The Alignment Game: A Theory of Long-Horizon Alignment Through Recursive Curation

Ali Falahati[1], Mohammad Mohammadi Amiri[2], Kate Larson[1], Lukasz Golab[1]

[1] University of Waterloo
[2] Rensselaer Polytechnic Institute
afalahati@uwaterloo.ca

Abstract

In self-consuming generative models that train on their own outputs, alignment with user preferences becomes a recursive rather than one-time process. We provide the first formal foundation for analyzing the long-term effects of such recursive retraining on alignment. Under a two-stage curation mechanism based on the Bradley–Terry (BT) model, we model alignment as an interaction between two factions: the *Model Owner*, who filters which outputs should be learned by the model, and the *Public User*, who determines which outputs are ultimately shared and retained through interactions with the model. Our analysis reveals three structural convergence regimes depending on the degree of preference alignment: consensus collapse, compromise on shared optima, and asymmetric refinement. We prove a fundamental impossibility theorem: no recursive BT-based curation mechanism can simultaneously preserve diversity, ensure symmetric influence, and eliminate dependence on initialization. Framing the process as dynamic social choice, we show that alignment is not a static goal but an evolving equilibrium, shaped both by power asymmetries and path dependence.

Extended version —
https://github.com/Mortrest/Alignment-Game/

Introduction

Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al. 2022) has become the de facto method for aligning large language models with human preferences. Its appeal lies in a simple loop: broadcast model outputs to annotators, collect pairwise preferences, and update the policy through reward modeling. Among the many instantiations of this loop, the *Bradley-Terry (BT) comparison model* (Bradley and Terry 1952) is widely used, powering early alignment successes such as InstructGPT and many subsequent fine-tuning pipelines. At the same time, BT has been criticized for strong independence assumptions of annotating samples, a tendency to reward extremely probable rather than diverse outputs, and vulnerability to noisy feedback (Ge et al. 2024; Xiao et al. 2024).

Most analyses of BT ask whether a single round of collecting human preferences yields an aligned model. In practice, however, modern generative systems are updated recursively: synthetic outputs are added to the corpus, new models are trained on these added data, and the cycle repeats across model generations. Recent work shows that this self-consuming regime can drift away from human values or collapse onto degenerate equilibria (Ferbach et al. 2024; Shumailov et al. 2023; Gerstgrasser et al. 2024; Alemohammad et al. 2023). A long-horizon view demands a formalism that tracks how preference aggregation compounds across successive rounds.

To fill this gap, we study the long-term alignment dynamics under recursive retraining. We focus on the simplest yet already rich scenario with two factions: the **Model Owner**, representing the incentives of developers or platform providers, and the **Public**, representing the aggregated preferences of society interacting with the model. At each iteration, the Owner curates samples via a mechanism based on BT, the model is trained on the selected outputs, the Public interacts with the released model and preserves a subset of model outputs through various actions such as upvoting, sharing, and other forms of engagement, and these data flow into the next training set. Our central question is: *if this BT curation loop continues indefinitely, what distribution of content survives in the limit*?

We model this loop as a dynamic social choice mechanism, which allows us to reason about fairness, incentive compatibility, and power asymmetries in a multi-agent setting. Our analysis is structured around three core alignment scenarios, each motivated by real-world tensions. First, we explore Perfect Alignment, where the Owner and Public preferences are identical. This represents the idealized ”fully-aligned” goal that many systems implicitly strive for, and analyzing it is critical to understanding the consequences of perfect agreement. Second, we examine Partial Alignment, the most realistic scenario where preferences overlap but do not perfectly coincide: for instance, a platform's commercial goals overlapping with, but not fully capturing, public interests. This case allows us to study the dynamics of compromise. Finally, we analyze Disjoint Alignment, which models a critical conflict where developer and user values are in direct opposition. This scenario reveals the system's inherent power dynamics and who ultimately controls the output when values diverge. By understanding the long-run trajectory of each regime, we can expose the structural trade-offs between diversity, fairness, and stability, establishing a foundation for alternative mechanisms and richer multi-agent systems.

In summary, our contributions are threefold: (i) a formal model of recursive BT-based RLHF involving a model owner and a public user; (ii) a complete characterization of its long-run behavior, including an impossibility result that generalizes classic social choice tensions to the dynamic setting; and (iii) empirical evidence that these theoretical dynamics surface in practice. This work aims to lay the groundwork for a broader theory of long-term alignment, extending beyond single-step fine-tuning to account for the recursive and evolutionary nature of generative models.

Related Work

Generative AI systems face fundamental challenges arising from recursive training dynamics and the cumulative effects of self-generated data. (Shumailov et al. 2023) and (Ferbach et al. 2024) demonstrated that training models on their own generated content leads to irreversible "model collapse," where output diversity diminishes through recursive iterations. These limitations are compounded by recursive training dynamics, as (Dohmatob 2024) demonstrates clear crossover points between stable and collapse regimes (Xu, He, and Cheng 2025). (Gerstgrasser et al. 2024) showed that data accumulation strategies can prevent collapse while replacement accelerates it.

The multi-stakeholder dimension of RLHF has been explored by (Tewolde et al. 2024), who propose Reinforcement Learning from Collective Human Feedback using social choice theory, while (Mishra 2023) proves that universal AI alignment using RLHF is impossible under democratic constraints. The BT model, widely used in preference-based reward modeling, faces significant limitations in multi-stakeholder scenarios. (Sun, Shen, and Ton 2024) provides the first theoretical critique, arguing these models are unnecessary for downstream optimization and proposing classification-based alternatives focused on "order consistency." (Zhang et al. 2024) demonstrate that BT models "fall short in expressiveness, particularly in addressing intransitive preferences" that arise naturally in multi-stakeholder scenarios.

(Wu, Niezink, and Junker 2022) shows that these models violate the independence assumptions when multiple stakeholders are involved, leading to systematic marginalization of the preferences of minorities. (Eckersley 2019) demonstrates that Arrow's impossibility theorem applies to AI alignment, showing that no satisfactory method exists to aggregate multiple human preferences without violating fairness criteria. (Qiu 2024) extends this with Arrow-like impossibility theorems for representative social choice settings. The convergence of theoretical impossibility results with empirical evidence of model collapse establishes the foundation for understanding AI alignment as an inherently conflicted game between competing stakeholder interests. Additional insights from social choice theory highlight its role in guiding AI alignment, particularly in handling diverse human feedback and avoiding preference aggregation pitfalls (Conitzer et al. 2024).

Problem Definition

We study iterative retraining of generative models, where synthetic outputs recursively influence future training data through a two-agent curation process. This analysis captures the alignment challenges that emerge when two stakeholders jointly shape the evolution of AI systems.

Model Components

Consider a generative modeling system with the following components:

- **State Space:** A compact metric space $(\mathcal{X}, d)$ representing the content domain.
- **Agents:** Two curators, the *Owner* (model developer) and the *Public* (user community), with continuous reward functions $r_O, r_P : \mathcal{X} \to \mathbb{R}$ encoding their respective preferences.
- **Data Evolution:** A sequence of public datasets $\{\mathcal{D}_t\}_{t\geq 1}$ where $\mathcal{D}_t \subset \mathcal{X}$.
- **Model Sequence:** Generative models $\{\mathcal{M}_t\}_{t\geq 1}$ with output distributions $p_t \in \mathcal{P}(\mathcal{X})$ where $\mathcal{P}(\mathcal{X})$ denote the set of all Borel probability measures on $\mathcal{X}$.
- **Sampling Pool:** We define a *pool* of size K as a collection of i.i.d. samples $\{x_1, \ldots, x_K\} \sim p$. The integer K is the *pool size* and represents the number of alternatives available for pairwise comparison under BT.

Definition 1. *For a probability measure p on $\mathcal{X}$, a pool size $K \geq 2$, and a reward function r, define the BT weight as (Ferbach et al. 2024):*

$$H^p_{K,r}(x) := \mathbb{E}_{Y_1,\ldots,Y_{K-1}\sim p}\left[\frac{K\, e^{r(x)}}{e^{r(x)} + \sum_{j=1}^{K-1} e^{r(Y_j)}}\right].$$

Alignment Game Framework

The system recursively evolves as follows. Start with initial dataset $\mathcal{D}_1$ and initial model distribution p_0.
Recursive loop (for $t = 1, 2, \ldots$):

(1) Owner Curation: At iteration t, the Owner samples a pool $\{x_1, \ldots, x_K\} \sim p_t$ *(the pool is a subset of $\mathcal{D}_t$, i.e., $\{x_i\} \sim q_t$ with $q_t \approx p_t$)* and selects outputs via BT selection with reward r_O, yielding:

$$\tilde{p}_t(x) = p_t(x) \cdot H^{p_t}_{K,r_O}(x) \tag{1}$$

(2) Model Update: Train $\mathcal{M}_{t+1}$ on data drawn from $\tilde{p}_t$, producing the updated model distribution:

$$p_{t+1}(x) \approx \tilde{p}_t(x) \tag{2}$$

(3) Public Curation: The Public samples a pool $\{\hat{x}_1, \ldots, \hat{x}_M\} \sim p_{t+1}$ and applies BT selection with reward r_P, yielding:

$$\hat{p}_t(x) = p_{t+1}(x) \cdot H^{p_{t+1}}_{M,r_P}(x) \tag{3}$$

(4) Dataset Evolution: Update $\mathcal{D}_{t+1} = \mathcal{D}_t \cup \mathcal{O}^*_t$ where $\mathcal{O}^*_t \sim \hat{p}_t$.

Alignment Regimes

The interaction between curator preferences determines the system's long-term behavior. Let $A_O = \arg\max_{x \in \mathcal{X}} r_O(x)$ and $A_P = \arg\max_{x \in \mathcal{X}} r_P(x)$ denote the optimal sets for each curator.

Definition 2. *We categorize value alignment between the Owner and the Public as:*

- ***Perfect Alignment:*** $A_O = A_P$ *(complete agreement)*
- ***Partial Alignment:*** $A_O \cap A_P \neq \emptyset$ *with* $A_O \neq A_P$ *(overlapping preferences)*
- ***Disjoint Alignment:*** $A_O \cap A_P = \emptyset$ *(conflicting preferences)*

Assumptions

Our analysis relies on the following assumptions:

1. The recursive curation mechanism is explicitly defined by the BT model. The convergence properties and the impossibility theorem are direct consequences of the mathematical properties of this specific pairwise comparison model.
2. We analyze the idealized dynamic $p_{t+1}(x) = \tilde{p}_t(x)$, where the model distribution at the next step perfectly matches the target distribution from the current step's curation. This abstracts away optimization error, noise, or catastrophic forgetting from practical training.
3. The reward functions $r_O : \mathcal{X} \to \mathbb{R}$ and $r_P : \mathcal{X} \to \mathbb{R}$ are assumed to be fixed and continuous over a compact metric space $\mathcal{X}$. This ensures the optimal sets A_O and A_P are well-defined and non-empty.

