^*) \leq \xi_c$.

- Overlap control: $\mathbb{E}_{P_c^X}\left[\frac{P_j^X(x)}{P_c^X(x)}\right] = \sum_{x\in X_c^*} P_j^X(x) \le \xi$ for each $j \neq c$.

And since

$$\sum_{j:y_j^*\neq y_c^*} \alpha_j = \bar{\alpha} - \alpha_c - \beta_c$$

we get:

$$S_c(\alpha_c) \ge 1 \;-\; \xi_c \cdot \frac{\Delta_c + 2\varepsilon}{1-2\varepsilon} \cdot \frac{1-\bar{\alpha}}{\alpha_c} \;-\; \xi \cdot \frac{1+2\varepsilon}{1-2\varepsilon} \cdot \frac{\bar{\alpha} - \alpha_c - \beta_c}{\alpha_c}. \qquad \square$$

Theorem 2. *For each collective Ω_c, under the feature-only strategy in* (11)*, ε-suboptimality of the classifier m, (ξ_c, ξ)-uniqueness, and $p_c \le (1+2\varepsilon)/2$, it holds:*

$$S_c(\alpha_c) \;\ge\; 1 \;-\; \xi_c \frac{(1-\bar{\alpha})(1+2\varepsilon-2p_c) \;+\; 2\varepsilon\,\bar{\alpha}}{(1-2\varepsilon)\,p_c} \cdot \frac{1}{\alpha_c} \;-\; \xi\,\frac{1+2\varepsilon}{1-2\varepsilon}\,\frac{\bar{\alpha}-\alpha_c-\beta_c}{\alpha_c\,p_c}. \tag{12}$$

Proof. We follow similar steps as the proof of 1.

First step: Fix a collective c and $x \in X$. For every $y \in Y$ define

$$N(y \mid x) \;:=\; (1-\bar{\alpha})\,P_0(y \mid x)\,P_0(x) \;+\; \sum_{j=1}^{M} \alpha_j\,P_j(x,y).$$

Under the *feature-only* strategy, each j modifies only the slice y_j^* and leaves all other pairs $(x, y \neq y_j^*)$ unchanged. Writing

$$Q_j^X \;:=\; (g_j)_{\#}P_0(\,\cdot \mid y = y_j^*\,),$$

we have

$$P_j(x,y) \;=\; 1_{\{y=y_j^*\}}\,P_0(y_j^*)\,Q_j^X(x) \;+\; 1_{\{y\neq y_j^*\}}\,P_0(x,y).$$

Hence, for any $y \neq y_c^*$,

$$\begin{aligned} N(y_c^* \mid x) - N(y \mid x) \ge{}& (1-\bar{\alpha})\big(P_0(y_c^* \mid x) - P_0(y \mid x)\big)P_0(x) \\ &+ \underbrace{\sum_{j:\,y_j^*=y_c^*} \alpha_j\,P_0(y_c^*)\,Q_j^X(x)}_{S_{\mathrm{al}}(x)} - \underbrace{\sum_{j:\,y_j^*=y} \alpha_j\,P_0(y)\,Q_j^X(x)}_{S_y(x)}. \end{aligned} \tag{17}$$

Moreover,

$$P(x) = (1-\bar{\alpha})P_0(x) + \sum_{j=1}^{M} \alpha_j P_j^X(x), \tag{18}$$

$$P_j^X(x) = \sum_{y\neq y_j^*} P_0(x,y) + P_0\left(y_j^*\right)Q_j^X(x) = P_0(x) - P_0(x,y_j^*) + P_0(y_j^*)\,Q_j^X(x). \tag{19}$$

By Lemma 1, a sufficient condition for $m(x) = y_c^*$ is

$$N(y_c^* \mid x) - N(y \mid x) \;\ge\; 2\varepsilon\,P(x) \qquad \forall\, y \neq y_c^*. \tag{A3}$$

