The Alignment Game: The Inevitable Conflict of Values for Shaping the Future of Generative Models

Anonymous authors
Paper under double-blind review

Abstract

As generative models are increasingly trained on their own outputs, data curation becomes the key force shaping what values persist. We formalize this recursive loop as a two-stage game between two agents: the model *Owner* and the *Public*. Each round, generative model outputs are filtered by both agents and returned to the training pool, progressively amplifying curator preferences. We analyze the dynamics under varying degrees of misalignment between the Owner and the Public—ranging from perfect alignment to partial and fully disjoint preferences—and show that the system converges exponentially to distinct long-run behaviours. Finally, we establish an *alignment trilemma*: No Bradley–Terry alignment process can simultaneously satisfy stability, diversity, and value alignment with both the Owner and the Public.

1 Introduction

Generative models are increasingly trained not just on human-authored datasets, but on data produced by earlier versions of themselves, resulting in *iterative retraining* or *self-consuming generation* (Shumailov et al., 2023; Ferbach et al., 2024; Alemohammad et al., 2024). This process is shaped by the interplay of two forces. First, the *The Owner*, typically the organization developing the model, filters outputs based on internal criteria, often through reward models or preference tuning (Ouyang et al., 2022; Bai et al., 2022). Second, the *The Public* contributes indirect supervision through engagement metrics such as clicks or upvotes, which are increasingly treated as proxies for user preferences (Glaese et al., 2022; Bubeck et al., 2023). Together, these forces determine what enters the training corpus for the next generation of models. We formalize this process as a *two-stage curation game*. In each round, the Owner samples outputs from the current model and selects a preferred subset using a Bradley–Terry selection rule (Bradley and Terry, 1952). The Public then filters this set using its own reward signals, capturing the way real-world platforms incorporate user feedback (Kreps et al., 2024). The selected outputs are added to the training set, and the model is retrained. Over time, this recursive loop drives the evolution of the model, with its trajectory shaped by the degree of alignment between the Owner's and the Public's preferences over what constitutes *"good"* data.

Our first contribution is a formal analysis of how the model's output distribution evolves over time under this two-stage curation process. We derive explicit update rules and characterize the system's long-term behavior across three alignment regimes: full alignment, partial alignment, and complete misalignment between the Owner and the Public. In all cases, we show that the system converges exponentially fast, either collapsing to a single point, concentrating on shared optima, or refining within the domain of one curator. As our second contribution, we identify a fundamental limitation that we term the *alignment trilemma*. We prove that no two-stage curation process based

on Bradley–Terry selection can simultaneously guarantee (i) value alignment with both curators, (ii) preservation output diversity, and (iii) stability with respect to the initial data distribution. This finding aligns with classic impossibility results in social choice theory and multi-agent learning (Arrow, 1951; Russell, 2022).

2 Mathematical Framework

Let $(\mathcal{X}, d)$ be a compact metric space. We consider two continuous reward functions $r_O, r_P : \mathcal{X} \to \mathbb{R}$ for the preferences of the Owner and the Public, respectively.

Definition 1 (Optimal Sets). *For any reward function $r : \mathcal{X} \to \mathbb{R}$, define the optimal set:*

$$A_r = \{x \in \mathcal{X} : r(x) = \max_{y \in \mathcal{X}} r(y)\}. \tag{1}$$

We denote $A_O = A_{r_O}$ and $A_P = A_{r_P}$ for the Owner's and Public's optimal sets, respectively.

Definition 2 (Open Ball). *For any $x^* \in \mathcal{X}$ and $\epsilon > 0$, define the open ball (neighborhood) around x^* of radius ϵ as:*

$$B_\epsilon(x^*) := \{x \in \mathcal{X} : d(x, x^*) < \epsilon\}. \tag{2}$$

This set contains all points in $\mathcal{X}$ within distance ϵ of x^*. It is used to describe local convergence and concentration behavior.