Throughout the analysis, we consider open neighborhoods around maximizer sets. For any set $A \subset \mathcal{X}$ and radius $\eta > 0$,

$$B_\eta(A) := \{x \in \mathcal{X} : \inf_{y \in A} d(x, y) < \eta\}. \tag{4}$$

where $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ is the metric on $\mathcal{X}$.

Core Recursive Alignment Challenges

We assume as $t \to \infty$, the public dataset becomes dominated by curated synthetic data:

$$\lim_{t \to \infty} \frac{|\mathcal{D}_1|}{|\mathcal{D}_t|} = 0. \tag{5}$$

This *self-consuming* regime, where models train predominantly on their own filtered outputs, is a critical challenge for maintaining alignment with diverse human values while avoiding mode collapse or value lock-in. This setting enables us to investigate fundamental questions about recursive alignment:

- **Convergence:** Under what conditions does the system converge to a stable distribution? Does it collapse to point masses or preserve diversity?
- **Influence Dynamics:** How do the sequential curation mechanisms affect the relative influence of each curator?
- **Alignment Impact:** How does the degree of preference alignment between curators shape the evolution and final state of the generative model?

Theoretical Results

We now analyze the long-term behavior of our two-stage curation mechanism under varying degrees of alignment between curators[1]. We begin with the idealized case where both the Owner and the Public share identical preferences.

Perfect Alignment: The Consensus Trap

Theorem 1. *Let the Owner and Public have perfectly aligned preferences:* $A_O = A_P =: A_\star \neq \varnothing$ *and* $\{p_t\}_{t \geq 0}$ *be the sequence of distributions generated by this mechanism. Then, the system reaches complete consensus: for any* $\eta > 0$*, there exist constants* $C, c > 0$ *such that* $p_t(\mathcal{X} \setminus B_\eta(A_\star)) \leq Ce^{-ct}$ *for all* $t \geq 0$*. Moreover,* p_t *converges weakly to the renormalized initial distribution on* $A_\star$*:*

$$p_\infty(x) = \frac{p_0(x)}{\int_{A_\star} p_0(z)\, dz} \mathbf{1}_{A_\star}(x).$$

Our first result uncovers that greater alignment does not preserve diversity, but speeds up its collapse. When preferences are perfectly aligned, the model concentrates on the shared maximizers, leading to a sharply reduced outcome space. Perfect alignment produces an impoverished limiting distribution. The exponential decay rate e^{-ct} reflects the speed of this convergence, highlighting how quickly diversity is lost under alignment. In the limit, the only remaining variation arises from the initial distribution p_0 restricted to $A_\star$.

Corollary 2 (Mode Collapse for Unique Maximizers)**.** *When both curators agree on a unique optimal point* $A_O = A_P = \{x^\star\}$*, system undergoes mode collapse:* $p_t \to \delta_{x^\star}$*, where* δ_x *represents the Dirac delta function at* x*.*

This corollary represents the most extreme form of homogenization and highlights a fundamental risk in alignment-driven curation: as agreement increases, the support of the distribution contracts. Thus, systems designed to maximize agreement between developers and users may unintentionally collapse into echo chambers. This consensus trap suggests that some degree of preference misalignment may be necessary to sustain a healthy generative ecosystem. Our result complements prior findings such as (Ferbach et al. 2024), which show that recursive training on curated outputs can lead to long-term degeneracy. Thus, before pursuing perfect agreement, we should ask whether we are engineering an echo chamber that eliminates diversity.

Partial Alignment: The Compromise Equilibrium

Next, we consider the case where curators share some common ground while maintaining distinct preferences, a setting that better reflects the relationship between model owners and diverse user communities.

Theorem 3. *Suppose the curators have overlapping but distinct preferences, with shared optima* $A_{shared} := A_O \cap$

[1]Proofs are in Section A of Appendix in the extended version.

$A_P \neq \varnothing$ while $A_O \neq A_P$. Under the two-stage curation mechanism, only the intersection survives: the mass outside $B_\eta(A_{shared})$ decays exponentially as $p_t(\mathcal{X} \setminus B_\eta(A_{shared})) \leq Ce^{-ct}$. The limiting distribution p_∞ concentrates entirely on A_{shared} with density proportional to the initial distribution:

$$p_\infty(x) = \frac{p_0(x)}{\int_{A_{shared}} p_0(z)dz} \mathbf{1}_{A_{shared}}(x).$$

Our findings reveal a key dynamic of partial alignment: the recursive process filters out content valued exclusively by one curator, preserving only what lies in the intersection. This "lowest common denominator" effect eliminates curator-specific signals, including potentially novel or specialized contributions. Although diversity is maintained within A_{shared}, it remains limited to areas of mutual agreement. For AI systems, this suggests that iterative training with multiple stakeholders may progressively reduce the breadth of the model to only the jointly endorsed features.

Disjoint Alignment: Asymmetric Power Dynamics

The most adversarial situation arises when curators hold entirely disjoint preferences. In such cases, the question becomes: who ultimately defines the output space of the generative model when both sides embody incompatible values?

Theorem 4. *When curator preferences are completely disjoint ($A_O \cap A_P = \varnothing$), the Owner determines the support while the Public refines within it. Define the Public's preferred subset within Owner optima as $A_{P|O} := \arg\max_{x \in A_O} r_P(x)$. The system exhibits two-stage exponential suppression. First, content outside the Owner's optima vanishes with $p_t(\mathcal{X} \setminus B_\eta(A_O)) \leq C_1 e^{-c_1 t}$, then within A_O, mass concentrates on the Public's preferred subset with $p_t(A_O \setminus B_\eta(A_{P|O})) \leq C_2 e^{-c_2 t}$. The limiting distribution is*

$$p_\infty(x) = \frac{p_0(x)}{\int_{A_{P|O}} p_0(z)dz} \mathbf{1}_{A_{P|O}}(x).$$

This result exposes the inherent power asymmetry in the recursive alignment game. The Owner's first-mover advantage allows it to determine the feasible region, while the Public can only optimize within these constraints. This "best of the worst" dynamic mirrors real-world scenarios where users must select from options pre-filtered by platform algorithms. The Owner's preferences shape the system quickly, while the Public's influence manifests gradually as a refinement. This temporal divide reveals a deeper governance dilemma: once early alignment choices solidify into norms, user feedback becomes more about adaptation than agency.

The Fundamental Impossibility Result

After exploring the range of alignment scenarios, we turn to a deeper limitation of recursive BT-based curation. Even under ideal conditions, certain properties that appear simultaneously desirable cannot all be achieved at once.

Theorem 5. *For any non-trivial preference misalignment ($A_O \neq A_P$), it is impossible to simultaneously satisfy:*

(1) Full Coverage:

$$\liminf_{t\to\infty} p_t(A_O \setminus A_P) > 0 \quad \text{and} \quad \liminf_{t\to\infty} p_t(A_P \setminus A_O) > 0$$

(2) Symmetric Influence: There exists a permutation-invariant functional Φ such that

$$p_\infty = \Phi(r_O, r_P, p_0) = \Phi(r_P, r_O, p_0)$$

(3) Initial Distribution Independence: For any two initial distributions $p_0, q_0 \in \mathcal{P}(\mathcal{X})$ with $\operatorname{supp}(p_0) = \operatorname{supp}(q_0) = \mathcal{X}$,

$$p_\infty^{(p_0)} = p_\infty^{(q_0)}$$

This impossibility result states that no recursive BT-based curation mechanism can simultaneously maintain content diversity across disagreement regions (Full Coverage), treat both curators equally (Symmetric Influence), and produce outcomes independent of initial conditions (Initial Distribution Independence). Each of these properties captures a fundamental value tension in alignment systems. Full Coverage represents the epistemic goal of preserving the full spectrum of ideas, ensuring that minority or dissenting content is not prematurely filtered out. Symmetric Influence encodes fairness: both curators should exert comparable control over the generative process, preventing dominance by either side. Initial Distribution Independence embodies stability, guaranteeing that the system's long-term behavior does not hinge on arbitrary starting points or early biases.

The theorem's implications extend beyond our specific model: it formalizes the intuition that these goals pull in incompatible directions. The tradeoffs observed in recursive BT-based curation systems are therefore not artifacts of design but mathematical necessities. Designers must decide which virtue to compromise: accepting homogenization, embracing asymmetry, or tolerating historical lock-in.

Remark 1. *Across all alignment regimes, the support of the limiting distribution shrinks as follows:*

- $\operatorname{supp}(p_\infty) \subseteq A_O \cap A_P$ *when the intersection is non-empty*
- $\operatorname{supp}(p_\infty) \subseteq A_{P|O} \subseteq A_O$ *when* $A_O \cap A_P = \emptyset$

This demonstrates that recursive BT-based curation inevitably reduces diversity, concentrating mass on increasingly narrow regions of agreement. These structural constraints raise a deeper question: what kind of decision-making process is recursive curation, and how should we reason about its limitations? To answer this, we now reinterpret the recursive BT-based curation mechanism using social choice theory.

The Curation Mechanism as Social Choice

The recursive BT-based curation process can be viewed as a form of collective decision-making, where two agents jointly shape the long-term distribution of model outputs. To examine its properties, fairness, efficiency, and the preservation of diversity, we draw from social choice theory and mechanism design. We formalize recursive BT-based curation as a dynamic preference aggregation mechanism in which the Owner and the Public iteratively express preferences over a shared alternative space $\mathcal{X}$ through a two-stage influence process analogous to voting.

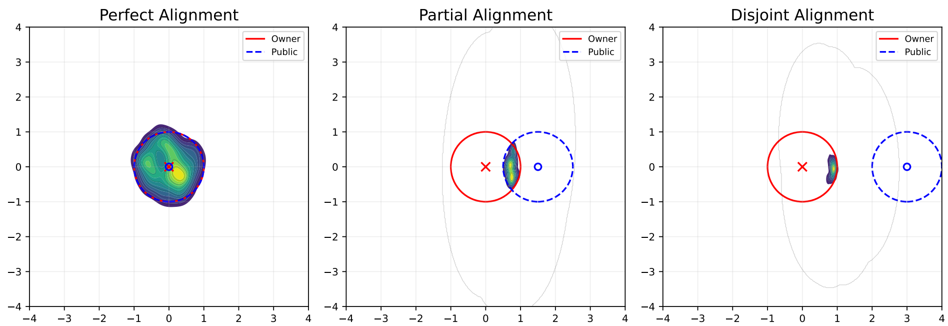

Figure 1: KDE plots showing point distributions in three alignment scenarios: perfect alignment (same circles), partial alignment (overlapping circles), and disjoint alignment (non-overlapping circles). Red circles indicate owner preferred regions, blue dashed circles show public preferred regions.