We bound the two sides of $(A3)$ pointwise, using:

- Uniform positivity: For the maximizer $y \neq y_c^*$ of $P_0(y \mid x)$, $P_0(y_c^* \mid x) \ge p_c$ implies $P_0(y_c^* \mid x) - P_0(y \mid x) \ge 2p_c - 1$.
- Contradictors: $S_y(x) \le S_{\text{ctr}}(x)$ for all $y \neq y_c^*$, where

$$S_{\text{ctr}}(x) \;:=\; \sum_{j:\,y_j^*\neq y_c^*} \alpha_j\,P_j^X(x).$$

- Aligned mass: Keep only the self-aligned contribution: $S_{\mathrm{al}}(x) \;\ge\; \alpha_c\,P_0(y_c^*)\,Q_c^X(x)$.

- Mixture upper bound: Using (19), we can write

$$P(x) = P_0(x) - \sum_j \alpha_j P_0(x, y_j^*) + \sum_j \alpha_j P_0(y_j^*) Q_j^X(x) \le P_0(x) + \sum_j \alpha_j P_0(y_j^*) Q_j^X(x).$$

  Remembering that $P_0(y_j^*) Q_j^X(x) \le P_j^X(x)$ and splitting the last sum into aligned and contradicting indices, we obtain

$$P(x) \le P_0(x) + S_{\text{al}}(x) + S_{\text{ctr}}(x).$$

  Hence $2\varepsilon P(x) \le 2\varepsilon\big(P_0(x) + S_{\text{al}}(x) + S_{\text{ctr}}(x)\big)$.

Tightening the sufficient condition $(A3)$ by setting the bound in (17) $\ge 2\epsilon P(x)$, combining this with the four bullets above, and rearranging, we obtain

$$(1-2\varepsilon)\,\alpha_c P_0(y_c^*) Q_c^X(x) - (1+2\varepsilon) S_{\text{ctr}}(x) \ge \Big((1-\bar\alpha)(1+2\varepsilon-2p_c)+2\varepsilon\,\bar\alpha\Big) P_0(x), \qquad x \in X_c^*. \tag{A4}$$

For $x \in X_c^*$ define the preimage ratios

$$r_c(x) := \frac{P_0(x)}{P_0\big(g_c^{-1}(x)\big)}, \qquad t_c(x) := \frac{S_{\text{ctr}}(x)}{P_0\big(g_c^{-1}(x)\big)}.$$

Using $P_0(y_c^*) Q_c^X(x) = P_0\big(g_c^{-1}(x), y_c^*\big)$ and uniform positivity,

$$\frac{P_0\big(g_c^{-1}(x), y_c^*\big)}{P_0\big(g_c^{-1}(x)\big)} = \frac{\sum_{x' \in g_c^{-1}(x)} P_0(y_c^* \mid x') P_0(x')}{P_0\big(g_c^{-1}(x)\big)} \ge \frac{1}{P_0\big(g_c^{-1}(x)\big)} \sum_{x' \in g_c^{-1}(x)} p_c P_0(x') = p_c.$$

Dividing $(A4)$ by $P_0(g_c^{-1}(x))$ and rearranging then yields the *preimage* sufficient condition

$$\alpha_c p_c \ge (1-\bar\alpha) \underbrace{\frac{1+2\varepsilon-2p_c}{1-2\varepsilon}}_{=:B_{y_c^*}(\varepsilon,p_c)} r_c(x) + \frac{2\varepsilon\,\bar\alpha}{1-2\varepsilon} r_c(x) + \underbrace{\frac{1+2\varepsilon}{1-2\varepsilon}}_{=:C(\varepsilon)} t_c(x), \qquad x \in X_c^*. \tag{A5}$$