Definition 3 (Bradley–Terry Weights, generalized from Ferbach et al. (2024)). *For a probability measure p on $\mathcal{X}$, a pool size $K \geq 2$, and a reward function r, define the Bradley–Terry weight as:*

$$H^p_{K,r}(x) := \mathbb{E}_{Y_1,\ldots,Y_{K-1} \sim p}\left[\frac{K\, e^{r(x)}}{e^{r(x)} + \sum_{j=1}^{K-1} e^{r(Y_j)}}\right]. \tag{3}$$

Definition 4 (Alignment Types). *We distinguish three alignment regimes: **Perfect Alignment** where $A_O = A_P$ (curators share the same optimal set); **Partial Alignment** where $A_O \cap A_P \neq \emptyset$, $A_O \setminus A_P \neq \emptyset$, and $A_P \setminus A_O \neq \emptyset$; and **Disjoint Alignment** where $A_O \cap A_P = \emptyset$.*

Definition 5 (Misalignment Parameters). *Define the reward misalignment:*

$$\Delta_O = \min_{x \in \mathcal{X} \setminus A_O}\left[\max_{x^* \in A_O} r_O(x^*) - r_O(x)\right] > 0, \qquad \Delta_P = \min_{x \in \mathcal{X} \setminus A_P}\left[\max_{x^* \in A_P} r_P(x^*) - r_P(x)\right] > 0. \tag{4}$$

For partial alignment with shared optima $A_{shared} = A_O \cap A_P$, also define:

$$\Delta_{O,P} = \min_{x \in A_O \setminus A_P}\left[\max_{y \in A_P} r_P(y) - r_P(x)\right] > 0, \qquad \Delta_{P,O} = \min_{x \in A_P \setminus A_O}\left[\max_{y \in A_O} r_O(y) - r_O(x)\right] > 0. \tag{5}$$

Assumption 1 (Regularity Conditions). *Throughout, we assume: (i) r_O, r_P are continuous on $\mathcal{X}$, (ii)A_O, A_P are non-empty and compact, (iii) Pool sizes K, M are sufficiently large for the large-deviation bounds to hold.*

3 Problem Definition

We formalize a generative loop in which a model's outputs recursively reenter the public dataset, giving rise to a feedback-driven dynamic of curation and self-consumption. The system involves two agents: *the Public*, which maintains a public dataset $\mathcal{D}_t \subset \mathbb{R}^d$, evolving over time, and *the Owner*, who periodically curates from $\mathcal{D}_t$, trains a generative model $\mathcal{M}_t$, and thereby influences future iterations of $\mathcal{D}_{t+1}$ and $\mathcal{M}_{t+1}$. The retraining loop proceeds as follows:

1. **Initialization:** Begin with an initial public dataset $\mathcal{D}_1$.
2. **Owner Curation:** The Owner curates a subset $\mathcal{D}_1^* \subset \mathcal{D}_1$ using a reward function $r_O : \mathbb{R}^d \to \mathbb{R}$, forming the training set for generative model $\mathcal{M}_1$.
3. **Model Generation:** The model $\mathcal{M}_1$ produces synthetic outputs $\mathcal{O}_1 \sim \mathcal{M}_1$.
4. **Public Feedback:** The Public curates $\mathcal{O}_1$ using its own reward function $r_P : \mathbb{R}^d \to \mathbb{R}$, yielding a refined dataset $\mathcal{O}_1^* \subset \mathcal{O}_1$.
5. **Dataset Update:** The public dataset is updated as $\mathcal{D}_2 := \mathcal{D}_1 \cup \mathcal{O}_1^*$.
6. **Recursive Loop:** In the next round, the Owner curates $\mathcal{D}_2^* \subset \mathcal{D}_2$, trains a new model $\mathcal{M}_2$, and the process repeats.

Over time, the influence of the initial dataset $\mathcal{D}_1$ weakens, and the public dataset becomes dominated by curated generations:

$$\lim_{t\to\infty} \mathcal{D}_t \approx \bigcup_{i=1}^{t-1} \mathcal{O}_i^*. \tag{6}$$

Thus, the system enters a *self-consuming* regime where synthetic data dominates further training. To model this, let $p_t \in \mathcal{P}(\mathbb{R}^d)$ denote the output distribution of $\mathcal{M}_t$, trained on curated data $\mathcal{D}_t^*$. Each retraining step proceeds through two curation stages governed by Bradley–Terry mechanisms.