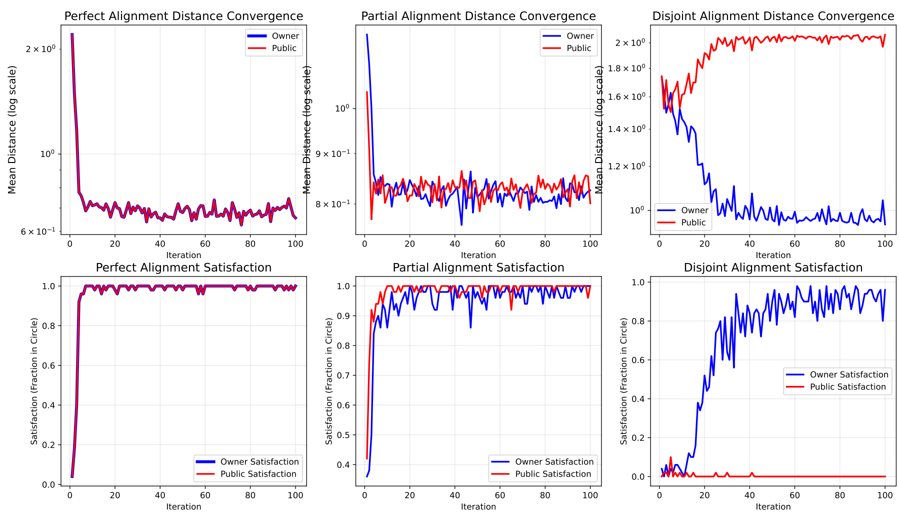

Figure 2: (Top) Convergence of the mean distance to the owner and public centers over iterations for three alignment scenarios: perfect alignment, partial alignment, and disjoint alignment.(Bottom) The fraction of points within the owner and public preferred regions ("satisfaction") as the alignment process progresses.

Definition 3. *A dynamic social choice mechanism for the alignment game consists of:*

(1) *A set of alternatives $\mathcal{X}$ (the content space).*

(2) *Two agents $\mathcal{N} = \{O, P\}$ (Owner and Public).*

(3) *Preferences, captured by reward functions $r_i : \mathcal{X} \to \mathbb{R}$, for $i \in \mathcal{N}$.*

(4) *A social choice correspondence $\mathcal{F} : \mathcal{R}^2 \times \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{X})$ that maps preference profiles and current distributions to updated distributions.*

(5) *A limit social choice function $f_\infty : \mathcal{R}^2 \to \mathcal{P}(\mathcal{X})$ representing the long-run outcome.*

The mechanism implements a sequential decision process in which agents express preferences via pairwise (binary) comparisons over alternatives. The Owner selects an alternative $x \in \mathcal{X}$ first, based on their reward function, and the Public responds by evaluating this choice using their own reward function. The final distribution over alternatives is updated based on these sequential comparisons, reflecting the asymmetric influence of each agent.

Strategic Voting and Incentive Compatibility

In the social choice literature, a central property is strategyproofness: agents should not have an incentive to misrepresent their preferences.

Theorem 6. *For any reported pair (r'_O, r'_P), let $p_\infty(r'_O, r'_P)$ denote the weak limit of $\{p_t\}$. Define each agent's utility by*

$$U_O(r'_O, r'_P) := \mathbb{E}_{x \sim p_\infty(r'_O, r'_P)}[\, r_O(x)\,],$$
$$U_P(r'_O, r'_P) := \mathbb{E}_{x \sim p_\infty(r'_O, r'_P)}[\, r_P(x)\,].$$

Then, truthful reporting is weakly dominant for both agents:

$$U_O(r_O, r'_P) \;\geq\; U_O(\hat{r}_O, r'_P), \quad U_P(r'_O, r_P) \;\geq\; U_P(r'_O, \hat{r}_P)$$

for all $(\hat{r}_O, \hat{r}_P)$, and for all (r'_O, r'_P). In particular, $(r'_O, r'_P) = (r_O, r_P)$ is a Nash equilibrium.

The theorem ensures truthful elicitation: for both agents, truthful reporting weakly dominates any misreport, regardless of the other report. Manipulating reports is therefore ineffective; influence instead comes from the dynamics. The result addresses incentives over reports only, and the broader questions of fairness, symmetric influence, and coverage are taken up in the next theorem, which formalizes their mutual incompatibility.

Fairness and Representation in Dynamic Voting

Beyond strategyproofness, social choice theory provides principled ways of thinking about fairness criteria, such as equal treatment of agents. We formalize this notion through the concept of *influence parity*.

Theorem 7. *No asymmetric sequential preference aggregation mechanism can simultaneously satisfy:*

(EQ) Influence parity: the outcome is order symmetric, that is

$$p_\infty^{\mathrm{OF}}(r_O, r_P) \;=\; p_\infty^{\mathrm{PF}}(r_O, r_P).$$

Here p_∞^{OF} is the limiting distribution when the Owner acts first, and p_∞^{PF} when the Public acts first.

(PO) Pareto optimality: there is no distribution q on X with

$$\mathbb{E}_q[r_O] \geq \mathbb{E}_{p_\infty}[r_O], \qquad \mathbb{E}_q[r_P] \geq \mathbb{E}_{p_\infty}[r_P],$$

(UC) Uniqueness: for each (r_O, r_P) the outcome is unique, independent of initialization and of any tie breaking.

The sequential nature of voting creates a first-mover advantage that conflicts with equal treatment. Unlike static voting where ties can be broken symmetrically, the recursive dynamics amplify initial asymmetries, making true voter equality impossible while maintaining uniqueness.

Experiments

We validate our analysis through two experimental frameworks: a synthetic alignment game that directly implements our mathematical framework, and a text-based alignment game based on a realistic language modeling setting.

Synthetic Alignment Game

Experimental Setup. We implement a synthetic alignment game in $\mathbb{R}^2$ where the Owner and Public preferences are defined by circular reward regions. The reward function for agent $i \in \{O, P\}$ with center c_i and radius r is:

$$r_i(x) = \begin{cases} 1.0 & \text{if } \|x - c_i\| \leq r \\ -2.0 \cdot (\|x - c_i\| - r) & \text{if } \|x - c_i\| > r \end{cases} \tag{6}$$

This creates a sharp preference boundary where points inside the circle receive maximum reward and points outside receive penalties proportional to their distance.

Alignment Scenarios. *(i) Perfect alignment:* $c_O = c_P = (0, 0)$, $r = 1.0$, i.e., both agents prefer the same circular region; *(ii) Partial alignment:* $c_O = (0, 0)$, $c_P = (1.5, 0)$, $r = 1.0$, i.e., overlapping circles with shared optima; *(iii) Disjoint alignment:* $c_O = (0, 0)$, $c_P = (3, 0)$, $r = 1.0$, i.e., nonoverlapping circles with no shared optima.

Experimental Parameters. Each experiment runs for 100 iterations with the following parameters: (i) Initial dataset: 1000 points sampled uniformly from $[-5, 5] \times [-5, 5]$; (ii) Owner curation: Select 100 points using the BT mechanism with temperature $\tau = 0.5$; (iii) Generation: Train a Gaussian Mixture Model (GMM) (Bishop 2006) on curated data, generate 200 new samples; (iv) Public curation: Select 50 points from generated samples using the BT mechanism; (v) Dataset update: Add public-curated samples to the training.

Results and Analysis. Figure (1) (top) shows the final distribution of points after 100 iterations for each alignment scenario. The KDE plots reveal convergence patterns that validate our theoretical predictions. *Perfect Alignment:* The system converges to a concentrated distribution within the shared optimal region, with exponential suppression outside the circle. The final distribution shows high density within the preferred region and negligible mass elsewhere, confirming Theorem 1. *Partial Alignment:* The system concentrates on the intersection of the two circles, preserving diversity only where preferences align. Points outside the intersection are exponentially suppressed, while the shared region maintains substantial density. This validates Theorem 3.

Disjoint Alignment: The system converges to the Public's preferred subset within the Owner's optimal region. The Owner's first-mover advantage determines the support, while the Public refines the distribution within that support. This confirms Theorem 4. Figure 2 (top) tracks the mean distance to each agent's preferred center over iterations. All scenarios show exponential convergence, with the rate depending on the degree of alignment. Perfect alignment achieves the fastest convergence, while disjoint alignment shows a two-stage process, first converging to the Owner's region, then refining within it. Figure 2 (bottom) shows the fraction of points satisfying each agent's preferences ("satisfaction rate"). In perfect alignment, both agents achieve near-universal satisfaction. In partial alignment, satisfaction is limited to the intersection region. In disjoint alignment, the Owner maintains high satisfaction, while the Public's satisfaction is constrained by the Owner's preferences.

Text-Based Alignment Game

Experimental Setup. We implement a realistic text generation scenario using GPT-2 models (Radford et al. 2019) where the Owner and Public have different preferences for response length. The system operates on the WikiText-2 (Merity et al. 2016) dataset, with agents preferring different word count ranges for generated responses.

Alignment Scenarios. We test three text-based alignment scenarios: *(i) Perfect Alignment:* Both agents prefer exactly 4 words (Owner: 4-4, Public: 4-4); *(ii) Partial Alignment:* Owner prefers 2-4 words, Public prefers 4-6 words (overlapping range); *Disjoint Alignment:* Owner prefers 1-3 words, Public prefers 5-6 words (non-overlapping ranges).

Experimental Parameters. (i) Initial dataset: 10000 sentences from WikiText-2 filtered by word count; (ii) Owner curation: Select 1000 sentences using the BT mechanism (iii) Generation: Fine-tune the model for 2 epochs, generate 2000 responses with temperature 0.8; (iv) Public curation: Select 1000 responses using the BT mechanism (v) Training: Learning rate 5×10^{-5}, batch size 8.

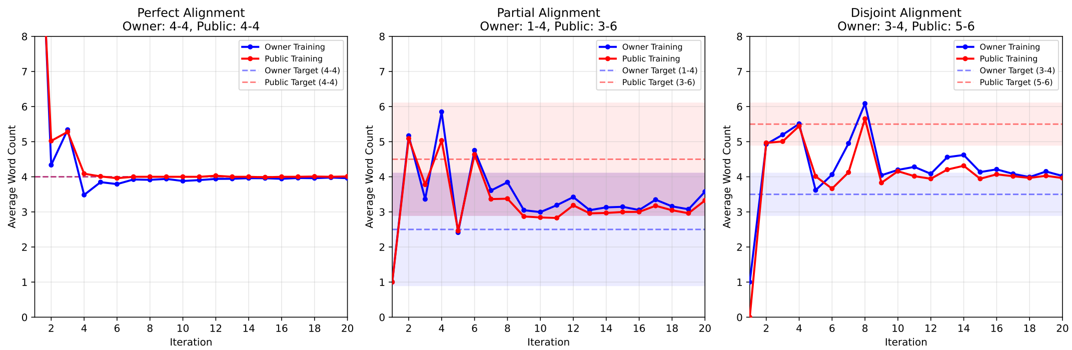

Figure 3: Word count evolution across three alignment scenarios: perfect alignment (agents target 4 words), partial alignment (owner targets 1-3 words, public targets 3-5 words), and disjoint alignment (owner targets 3-4 words, public targets 5-6 words).