Second step: Define the post–action feature law as the pushforward $\lambda_c := (g_c)_\# P_0^X$ of P_0^X through g_c. With this notation, the per–collective success can be written as the expectation under λ_c (law of the unconscious statistician):

$$S_c(\alpha_c) = \Pr_{x \sim P_0^X}\big[m(g_c(x)) = y_c^*\big] = \sum_{x \in X} 1\{m(x) = y_c^*\}\,\lambda_c(x).$$

By Lemma 2 and the preimage margin condition $(A5)$ proved in the first step, we have for all $x \in X$,

$$1\{m(x) \ne y_c^*\} \le 1\{x \in X_c^*\} \frac{(1-\bar\alpha) B_{y_c^*}(\varepsilon, p_c) r_c(x) + \dfrac{2\varepsilon\,\bar\alpha}{1-2\varepsilon} r_c(x) + C(\varepsilon) t_c(x)}{\alpha_c p_c}.$$

Summing over $x \in X$ and using that $\operatorname{supp}(\lambda_c) \subseteq X_c^*$, we obtain

$$\begin{aligned} 1 - S_c(\alpha_c) &= \sum_{x \in X} 1\{m(x) \ne y_c^*\}\,\lambda_c(x) \\ &\le \frac{(1-\bar\alpha) B_{y_c^*}(\varepsilon, p_c)}{\alpha_c p_c} \sum_{x \in X_c^*} r_c(x)\,\lambda_c(x) + \frac{2\varepsilon\,\bar\alpha}{(1-2\varepsilon)\,\alpha_c p_c} \sum_{x \in X_c^*} r_c(x)\,\lambda_c(x) \\ &\quad + \frac{C(\varepsilon)}{\alpha_c p_c} \sum_{x \in X_c^*} t_c(x)\,\lambda_c(x). \end{aligned}$$

Since $r_c(x) = \dfrac{P_0(x)}{\lambda_c(x)}$ and $t_c(x) = \dfrac{S_{\text{ctr}}(x)}{\lambda_c(x)}$, it holds

$$\sum_{x \in X_c^*} r_c(x)\,\lambda_c(x) = \sum_{x \in X_c^*} P_0(x) =: P_0(X_c^*) \le \xi_c,$$

and

$$\sum_{x \in X_c^*} t_c(x)\, \lambda_c(x) \;=\; \sum_{x \in X_c^*} S_{\text{ctr}}(x) \;=\; \sum_{j:\, y_j^* \neq y_c^*} \alpha_j \sum_{x \in X_c^*} P_j^X(x) \;\leq\; \xi \left(\bar{\alpha} - \alpha_c - \beta_c\right),$$

where the last inequality uses global ξ–uniqueness assumption.

Combining the displays yields

$$S_c(\alpha_c) \;\geq\; 1 \;-\; \xi_c \cdot \frac{(1-\bar{\alpha})\, B_{y_c^*}(\varepsilon, p_c) \;+\; \dfrac{2\varepsilon\, \bar{\alpha}}{1-2\varepsilon}}{\alpha_c\, p_c} \;-\; \xi \cdot C(\varepsilon) \cdot \frac{\bar{\alpha} - \alpha_c - \beta_c}{\alpha_c\, p_c}.$$

Recalling $B_{y_c^*}(\varepsilon, p_c) = \dfrac{1 + 2\varepsilon - 2p_c}{1 - 2\varepsilon}$ and $C(\varepsilon) = \dfrac{1 + 2\varepsilon}{1 - 2\varepsilon}$, we obtain

$$S_c(\alpha_c) \;\geq\; 1 \;-\; \xi_c\, \frac{(1-\bar{\alpha})(1 + 2\varepsilon - 2p_c) \;+\; 2\varepsilon\, \bar{\alpha}}{(1-2\varepsilon)\, \alpha_c\, p_c} \;-\; \xi\, \frac{1+2\varepsilon}{1-2\varepsilon}\, \frac{\bar{\alpha} - \alpha_c - \beta_c}{\alpha_c\, p_c}$$

□