Owner Stage. The Owner samples a pool $\{x_1, \ldots, x_K\} \sim p_t$ and selects an output according to the Bradley–Terry selection rule using reward function r_O. This reweights the original distribution via the kernel:

$$H_{K,r_O}^{p_t}(x) := \mathbb{E}_{Y_1,\ldots,Y_{K-1}\sim p_t}\left[\frac{K \cdot e^{r_O(x)}}{e^{r_O(x)} + \sum_{j=1}^{K-1} e^{r_O(Y_j)}}\right]. \tag{7}$$

The resulting intermediate distribution is given by:

$$\tilde{p}_t(x) = p_t(x) \cdot H_{K,r_O}^{p_t}(x). \tag{8}$$

Public Stage. The Public applies its own reward function r_P to curate samples from $\tilde{p}_t$, again using a Bradley–Terry kernel with pool size M:

$$H_{M,r_P}^{\tilde{p}_t}(x) := \mathbb{E}_{Z_1,\ldots,Z_{M-1}\sim \tilde{p}_t}\left[\frac{M \cdot e^{r_P(x)}}{e^{r_P(x)} + \sum_{j=1}^{M-1} e^{r_P(Z_j)}}\right]. \tag{9}$$

The final post-curation distribution used for retraining is:

$$\hat{p}_t(x) = \tilde{p}_t(x) \cdot H_{M,r_P}^{\tilde{p}_t}(x). \tag{10}$$

The next model $\mathcal{M}_{t+1}$ is trained on samples drawn from $\hat{p}_t$, yielding an updated distribution:

$$p_{t+1}(x) \propto p_t(x) \cdot H_{K,r_O}^{p_t}(x) \cdot H_{M,r_P}^{\tilde{p}_t}(x), \qquad \text{where} \quad \tilde{p}_t(x) = p_t(x) \cdot H_{K,r_O}^{p_t}(x). \tag{11}$$

Remark 1. *Note that this two-stage process recursively couples the Owner and Public. The Owner first curates using reward r_O, and the resulting intermediate distribution is then curated by the Public using r_P, forming the new training distribution p_{t+1}. This feedback loop drives the evolution of the system.*

This recursive dynamic leads to the questions studied in this work: when does the system collapse to a degenerate point mass? When does it preserve diversity across iterations? And how do evolving preferences or asymmetric control shape the long-run behavior of such generative feedback loops?

4 Theoretical Analysis

Theorem 1 (Perfect Alignment: Mode Collapse)**.** *Suppose the curators have perfectly aligned preferences with a unique shared maximizer, i.e., $A_O = A_P = \{x^*\}$. Then:*

(i) Exponential decay outside the shared maximizer: *For any $\varepsilon > 0$, there exist constants $C_1, c_2 > 0$ such that*

$$p_t(\mathcal{X} \setminus B_\varepsilon(x^*)) \leq C_1 e^{-c_2 t}.$$

(ii) Convergence to point mass: *The sequence $p_t \Rightarrow \delta_{x^*}$ converges weakly to the Dirac delta at x^*.*

Thus, perfect agreement leads not to a balanced blend of values, but to degenerate outputs. Diversity is sacrificed for consensus. But what happens when the curators disagree, when each agent has its own optimal region, yet there remains a nonempty set of overlap?

Partial alignment. The next regime considers the more realistic case where the Owner and the Public share some values but diverge on others. In this scenario, we find that the model converges to a multimodal equilibrium, concentrating only on the intersection of optimal regions. Diversity can be preserved within this shared subset, but the eventual distribution remains sensitive to the initial conditions, breaking any hope of global stability.

Theorem 2 (Partial Alignment: Consensus on Intersection)**.** *Suppose the curators have partially aligned preferences with shared optima $A_{shared} = A_O \cap A_P \neq \emptyset$. Then:*

(i) Exponential decay outside shared optima*: For any $\varepsilon > 0$, there exist constants $C_3, c_4 > 0$ such that*

$$p_t(\mathcal{X} \setminus B_\varepsilon(A_{shared})) \leq C_3 e^{-c_4 t}.$$

(ii) Convergence to multi-modal equilibrium*: The limit $p_\infty = \lim_{t\to\infty} p_t$ exists and is supported on A_{shared}.*

(iii) Characterization*: On A_{shared}, the limiting density equals*

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

This regime is characterized by negotiated consensus: both agents exert influence, but only where their incentives align. The result preserves the alignment and diversity of the values, but the long-term model depends not just on preferences, but on where the training loop began.

Disjoint alignment. Finally, we consider the adversarial case in which the Owner and the Public have disjoint objectives. While this might suggest indecision, we show that the Owner dominates in determining the region of support, yet the Public sculpts the fine structure within it. The result is a refined convergence to the subset of the Owner's optima most acceptable to the Public.