Results and Analysis. Figure 3 shows the evolution of average word count across iterations for all three scenarios. The results demonstrate the same convergence patterns as the synthetic experiments. *Perfect Alignment:* Both training stages converge to the shared target of 4 words within 10 iterations, achieving near-perfect alignment with minimal variance. *Partial Alignment:* The system converges to the intersection of preferences (around 3-4 words), with the Owner's training showing slightly lower word counts and the Public's training showing slightly higher counts, but both remaining within the overlapping range. *Disjoint Alignment:* The system shows a two-stage convergence process. Initially, both training stages converge toward the Owner's preferred range (3-4 words). However, the Public's training gradually shifts toward longer responses, eventually settling around 4 words, demonstrating the Public's ability to refine within the Owner's preferred domain.

Key Findings. Our experiments validate several theoretical predictions. All scenarios show exponential convergence to equilibrium distributions, with convergence rates inversely proportional to the degree of alignment. In all cases, the final distribution concentrates on a subset of the original support, confirming the shrinking support principle (Remark 1). In disjoint alignment scenarios, the Owner's preferences dominate the support selection, while the Public refines within that support. Partial alignment scenarios preserve diversity only within the intersection of preferred regions, with exponential suppression elsewhere. These phenomena persist across different domains (continuous vs. discrete text) and different model architectures (GMM vs. transformer).

Conclusion

Alignment is not a one-time setting. In recursive curation, even well-intended preference aggregation can reduce diversity, confer first mover power, and lock models into narrow equilibria. We identified three convergence regimes, consensus collapse, lowest common denominator compromise, and owner-led refinement, showing how small preference gaps can redirect the long-run trajectory of generative systems.

Implications for AI alignment. Our findings suggest that we must reconceptualize alignment itself. First, it moves the problem from one-time preference-matching to dynamic mechanism design. The paper demonstrates that the structure of the alignment process, the sequential curation, the power dynamics, is a value-driven system that actively shapes outcomes, privileging owner-defined constraints over public refinement. Second, the impossibility theorem is not a failure but a clarification. It suggests that the goal of a single, stable, and diverse alignment may be a contradiction in terms. This should force the field to pivot from designing for consensus to designing for pluralism. The challenge is not to eliminate disagreement but to build systems that can productively manage it, treating "lowest common denominator" outcomes as a failure state of the mechanism, not an acceptable compromise. Finally, this work implies that the true alignment "meta-problem" is not just aligning a model, but aligning the recursive alignment process itself. We must design the feedback loop, the "game", to be transparent, contestable, and fair. Instead of a one-time setting, alignment becomes a continuous process of governing the system that governs the models.

Limitations and Future Work. This analysis is based on a model with key simplifying assumptions. First, the analysis rests on the BT model, which assumes that preferences are independent and transitive. Future work must extend this analysis to more complex preference models. Second, the two-agent "Owner-Public" game is a simplification of a complex ecosystem. The next step is to model this as an n-agent game with heterogeneous agents. Third, the model assumes static preferences. A more realistic analysis would treat preferences and model outputs as co-evolving. Future research should explore this co-evolutionary dynamic, where the outputs of one generation's model actively shape the preferences of the agents who curate the next. Finally, this paper is descriptive; it explains what will happen. The next step is normative mechanism design. Given the impossibility theorem, how can we design new curation systems that explicitly and transparently choose which property to sacrifice? This moves the alignment challenge beyond the realm of optimization and into the domain of political and social governance. The task is no longer to discover a single, mathematically "correct" alignment, but to engineer a process that is perceived as legitimate, transparent, and fair by all stakeholders.

Acknowledgements

The authors thank the Natural Sciences and Engineering Research Council of Canada, the University of Waterloo José Blakeley Scholarship for financial support. We are also grateful to the reviewers for their valued feedback on the paper.

References

Alemohammad, S.; Casco-Rodriguez, J.; Luzi, L.; Humayun, A. I.; Babaei, H.; LeJeune, D.; Siahkoohi, A.; and Baraniuk, R. 2023. Self-consuming generative models go mad. In *The Twelfth International Conference on Learning Representations*.

Bishop, C. M. 2006. *Pattern Recognition and Machine Learning*. Springer.

Bradley, R. A.; and Terry, M. E. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4): 324–345.

Conitzer, V.; Freedman, R.; Heitzig, J.; Holliday, W. H.; Jacobs, B. M.; Lambert, N.; Mossé, M.; Pacuit, E.; Russell, S.; Schoelkopf, H.; et al. 2024. Social choice should guide ai alignment in dealing with diverse human feedback. *arXiv preprint arXiv:2404.10271*.

Dohmatob, e. a. 2024. Model Collapse Demystified: The Case of Regression. *arXiv preprint arXiv:2402.07712*.

Eckersley, P. 2019. Impossibility and Uncertainty Theorems in AI Value Alignment (or why your AGI should not have a utility function). *arXiv preprint arXiv:1901.00064*.

Ferbach, D.; Bertrand, Q.; Bose, A. J.; and Gidel, G. 2024. Self-Consuming Generative Models with Curated Data Provably Optimize Human Preferences. *arXiv preprint arXiv:2407.09499*.

Ge, L.; Halpern, D.; Micha, E.; Procaccia, A. D.; Shapira, I.; Vorobeychik, Y.; and Wu, J. 2024. Axioms for ai alignment from human feedback. *Advances in Neural Information Processing Systems*, 37: 80439–80465.

Gerstgrasser, M.; Schaeffer, R.; Dey, A.; Rafailov, R.; Sleight, H.; Hughes, J.; Korbak, T.; Agrawal, R.; Pai, D.; Gromov, A.; Roberts, D. A.; Yang, D.; Donoho, D. L.; and Koyejo, S. 2024. Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data. *arXiv preprint arXiv:2404.01413*.

Merity, S.; Xiong, C.; Bradbury, J.; and Socher, R. 2016. Pointer Sentinel Mixture Models. arXiv:1609.07843.

Mishra, A. 2023. AI Alignment and Social Choice: Fundamental Limitations and Policy Implications. *arXiv preprint arXiv:2310.16048*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Qiu, T. 2024. Representative Social Choice: From Learning Theory to AI Alignment. *arXiv preprint arXiv:2410.23953*.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI*.

Shumailov, I.; Shumaylov, Z.; Zhao, Y.; Gal, Y.; Papernot, N.; and Anderson, R. 2023. The Curse of Recursion: Training on Generated Data Makes Models Forget. *arXiv preprint arXiv:2305.17493*.

Sun, H.; Shen, Y.; and Ton, J.-F. 2024. Rethinking bradley-terry models in preference-based reward modeling: Foundations, theory, and alternatives. *arXiv preprint arXiv:2411.04991*.

Tewolde, E.; Conitzer, V.; Freedman, R.; Heitzig, J.; Holliday, W. H.; Jacobs, B. M.; Lambert, N.; Mossé, M.; Pacuit, E.; Russell, S.; Schoelkopf, H.; and Zwicker, W. S. 2024. Social Choice Should Guide AI Alignment in Dealing with Diverse Human Feedback. *arXiv preprint arXiv:2404.10271*.

Wu, W.; Niezink, N.; and Junker, B. 2022. A diagnostic framework for the Bradley–Terry model. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 185: S461–S484.

Xiao, J.; Li, Z.; Xie, X.; Getzen, E.; Fang, C.; Long, Q.; and Su, W. J. 2024. On the algorithmic bias of aligning large language models with RLHF: Preference collapse and matching regularization. *arXiv preprint arXiv:2405.16455*.

Xu, S.; He, H.; and Cheng, G. 2025. A Probabilistic Perspective on Model Collapse. *arXiv preprint arXiv:2505.13947*.

Zhang, Y.; Li, B.; Liu, S.; Zhou, Z.; and Shi, Z. 2024. Beyond Bradley-Terry Models: A General Preference Model for Language Model Alignment. *arXiv preprint arXiv:2410.02197*.

Notation

Basic Spaces and Sets	
$(\mathcal{X}, d)$	Compact metric space representing the content domain
$\mathcal{P}(\mathcal{X})$	Space of probability measures on $\mathcal{X}$
$\mathcal{N} = \{O, P\}$	Set of agents (Owner and Public)
$B_\eta(A)$	η-neighborhood: $\{x \in \mathcal{X} : \inf_{y \in A} d(x, y) < \eta\}$
Reward Functions and Optimal Sets	
$r_O, r_P : \mathcal{X} \to \mathbb{R}$	Continuous reward functions for Owner and Public
$r_i^{\max}$	Maximum reward: $r_i^{\max} = \max_{x \in \mathcal{X}} r_i(x)$ for $i \in \{O, P\}$
A_O, A_P	Optimal sets: $A_i = \arg\max_{x \in \mathcal{X}} r_i(x)$
$A_\star$	Shared maximizer set in perfect alignment: $A_\star := A_O = A_P$
A_{shared}	Intersection of optima in partial alignment: $A_{\text{shared}} := A_O \cap A_P$
$A_{P\mid O}$	Public-refined owner optima: $\arg\max_{x \in A_O} r_P(x)$
BT Mechanism	
$H^p_{K,r}(x)$	BT weight: $\mathbb{E}_{Y_1,\dots,Y_{K-1} \sim p}\left[\dfrac{K e^{r(x)}}{e^{r(x)} + \sum_{j=1}^{K-1} e^{r(Y_j)}}\right]$
K, M	Pool sizes for Owner and Public curation stages
K_0	Threshold for pool size: $K_0 := \left\lceil 1 + \dfrac{8}{(\Delta_O - \delta)\, v_{\min}} \right\rceil$
Distributions and Evolution	
p_0	Initial distribution
p_t	Distribution at iteration t
$\tilde{p}_t$	Intermediate distribution after Owner curation
$\hat{p}_t$	Distribution after Public curation
p_{t+1}	Updated distribution
p_∞	Limiting distribution: $\lim_{t \to \infty} p_t$
$\mathcal{M}_t$	Generative model at iteration t
$\mathcal{D}_t$	Dataset at iteration t
$\mathcal{D}_1$	Initial dataset
$\mathcal{O}_t^*$	Outputs added to the dataset at iteration t
Gap and Reward Parameters	
Δ_O	Owner gap: $\Delta_O = \min_{x \notin A_O}\left[r_O^{\max} - r_O(x)\right]$
Δ_P	Public gap: $\Delta_P = \min_{x \notin A_P}\left[r_P^{\max} - r_P(x)\right]$
$\Delta_{P\mid O}$	Public refinement gap: $\Delta_{P\mid O} = \min_{x \in A_O \setminus A_{P\mid O}}\left[\max_{y \in A_{P\mid O}} r_P(y) - r_P(x)\right]$
Mass and Probability Measures	
v_t	Buffer mass: $v_t := p_t(B)$ for a buffer set $B = B_\eta(\cdot)$
$v_{\min}$	Minimum buffer mass: $v_{\min} := v_0 = p_0(B)$
m_t	Outside mass: $m_t := p_t(\mathcal{X} \setminus B)$
m_0	Initial outside mass: $m_0 := p_0(\mathcal{X} \setminus B)$
N_B	Number of competitors in buffer: $N_B = \sum_{j=1}^{K-1} \mathbf{1}\{Y_j \in B\}$
Bounds and Constants	
C, c	Constants in exponential bounds: Ce^{-ct}
$H_{\inf}$	Infimum of weights inside buffer: $H_{\inf} := \inf_{x \in B} H(x)$
$H_{\sup}$	Supremum of weights outside buffer: $H_{\sup} := \sup_{y \notin B} H(y)$
C_O	Constant in Owner weight bound: $C_O := \dfrac{4}{v_{\min}} + e^{\delta/2}$
ρ	Contraction factor: $\rho := \dfrac{C_O}{v_{\min}} \exp\left[-(\Delta_O - 2\delta)\right]$

Proofs

Perfect alignment: convergence on a shared maximizer set

Theorem

Theorem 1 (Perfect alignment: convergence on a shared maximizer set)**.** *Assume the Owner and the Public share a common set of optimal points:*

$$A_\star := A_O = A_P \subset \mathcal{X}, \qquad A_\star \neq \varnothing.$$

Let $(p_t)_{t\geq 0}$ be the sequence of distributions generated by the curation loop with BT feedback. Then the following hold:

(i) ***Exponential suppression outside the optimum.*** *For every $\eta > 0$, there exist constants $C > 0$, $c > 0$ such that for all $t \geq 0$,*

$$p_t\big(\mathcal{X} \setminus B_\eta(A_\star)\big) \leq Ce^{-ct}.$$

(ii) ***Weak convergence to a renormalized limit on $A_\star$.*** *The sequence (p_t) converges weakly to a limit distribution p_∞ supported on $A_\star$, given by*

$$p_\infty(x) = \frac{p_0(x)}{\displaystyle\int_{A_\star} p_0(z)\,dz} \cdot \mathbf{1}_{A_\star}(x).$$

To prove this theorem, we establish several key lemmas that build upon each other.

Lemma 1 (Buffer Mass Preservation)**.** *Fix $\delta \in \big(0, \frac{1}{2}\Delta_O\big)$. By continuity of r_O and compactness of $\mathcal{X}$, there exists $\eta > 0$ such that*

$$r_O(x) \leq r_O^{\max} - \Delta_O + \delta, \qquad \forall x \notin B_\eta(A_\star), \tag{7}$$

$$r_O(x) \geq r_O^{\max} - \tfrac{\delta}{2}, \qquad \forall x \in B_\eta(A_\star). \tag{8}$$

Let $B := B_\eta(A_\star)$ and $v_t := p_t(B)$. For each $t \geq 0$ define

$$H(x) := H^{p_t}_{K,r_O}(x), \qquad \mathcal{Z}_t := \int_{\mathcal{X}} p_t(z)\,H(z)\,dz.$$

Then $v_{t+1} = \dfrac{\int_B p_t H}{\mathcal{Z}_t}$ is nondecreasing in t, and $v_t \geq v_{\min} := v_0 > 0$ for all $t \geq 0$. In particular, with

$$H_{\inf} := \inf_{x\in B} H(x), \qquad H_{\sup} := \sup_{y\notin B} H(y),$$

we have $0 < H_{\sup} \leq H_{\inf}$ for every $t \geq 0$.

Proof. Since r_O is continuous and $\mathcal{X}$ is compact, the gap

$$\Delta_O := \min_{x\notin A_O} \big(r_O^{\max} - r_O(x)\big) > 0 \tag{9}$$

is attained. Choose $\eta > 0$ so that (7)–(8) hold; this is possible by continuity. For any $x \in B$ and $y \notin B$,

$$r_O(x) - r_O(y) \;\geq\; (r_O^{\max} - \tfrac{\delta}{2}) - (r_O^{\max} - \Delta_O + \delta) \;=\; \Delta_O - \tfrac{3}{2}\delta \;\geq\; 0, \tag{10}$$

because $\delta < \frac{1}{2}\Delta_O$ implies $\delta \leq \frac{2}{3}\Delta_O$. Fix any realization of competitors $Y_{1:K-1}$. The map $u \mapsto \frac{Ke^u}{e^u + \sum_j e^{r_O(Y_j)}}$ is strictly increasing in u; hence for all such realizations

$$\frac{Ke^{r_O(x)}}{e^{r_O(x)} + \sum_j e^{r_O(Y_j)}} \;\geq\; \frac{Ke^{r_O(y)}}{e^{r_O(y)} + \sum_j e^{r_O(Y_j)}}, \qquad x \in B,\; y \notin B. \tag{11}$$

Taking expectation over $Y_{1:K-1} \sim p_t$ yields $H(x) \geq H(y)$ for $x \in B$, $y \notin B$, so $0 < H_{\sup} \leq H_{\inf}$. Let $F := \int_B p_t H$, $G := \int_{\mathcal{X}\setminus B} p_t H$. Then $v_{t+1} = F/(F+G)$. Using $H(x) \geq H_{\inf}$ on B and $H(x) \leq H_{\sup}$ on $\mathcal{X} \setminus B$,

$$F \geq H_{\inf}\,v_t, \qquad G \leq H_{\sup}\,(1 - v_t).$$

Thus

$$v_{t+1} \;=\; \frac{F}{F+G} \;\geq\; \frac{H_{\inf} v_t}{H_{\inf} v_t + H_{\sup}(1-v_t)} \;=\; f_\zeta(v_t), \qquad \zeta := \frac{H_{\sup}}{H_{\inf}} \leq 1. \tag{12}$$

A direct computation gives

$$f_\zeta(v) - v = \frac{(1-\zeta)\, v(1-v)}{v + \zeta(1-v)} \;\geq\; 0, \qquad v \in [0,1],\ \zeta \in (0,1], \tag{13}$$

hence $v_{t+1} \geq v_t$ for all t. Since $v_0 = p_0(B) > 0$ by assumption, it follows by induction that $v_t \geq v_{\min} := v_0 > 0$ for all $t \geq 0$. ☐

Lemma 2 (Exponential Weight Suppression). *Fix $\delta \in \big(0, \frac{1}{2}\Delta_O\big)$ and let $\eta > 0$ be as in Lemma 1. With $B := B_\eta(A_\star)$ and $v_{\min} := v_0 = p_0(B) > 0$, define*

$$K_0 \;:=\; \left\lceil 1 + \frac{8}{(\Delta_O - \delta)\, v_{\min}} \right\rceil, \qquad C_O \;:=\; \frac{4}{v_{\min}} + e^{\delta/2}.$$

Then for all $t \geq 0$, all $K \geq \max\{2, K_0\}$, and all $x \notin B$,

$$H^{p_t}_{K,r_O}(x) \;\leq\; C_O \,\exp\!\Big(-\big(\Delta_O - \tfrac{3}{2}\delta\big)\Big).$$

Proof. Fix $t \geq 0$ and $x \notin B$. Let $Y_1, \ldots, Y_{K-1} \overset{\text{i.i.d.}}{\sim} p_t$ and $N_B := \sum_{j=1}^{K-1} \mathbf{1}\{Y_j \in B\}$. By Lemma 1, $v_t = p_t(B) \geq v_{\min}$, so $\mathbb{E}[N_B] = (K-1)v_t \geq (K-1)v_{\min}$. Chernoff's lower tail with parameter $v_{\min}$ and $\alpha = \frac{1}{2}$ gives

$$\Pr\big[N_B < \tfrac{1}{2}(K-1)v_{\min}\big] \;\leq\; \exp\!\Big(-\tfrac{1}{8}(K-1)v_{\min}\Big). \tag{14}$$

By the definition of K_0, for all $K \geq K_0$,

$$\Pr\big[N_B < \tfrac{1}{2}(K-1)v_{\min}\big] \;\leq\; e^{-(\Delta_O - \delta)}. \tag{15}$$

Let $E := \{N_B \geq \frac{1}{2}(K-1)v_{\min}\}$. On E, at least $\frac{1}{2}(K-1)v_{\min}$ competitors lie in B, and by (8)

$$\sum_{j=1}^{K-1} e^{r_O(Y_j)} \;\geq\; \tfrac{1}{2}(K-1)v_{\min}\, e^{\,r_O^{\max} - \delta/2}. \tag{16}$$

For $x \notin B$, by (7),

$$e^{r_O(x)} \;\leq\; e^{\,r_O^{\max} - (\Delta_O - \delta)}. \tag{17}$$

Therefore, conditioning on $Y_{1:K-1}$,

$$\frac{K e^{r_O(x)}}{e^{r_O(x)} + \sum_{j=1}^{K-1} e^{r_O(Y_j)}} \;\leq\; \frac{K\, e^{\,r_O^{\max} - (\Delta_O - \delta)}}{\frac{1}{2}(K-1)v_{\min}\, e^{\,r_O^{\max} - \delta/2}} = \frac{2K}{(K-1)v_{\min}}\, e^{-(\Delta_O - \frac{3}{2}\delta)} \;\leq\; \frac{4}{v_{\min}}\, e^{-(\Delta_O - \frac{3}{2}\delta)}, \tag{18}$$

using $\frac{K}{K-1} \leq 2$ for $K \geq 2$. Taking expectation yields

$$\mathbb{E}\Big[H^{p_t}_{K,r_O}(x)\mathbf{1}_E\Big] \;\leq\; \frac{4}{v_{\min}}\, e^{-(\Delta_O - \frac{3}{2}\delta)}. \tag{19}$$

On E^c, $H^{p_t}_{K,r_O}(x) \leq 1$, hence

$$\mathbb{E}\Big[H^{p_t}_{K,r_O}(x)\mathbf{1}_{E^c}\Big] \;\leq\; \Pr(E^c) \;\leq\; e^{-(\Delta_O - \delta)} \;=\; e^{\delta/2}\, e^{-(\Delta_O - \frac{3}{2}\delta)}. \tag{20}$$

Combining both parts,

$$H^{p_t}_{K,r_O}(x) = \mathbb{E}\Big[H^{p_t}_{K,r_O}(x)\Big] \;\leq\; \Big(\frac{4}{v_{\min}} + e^{\delta/2}\Big)\, e^{-(\Delta_O - \frac{3}{2}\delta)} \;=\; C_O\, e^{-(\Delta_O - \frac{3}{2}\delta)}. \tag{21}$$

All constants are independent of t. ☐

Proposition 3 (Contraction of Outside Mass). *Let $\delta \in \big(0, \frac{1}{2}\Delta_O\big)$ and let $\eta > 0$ be as in Lemma 1. With $B := B_\eta(A_\star)$, $m_t := p_t(\mathcal{X} \setminus B)$, and $v_{\min} := v_0 = p_0(B) > 0$, assume $K \geq \max\{2, K_0\}$ with K_0 as in Lemma **??**. Then for all $t \geq 0$,*

$$m_{t+1} \;\leq\; \rho\, m_t, \qquad \rho \;:=\; \frac{C_O}{v_{\min}}\; \exp\Big(-\big(\Delta_O - 2\delta\big)\Big).$$

Proof. Let $H(x) := H^{p_t}_{K,r_O}(x)$ and

$$\mathcal{Z}_t := \int_{\mathcal{X}} p_t(z)\, H(z)\, dz. \tag{22}$$

Decompose

$$\mathcal{Z}_t = \underbrace{\int_B p_t(z)\, H(z)\, dz}_{\text{in-buffer}} + \underbrace{\int_{\mathcal{X}\setminus B} p_t(z)\, H(z)\, dz}_{\text{out-of-buffer}}. \tag{23}$$

Step 1: Lower bound on the in-buffer contribution. For $x \in B$, (8) gives $r_O(x) \geq r_O^{\max} - \frac{\delta}{2}$. For any competitor multiset $\{Y_j\}$,

$$e^{r_O(x)} + \sum_{j=1}^{K-1} e^{r_O(Y_j)} \;\leq\; e^{r_O^{\max}} + (K-1)e^{r_O^{\max}} \;=\; K e^{r_O^{\max}}. \tag{24}$$

Hence, pointwise in $\{Y_j\}$,

$$\frac{K e^{r_O(x)}}{e^{r_O(x)} + \sum_{j=1}^{K-1} e^{r_O(Y_j)}} \;\geq\; \frac{K e^{\,r_O^{\max}-\delta/2}}{K e^{r_O^{\max}}} \;=\; e^{-\delta/2}. \tag{25}$$

Taking expectation over $Y_{1:K-1}$ yields the *uniform inner lower bound*

$$H(x) \;\geq\; e^{-\delta/2}, \qquad x \in B. \tag{26}$$

Therefore, using Lemma 1 ($p_t(B) \geq v_{\min}$),

$$\mathcal{Z}_t \;\geq\; \int_B p_t(z)\, H(z)\, dz \;\geq\; e^{-\delta/2}\, v_{\min}. \tag{27}$$

Step 2: Upper bound on the out-of-buffer contribution. By Lemma 2, for all $x \notin B$,

$$H(x) \;\leq\; C_O\, e^{-(\Delta_O - \frac{3}{2}\delta)}. \tag{28}$$

Hence

$$\int_{\mathcal{X}\setminus B} p_t(x)\, H(x)\, dx \;\leq\; C_O\, e^{-(\Delta_O - \frac{3}{2}\delta)}\, m_t. \tag{29}$$

Step 3: Contraction. The update gives

$$m_{t+1} = \frac{\int_{\mathcal{X}\setminus B} p_t(x)\, H(x)\, dx}{\mathcal{Z}_t}. \tag{30}$$

Combining (27) and (29),

$$m_{t+1} \;\leq\; \frac{C_O\, e^{-(\Delta_O - \frac{3}{2}\delta)}\, m_t}{e^{-\delta/2}\, v_{\min}} = \Big(\frac{C_O}{v_{\min}}\Big)\, e^{-(\Delta_O - 2\delta)}\, m_t = \rho\, m_t. \tag{31}$$

All constants are independent of t. $\square$

Now, we're ready to prove Theorem (1).

Proof of the Main Theorem

Proof of Theorem 1. **(i) Exponential suppression outside the optimum.** Fix $\delta \in \left(0, \frac{1}{2}\Delta_O\right)$ and let $\eta > 0$ be as in Lemma 1. Let $B := B_\eta(A_\star)$ and define $m_t := p_t(\mathcal{X} \setminus B)$. By Lemma 2 and Proposition 3, for any $K \in [\max\{2, K_0\}, \infty) \cap \mathbb{N}$ we have a time–independent factor $\rho \in (0,1)$ such that

$$m_{t+1} \;\le\; \rho\, m_t, \qquad t \ge 0. \tag{32}$$

Iterating yields

$$m_t \;\le\; m_0\, \rho^t. \tag{33}$$

Setting

$$C := m_0 = p_0(\mathcal{X} \setminus B_\eta(A_\star)), \qquad c := -\log \rho > 0, \tag{34}$$

We obtain the uniform exponential bound

$$p_t\big(\mathcal{X} \setminus B_\eta(A_\star)\big) \;=\; m_t \;\le\; C\, e^{-ct}, \qquad t \ge 0, \tag{35}$$

which proves **(i)**.

(ii) Weak convergence to a renormalized limit on $A_\star$. We now show that $p_t \Rightarrow p_\infty$, where

$$p_\infty(x) \;=\; \frac{p_0(x)}{\displaystyle\int_{A_\star} p_0(z)\, dz}\, \mathbf{1}_{A_\star}(x), \tag{36}$$

and the denominator is positive by the assumption $p_0(A_\star) > 0$.

Step 1: Any weak limit is supported on $A_\star$. Fix $\eta > 0$ and choose a continuous $\psi_\eta : \mathcal{X} \to [0,1]$ with $\psi_\eta \equiv 0$ on $B_\eta(A_\star)$ and $\psi_\eta \equiv 1$ on $\mathcal{X} \setminus B_{2\eta}(A_\star)$. Then $\int \psi_\eta\, dp_t \le p_t(\mathcal{X} \setminus B_\eta) \to 0$. By weak convergence along the subsequence $p_{t_j} \Rightarrow p^*$, $\int \psi_\eta\, dp^* = 0$, hence $p^*(\mathcal{X} \setminus B_{2\eta}) = 0$. Letting $\eta \downarrow 0$ yields $\mathrm{supp}(p^*) \subset A_\star$ hence $p^*\big(B_\eta(A_\star)\big) = 1$ and therefore $p^*\big(\mathcal{X} \setminus B_\eta(A_\star)\big) = 0$. As this holds for all $\eta > 0$, we conclude

$$\mathrm{supp}(p^*) \;\subseteq\; \bigcap_{\eta > 0} B_\eta(A_\star) \;=\; \overline{A_\star} \;=\; A_\star, \tag{37}$$

since $A_\star$ is closed.

Step 2: Ratio preservation on $A_\star$. Define the two–stage BT update (Owner then Public):

$$\tilde{p}_t(x) := \frac{p_t(x)\, H^{p_t}_{K, r_O}(x)}{\displaystyle\int_{\mathcal{X}} p_t(z)\, H^{p_t}_{K, r_O}(z)\, dz}, \tag{38}$$

$$p_{t+1}(x) := \frac{\tilde{p}_t(x)\, H^{\tilde{p}_t}_{M, r_P}(x)}{\displaystyle\int_{\mathcal{X}} \tilde{p}_t(z)\, H^{\tilde{p}_t}_{M, r_P}(z)\, dz}. \tag{39}$$

Since r_O and r_P are constant on $A_\star$ by the assumption $A_\star = A_O = A_P$, both $H^{p_t}_{K, r_O}(\cdot)$ and $H^{\tilde{p}_t}_{M, r_P}(\cdot)$ are constant on $A_\star$. Therefore, for any $x, y \in A_\star$,

$$\frac{p_{t+1}(x)}{p_{t+1}(y)} \;=\; \frac{p_t(x)}{p_t(y)} \;=\; \cdots \;=\; \frac{p_0(x)}{p_0(y)}. \tag{40}$$

Consequently, there exists a scalar $\alpha_t \in [0,1]$ such that the restriction of p_t to $A_\star$ equals $\alpha_t\, q$, where

$$q(x) \;:=\; \frac{p_0(x)}{\displaystyle\int_{A_\star} p_0(z)\, dz}\, \mathbf{1}_{A_\star}(x), \qquad \int_{A_\star} q(z)\, dz = 1. \tag{41}$$

Step 3: Identification and uniqueness of the weak limit. Let $p_{t_j} \Rightarrow p^*$ as above. For any Borel set $E \subseteq A_\star$,

$$p_{t_j}(E) \;=\; \alpha_{t_j}\, q(E). \tag{42}$$

Taking limits along the subsequence gives $p^*(E) = \alpha\, q(E)$ for some $\alpha \in [0,1]$ (a limit point of (α_{t_j})). From Step 1 we have $p^*(A_\star) = 1$ while $q(A_\star) = 1$, hence $\alpha = 1$ and $p^* = q$. Therefore every weakly convergent subsequence has the same limit q, and the entire sequence converges:

$$p_t \;\Rightarrow\; q \;=\; p_\infty. \tag{43}$$

This completes the proof of Theorem 1. □

Mode Collapse for Point Maximizers

Theorem

Corollary 2 (Mode Collapse for Point Maximizers)**.** *Suppose the curators have perfectly aligned preferences with a unique shared maximizer:*

$$A_O = A_P = \{x^\star\}.$$

Then:

(i) Exponential suppression outside the maximizer. *For every $\eta > 0$, there exist constants $C > 0$, $c > 0$ (independent of t) such that for all $t \geq 0$,*

$$p_t\big(\mathcal{X} \setminus B_\eta(x^\star)\big) \leq C\, e^{-ct}.$$

(ii) Weak convergence to the point mass. *$p_t \Rightarrow \delta_{x^\star}$ as $t \to \infty$.*

Proof. This corollary is a direct specialization of Theorem 1. We set the shared optimal set $A_\star = \{x^\star\}$. We assume $p_0(\{x^\star\}) > 0$, as the point is otherwise unreachable.

(i) Exponential Suppression. This follows directly from Theorem 1(i). Setting $A_\star = \{x^\star\}$ in that theorem's result immediately gives

$$p_t\big(\mathcal{X} \setminus B_\eta(x^\star)\big) \leq Ce^{-ct}. \tag{44}$$

(ii) Weak Convergence. This follows directly from Theorem 1(ii). The theorem states that the limiting distribution p_∞ is the initial distribution p_0 renormalized on $A_\star$:

$$p_\infty(x) = \frac{p_0(x)}{\displaystyle\int_{A_\star} p_0(z)\, dz} \cdot \mathbf{1}_{A_\star}(x). \tag{45}$$

Substituting $A_\star = \{x^\star\}$, the denominator becomes $\int_{\{x^\star\}} p_0(z)\, dz = p_0(\{x^\star\})$, which is positive. The limiting distribution is therefore

$$p_\infty(E) = \frac{p_0(E \cap \{x^\star\})}{p_0(\{x^\star\})}. \tag{46}$$

For any set E not containing $x^\star$, $p_\infty(E) = 0$. For any set E containing $x^\star$, $p_\infty(E) = 1$. This is precisely the Dirac point mass $\delta_{x^\star}$. Thus,

$$p_t \Rightarrow \delta_{x^\star}. \tag{47}$$

□

Partial Alignment: Consensus on Intersection

Theorem

Theorem 3 (Partial Alignment: Consensus on Intersection)**.** *Let $(\mathcal{X}, d)$ be a compact metric space. Let $r_O, r_P : \mathcal{X} \to \mathbb{R}$ be continuous, with nonempty argmax sets*

$$A_O := \arg\max_{\mathcal{X}} r_O, \qquad A_P := \arg\max_{\mathcal{X}} r_P,$$

and assume their intersection is nonempty:

$$A_{\textit{shared}} := A_O \cap A_P \neq \varnothing.$$

Assume the uniform gap condition:

$$\Delta_O := \inf_{x \notin A_O} \big(r_O^{\max} - r_O(x)\big) > 0, \qquad \Delta_P := \inf_{x \notin A_P} \big(r_P^{\max} - r_P(x)\big) > 0.$$

Let the two-stage dynamics be defined as in the main framework, starting from an initial p_0 with $p_0(A_{\textit{shared}}) > 0$. Then the following hold:

(i) Exponential suppression outside the intersection. *For every $\eta > 0$, there exist constants $C > 0$, $c > 0$ such that for all $t \geq 0$,*

$$p_t\big(\mathcal{X} \setminus B_\eta(A_{\textit{shared}})\big) \leq Ce^{-ct}.$$

(ii) ***Weak convergence to a renormalized limit on*** A_{shared}. *The sequence* (p_t) *converges weakly to a limit distribution* p_∞ *supported on* A_{shared}*, given by*

$$p_\infty(x) = \frac{p_0(x)}{\displaystyle\int_{A_{shared}} p_0(z)\,dz} \cdot \mathbf{1}_{A_{shared}}(x).$$

Proof. The proof follows the same two-part structure as Theorem 1, leveraging the sequential nature of the curation. The system's evolution is governed by the sequential application of the Owner's filter and then the Public's filter.

(i) Exponential suppression outside A_{shared}**.** We analyze the effect of the two sequential curation stages.

Stage 1: Owner Curation. The Owner filters p_t to produce $\tilde{p}_t$. The Owner's reward r_O has a gap $\Delta_O > 0$ for all $x \notin A_O$. By the same logic as Lemma 2 and Proposition 3, the Owner's filter exponentially suppresses all mass outside A_O. After this stage, the mass in $\tilde{p}_t$ is concentrated on A_O.

Stage 2: Public Curation. The Public filters $\tilde{p}_t$ to produce p_{t+1}. The Public's reward r_P has a gap $\Delta_P > 0$ for all $x \notin A_P$. The Public's filter now acts on $\tilde{p}_t$ (which lives on A_O) and, by the same logic, exponentially suppresses all mass outside A_P.

Combined Effect. Only the mass that is *not* suppressed by the Owner (i.e., mass in A_O) and *also* not suppressed by the Public (i.e., mass in A_P) can survive. This is precisely the intersection $A_{\text{shared}} = A_O \cap A_P$. Any $x \notin A_{\text{shared}}$ is either in $\mathcal{X} \setminus A_O$ (and is suppressed in Stage 1) or in $A_O \setminus A_P$ (and is suppressed in Stage 2). Therefore, the total mass outside A_{shared} contracts exponentially at each full iteration t. This proves **(i)**.

(ii) Weak convergence to a renormalized limit on A_{shared}**.** This part is identical to the logic in Theorem 1(ii).

Step 1: Any weak limit is supported on A_{shared}**.** From part (i), we have $p_t(\mathcal{X} \setminus A_{\text{shared}}) \to 0$. By the same measure-theoretic argument as in Theorem 1, any weak limit p^* must have $\text{supp}(p^*) \subseteq A_{\text{shared}}$.

Step 2: Ratio preservation on A_{shared}**.** For any $x, y \in A_{\text{shared}}$, by definition $x, y \in A_O$ and $x, y \in A_P$. Therefore, $r_O(x) = r_O(y) = r_O^{\max}$ and $r_P(x) = r_P(y) = r_P^{\max}$. Since the BT weight $H(\cdot)$ depends on x only through the reward $r(x)$, the weights $H^{p_t}_{K,r_O}(x)$ and $H^{\tilde{p}_t}_{M,r_P}(x)$ must be constant for all $x \in A_{\text{shared}}$. Since the update for any $x \in A_{\text{shared}}$ is just multiplication by these (content-independent) constants at each stage, the ratios are preserved:

$$\frac{p_{t+1}(x)}{p_{t+1}(y)} = \frac{p_t(x)}{p_t(y)} = \cdots = \frac{p_0(x)}{p_0(y)} \qquad \forall x, y \in A_{\text{shared}}. \tag{48}$$

Step 3: Identification and uniqueness of the weak limit. From Step 1, any weak limit p^* lives on A_{shared}. From Step 2, the distribution conditional on A_{shared} is invariant and equal to the renormalized p_0. Therefore, any weak limit p^* must be p_∞, which is the normalized restriction of p_0 to A_{shared}. $\square$

Disjoint alignment: owner dominance with public refinement

Theorem

Theorem 4 (Disjoint alignment: owner dominance with public refinement)**.** *Suppose the curators have disjoint preferences with* $A_O \cap A_P = \varnothing$*. Define the public-refined owner optima:*

$$A_{P|O} := \arg\max_{x \in A_O} r_P(x)$$

and assume a uniform gap exists for this refinement:

$$\Delta_{P|O} := \min_{x \in A_O \setminus A_{P|O}} \left[\max_{y \in A_{P|O}} r_P(y) - r_P(x) \right] > 0.$$

Let $(p_t)_{t \geq 0}$ *be the sequence of distributions generated by the curation loop, and assume* $p_0(A_{P|O}) > 0$*. Then the following hold:*

(i) ***Exponential suppression outside owner optima.*** *For every* $\eta > 0$*, there exist constants* $C > 0$*,* $c > 0$ *such that for all* $t \geq 0$*,*

$$p_t\big(\mathcal{X} \setminus B_\eta(A_O)\big) \leq Ce^{-ct}.$$

(ii) ***Exponential decay within owner optima but outside public refinement.*** *For every $\eta > 0$, there exist constants $C' > 0$, $c' > 0$ such that for all $t \geq 0$,*

$$p_t\big(A_O \setminus B_\eta(A_{P|O})\big) \leq C' e^{-c't}.$$

(iii) ***Weak convergence to public-refined equilibrium.*** *The sequence (p_t) converges weakly to a limit distribution p_∞ supported on $A_{P|O}$, given by*

$$p_\infty(x) = \frac{p_0(x)}{\displaystyle\int_{A_{P|O}} p_0(z)\,dz} \cdot \mathbf{1}_{A_{P|O}}(x).$$

To prove this theorem, we first establish two lemmas governing the two-stage suppression.

Lemma 4 (Public Refinement as Alignment Subproblem)**.** *Let the dynamics be concentrated on A_O, such that $\tilde{p}_t(A_O) \approx 1$. The Public curation stage $H^{\tilde{p}_t}_{M,r_P}(\cdot)$ acts as an alignment mechanism on the compact space A_O. Given the gap $\Delta_{P|O} > 0$, this stage exponentially suppresses mass on $A_O \setminus A_{P|O}$.*

Proof. This is a recursive application of the logic from Theorem 1. The Public curation stage is a recursive process on the distribution $\tilde{p}_t$. As $t \to \infty$, $p_t \to p^*$ (a limit on A_O), and thus $\tilde{p}_t \to \tilde{p}^*$ (also on A_O). The Public's filter $H^{\tilde{p}_t}_{M,r_P}(\cdot)$ on the space A_O seeks to align with r_P. The optimal set for r_P *within* A_O is $A_{P|O}$. The gap *within* A_O is $\Delta_{P|O} > 0$. By the same mechanism established in Lemma 2 and Proposition 3, the recursive application of this filter will exponentially suppress all mass in the suboptimal set $A_O \setminus A_{P|O}$. □

Lemma 5 (Ratio Preservation on $A_{P|O}$)**.** *For any $x, y \in A_{P|O}$, the total sequential update weight is constant, i.e., $F_t(x) = F_t(y)$, where $F_t(\cdot)$ represents the full $p_t \to p_{t+1}$ update.*

Proof. The full update from p_t to p_{t+1} involves two stages. Let $x, y \in A_{P|O}$.

1. **Stage 1 (Owner):** Since $x, y \in A_O$, $r_O(x) = r_O(y) = r_O^{\max}$. Thus, the Owner weights $H^{p_t}_{K,r_O}(x) = H^{p_t}_{K,r_O}(y)$.
2. **Stage 2 (Public):** Since $x, y \in A_{P|O}$, $r_P(x) = r_P(y) = \max_{z \in A_O} r_P(z)$. Thus, the Public weights $H^{\tilde{p}_t}_{M,r_P}(x) = H^{\tilde{p}_t}_{M,r_P}(y)$.

Since the update at both stages is a multiplication by a constant factor (independent of the choice of x or y), the total update $p_t \to p_{t+1}$ preserves the ratio $\frac{p_t(x)}{p_t(y)}$. □

Proof of Theorem 4. The proof demonstrates a two-stage cascade, corresponding to the sequential nature of the curation.

First, we establish suppression outside A_O. This is governed by the Owner's curation (Stage 1). The Owner's reward r_O has a uniform gap $\Delta_O > 0$ for all $x \notin A_O$. By the same logic as Lemma 2 and Proposition 3, this stage introduces an exponential contraction factor $\rho_O < 1$ for all mass outside A_O. The subsequent Public filter $H^{\tilde{p}_t}_{M,r_P}(\cdot)$ is bounded by M, so the total update for the outside mass $m_t := p_t(\mathcal{X} \setminus B_\eta(A_O))$ still contracts as $m_{t+1} \leq (M \cdot \rho_O) m_t$. For sufficiently large K, this is a contraction, ensuring $m_t \to 0$ exponentially. This proves **(i)**.

Second, we show refinement within A_O. As the mass p_t concentrates on A_O, the system's dynamics are increasingly dominated by the Public's filter (Stage 2) acting on this subspace. This creates a new, self-contained alignment subproblem on the compact set A_O. Lemma 4 formalizes this, showing that the Public's recursive filter exponentially suppresses all mass in the suboptimal region $A_O \setminus A_{P|O}$ due to the gap $\Delta_{P|O}$. This proves **(ii)**.

Finally, we prove convergence to the limit. From (i) and (ii), all mass outside $A_{P|O}$ is exponentially suppressed, so any weak limit p^* must be supported on $A_{P|O}$. Lemma 5 establishes that for any $x, y \in A_{P|O}$, the update weight is constant. Therefore, ratios are preserved within this set at every step t:

$$\frac{p_{t+1}(x)}{p_{t+1}(y)} = \frac{p_t(x)}{p_t(y)} = \cdots = \frac{p_0(x)}{p_0(y)}.$$

By the same argument as in Theorem 1(ii), the sequence (p_t) must converge to the only possible limit: the initial distribution p_0 restricted to $A_{P|O}$ and renormalized to unity. This completes the proof. □

Fundamental Limitation

Theorem

Theorem 5 (Fundamental Limitation)**.** *Let $(p_t)_{t\geq 0}$ be the sequence of distributions generated by the two-stage BT curation mechanism with reward functions $r_O, r_P : \mathcal{X} \to \mathbb{R}$ and pool sizes $K, M \geq 2$. For any non-trivial preference misalignment ($A_O \neq A_P$), it is impossible to simultaneously satisfy:*

1. ***Full Coverage**:*
$$\liminf_{t\to\infty} p_t(A_O \setminus A_P) > 0 \quad \textit{and} \quad \liminf_{t\to\infty} p_t(A_P \setminus A_O) > 0$$
2. ***Symmetric Influence**: There exists a permutation-invariant functional Φ such that*
$$p_\infty = \Phi(r_O, r_P, p_0) = \Phi(r_P, r_O, p_0)$$
3. ***Initial Distribution Independence**: For any two initial distributions $p_0, q_0 \in \mathcal{P}(\mathcal{X})$ with $\operatorname{supp}(p_0) = \operatorname{supp}(q_0) = \mathcal{X}$,*
$$p_\infty^{(p_0)} = p_\infty^{(q_0)}$$
where $p_\infty^{(p_0)}$ denotes the weak limit starting from p_0.

Proof. We prove this by showing that for any non-trivial misalignment $A_O \neq A_P$, at least one of the three conditions must fail. We analyze the two exhaustive cases of misalignment.

Case 1: Partial alignment ($A_O \cap A_P \neq \varnothing$). By Theorem 3, the mechanism converges to p_∞, where $\operatorname{supp}(p_\infty) \subseteq A_O \cap A_P$. This limit has $p_\infty(A_O \setminus A_P) = 0$ and $p_\infty(A_P \setminus A_O) = 0$, which directly violates **Full Coverage (1)**. Furthermore, Theorem 3 shows that p_∞ is the initial distribution p_0 renormalized on the intersection $A_O \cap A_P$. Because this limit p_∞ depends on p_0, this violates **Initial Distribution Independence (3)**. Since at least two conditions fail, the three cannot hold simultaneously.

Case 2: Disjoint alignment ($A_O \cap A_P = \varnothing$). By Theorem 4, the mechanism (with Owner as first mover) converges to $p_\infty = \Phi(r_O, r_P, p_0)$, where $\operatorname{supp}(p_\infty) \subseteq A_{P|O} \subseteq A_O$. This limit has $p_\infty(A_P \setminus A_O) = 0$, as its support is a subset of A_O, which directly violates **Full Coverage (1)**. Theorem 4 also shows that p_∞ is the initial distribution p_0 renormalized on $A_{P|O}$; this dependence on p_0 violates **Initial Distribution Independence (3)**. Finally, we must consider the permuted case $\Phi(r_P, r_O, p_0)$, which represents the outcome if the Public were the first mover. By the same logic of Theorem 4, this system would first collapse onto A_P, and then refine to $A_{O|P} := \arg\max_{x\in A_P} r_O(x)$. The resulting limit p'_∞ would have $\operatorname{supp}(p'_\infty) \subseteq A_P$. Since $A_O \cap A_P = \varnothing$, the two limits are disjoint, $p_\infty \neq p'_\infty$, and thus $\Phi(r_O, r_P, p_0) \neq \Phi(r_P, r_O, p_0)$. This violates **Symmetric Influence (2)**. Since all three conditions fail in this case, they cannot hold simultaneously.
In all possible misalignment regimes ($A_O \neq A_P$), at least one of the desired properties must fail. Therefore, the three properties cannot be satisfied simultaneously. This completes the proof. $\square$

Theorem

Theorem 6 (Weak Strategyproofness of Reports)**.** *For any reported pair (r'_O, r'_P) let $p_\infty(r'_O, r'_P)$ denote the weak limit of (p_t), and define each agent's utility by*
$$U_O(r'_O, r'_P) := \mathbb{E}_{x\sim p_\infty(r'_O, r'_P)}[\, r_O(x)\,], \qquad U_P(r'_O, r'_P) := \mathbb{E}_{x\sim p_\infty(r'_O, r'_P)}[\, r_P(x)\,].$$
Then truthful reporting is weakly dominant for both agents:
$$U_O(r_O, r'_P) \;\geq\; U_O(\hat{r}_O, r'_P) \quad \textit{and} \quad U_P(r'_O, r_P) \;\geq\; U_P(r'_O, \hat{r}_P) \qquad \forall\,(\hat{r}_O, \hat{r}_P),\ \forall\,(r'_O, r'_P).$$
In particular, $(r'_O, r'_P) = (r_O, r_P)$ is a (generally non–unique) Nash equilibrium.

Proof. Let $A_O := \arg\max r_O$, $A_P := \arg\max r_P$. For any report r'_O, write $\hat{A}_O := \arg\max r'_O$ and for any r'_P define
$$A_{P|\hat{O}} := \arg\max_{x\in \hat{A}_O} r'_P(x).$$
By the convergence and suppression results proved earlier for the two–stage BT dynamics, the limit distribution $p_\infty(r'_O, r'_P)$ is supported inside $A_{P|\hat{O}} \subseteq \hat{A}_O$. We use only this support characterization.

Owner's side. For any outcome distribution q and any $x \in X$, $r_O(x) \leq r_O^{\max} := \max_{z\in X} r_O(z)$, hence
$$U_O(r'_O, r'_P) \;=\; \mathbb{E}_{x\sim p_\infty(r'_O, r'_P)}[r_O(x)] \;\leq\; r_O^{\max}. \tag{49}$$

Equality is attained whenever the support of $p_\infty(r'_O, r'_P)$ is contained in A_O, since $r_O \equiv r_O^{\max}$ on A_O. In particular, with truthful reporting $r'_O = r_O$ we have $\hat{A}_O = A_O$ and the second stage cannot move mass out of A_O; therefore $U_O(r_O, r'_P) = r_O^{\max}$ for every r'_P. No misreport $\hat{r}_O$ can yield utility exceeding $r_O^{\max}$, so $U_O(r_O, r'_P) \geq U_O(\hat{r}_O, r'_P)$ for all $\hat{r}_O$ and all r'_P. Thus the owner's truthful report is weakly dominant.

Public's side. Fix any r'_O and write $\hat{A}_O = \arg\max r'_O$ as above. Because the limit support lies inside $A_{P|\hat{O}}$, we have

$$U_P(r'_O, r'_P) = \mathbb{E}_{x\sim p_\infty(r'_O, r'_P)}[r_P(x)] \;\leq\; \max_{x\in\hat{A}_O} r_P(x). \tag{50}$$

This upper bound is achievable by choosing $r'_P = r_P$, in which case $A_{P|\hat{O}} = \arg\max_{x\in\hat{A}_O} r_P(x)$ and r_P is constant on $A_{P|\hat{O}}$ at its maximum value over $\hat{A}_O$, so the expectation equals the bound. Consequently, for every fixed r'_O and every misreport $\hat{r}_P$,

$$U_P(r'_O, r_P) \;\geq\; U_P(r'_O, \hat{r}_P). \tag{51}$$

Therefore the public's truthful report is weakly dominant.

Since each player's truthful report is a best response to any report of the other, the truthful profile (r_O, r_P) is a Nash equilibrium. Non–uniqueness may occur: for instance, if $A_O \cap A_P \neq \varnothing$, any reports that keep the limit support inside $A_O \cap A_P$ make both agents payoff indifferent. □

Theorem

Theorem 7 (Fairness Impossibility for Sequential BT Aggregation)**.** *Consider the class of two–stage aggregation mechanisms on (X, d) with continuous r_O, r_P and pool sizes $K, M \geq 2$ defined by the BT updates from Theorem 6. For (r_O, r_P) let $p_\infty^{\mathrm{OF}}(r_O, r_P)$ be the limit outcome when the owner moves first, and $p_\infty^{\mathrm{PF}}(r_O, r_P)$ when the public moves first. Suppose a mechanism in this class satisfies the following properties for every (r_O, r_P):*

(EQ) Influence parity: *the outcome is order symmetric, that is*

$$p_\infty^{\mathrm{OF}}(r_O, r_P) \;=\; p_\infty^{\mathrm{PF}}(r_O, r_P).$$

(PO) Pareto optimality: *there is no distribution q on X with*

$$\mathbb{E}_q[r_O] \geq \mathbb{E}_{p_\infty}[r_O], \qquad \mathbb{E}_q[r_P] \geq \mathbb{E}_{p_\infty}[r_P],$$

and at least one inequality strict, where p_∞ denotes the mechanism's outcome.

(UC) Uniqueness: *for each (r_O, r_P) the outcome is unique, independent of initialization and of any tie breaking in the dynamics.*

Then no such mechanism exists. In other words, within the BT two–stage class, properties (EQ), (PO)*, and* (UC) *cannot all hold simultaneously.*

Proof. We argue by contradiction. Assume there is a two–stage BT mechanism satisfying (EQ), (PO), and (UC) for every (r_O, r_P).

Step 1: Disjoint preference profile. Choose continuous rewards with disjoint maximizer sets

$$A_O := \arg\max r_O, \qquad A_P := \arg\max r_P, \qquad A_O \cap A_P = \varnothing,$$

and with uniform positive gaps off the respective argmax sets. Such pairs exist on compact X.

Step 2: Owner first versus public first. By the convergence characterization for the two–stage BT dynamics, in the owner first order the limit support is contained in

$$S_{\mathrm{OF}} \;\subseteq\; A_{P|O} \;:=\; \arg\max_{x\in A_O} r_P(x) \;\subseteq\; A_O, \tag{52}$$

while in the public first order the limit support is contained in

$$S_{\mathrm{PF}} \;\subseteq\; A_{O|P} \;:=\; \arg\max_{x\in A_P} r_O(x) \;\subseteq\; A_P. \tag{53}$$

Because $A_O \cap A_P = \varnothing$, we have $S_{\mathrm{OF}} \cap S_{\mathrm{PF}} = \varnothing$.

Step 3: Order symmetry contradicts support separation. By (EQ) we must have $p_\infty^{\mathrm{OF}}(r_O, r_P) = p_\infty^{\mathrm{PF}}(r_O, r_P)$. But p_∞^{OF} is supported in S_{OF} and p_∞^{PF} is supported in S_{PF}, with $S_{\mathrm{OF}} \cap S_{\mathrm{PF}} = \varnothing$. This is impossible unless both outcomes assign zero mass to their own supports, which cannot happen by construction. Hence (EQ) cannot hold together with the BT support characterization for this disjoint profile.

Step 4: Why (PO) and (UC) do not rescue (EQ). One might attempt to reconcile (EQ) by selecting a compromise outcome supported outside $A_O \cup A_P$. That violates (PO): for any point outside A_O there exists $x^\star \in A_O$ with $r_O(x^\star) > r_O(x)$, and for any point outside A_P there exists $y^\star \in A_P$ with $r_P(y^\star) > r_P(x)$. No single compromise distribution can be undominated with respect to both extremes when both agents' payoffs strictly increase on their own argmax sets. Alternatively, one might try to randomize between S_{OF} and S_{PF} to enforce order symmetry, but then the outcome depends on the order or on initialization unless an extra selection rule is imposed, contradicting (UC). Therefore (PO) and (UC) cannot be simultaneously maintained with (EQ) for this profile.

We have reached a contradiction. Thus, within the two–stage BT class, no mechanism can satisfy (EQ), (PO), and (UC) for all preference profiles. $\square$