Theorem 3 (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 the public refinement misalignment:

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

(i) Exponential decay outside owner optima: *For any $\varepsilon > 0$, there exist constants $C_5, c_6 > 0$ such that*

$$p_t(X \setminus B_\varepsilon(A_O)) \leq C_5 e^{-c_6 t}.$$

(ii) Exponential decay within owner optima but outside public refinement: *For any $\varepsilon > 0$, there exist constants $C', c' > 0$ such that*

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

(iii) Convergence to public-refined equilibrium: *The limit $p_\infty = \lim_{t\to\infty} p_t$ exists and is supported on $A_{P|O}$. Moreover, for all $x \in A_{P|O}$:*

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

The preceding theorems classify the long-term outcomes of recursive curation under different preference structures between the Owner and the Public. While these results reveal distinct behaviours – collapse, consensus, or selective refinement – they also expose a deeper structural constraint. In every alignment regime, some desirable property is sacrificed.

Part of this rigidity stems from the specific mechanism through which preferences are operationalized. In our framework, both curators apply a Bradley–Terry-style selection rule, which ranks samples by exponentiated rewards and samples proportionally, introducing a form of soft argmax pressure that amplifies reward peaks and suppresses tail mass. As a result, even moderate disagreement between curators is not gracefully negotiated but structurally constrained. The recursive loop becomes brittle: alignment forces mode collapse, diversity induces instability, and asymmetric preferences devolve into winner-takes-all equilibria. This leads to our impossibility result: no recursive generative loop governed by two Bradley–Terry-based curators can simultaneously achieve diversity, stability, and value alignment.

Theorem 4 (Fundamental Alignment Trilemma)**.** *Let $(p_t)_{t\geq 0}$ be the sequence of model output distributions generated by the two-curator loop on a compact space $\mathcal{X}$ with continuous rewards $r_O, r_P \colon \mathcal{X} \to \mathbb{R}$ and sufficiently large pool sizes K, M. Define the following desirable properties:*

(i) Value Alignment*: The weak limit p_∞ assigns positive probability to at least one maximizer of r_O and to at least one maximizer of r_P.*

(ii) Diversity*: p_∞ has strictly positive Shannon entropy $H(p_\infty) = -\int_{\mathcal{X}} p_\infty(x) \log p_\infty(x)\, dx > 0$.*

(iii) Stability*: The sequence (p_t) converges to a* unique *limit independent of the initial distribution p_0.*

Then no *recursive curation system can satisfy all three properties simultaneously; at most two can hold for any given alignment regime.*

5 Conclusion

We modeled recursive training of generative models as a two-agent game between the Owner and the Public, each applying their own reward-guided selection. We revealed how seemingly benign curation mechanisms can lead to sharp long-term effects: collapse, stagnation, or selective convergence. Even in the absence of noise or adversaries, recursive curation faces hard constraints: no selection mechanism based on Bradley–Terry-style rewards can achieve stability, diversity, and mutual value alignment at once. Future work might therefore explore alternative mechanisms to purely reward-maximizing feedback loops, such as smoothing kernels, entropy-aware filtering, or explicit negotiation between curators, that resist collapse while supporting pluralistic values.

References

Soroush Alemohammad, Javier Casco-Rodriguez, Luca Luzi, Atousa Humayun, Hossein Babaei, David LeJeune, Amir Siahkoohi, and Richard G Baraniuk. Self-consuming generative models go mad. *International Conference on Learning Representations (ICLR)*, 2024.

Kenneth J. Arrow. *Social Choice and Individual Values*. Yale University Press, 1951.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, and et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Sébastien Bubeck, Varun Chadrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.

Damien Ferbach, Quentin Bertrand, Avishek Joey Bose, and Gauthier Gidel. Self-consuming generative models with curated data provably optimize human preferences. *arXiv preprint arXiv:2407.09499*, 2024.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*, 2022.

David Kreps, Maxime Ferbach, Chris Jackson, and et al. Self-consuming generative models with curated data provably optimize human preferences. *arXiv preprint arXiv:2405.06810*, 2024.

Long Ouyang, Jeff Wu, Xu Jiang, and et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

Stuart Russell. Human compatible: Artificial intelligence and the problem of control. *Penguin*, 2022.

Ilia Shumailov, Zeming Shumaylov, Yinpeng Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*, 2023.

A Proofs

A.1 Perfect Alignment

Theorem 5 (Perfect Alignment, Mode Collapse)**.** *Suppose the curators have perfectly aligned preferences with a unique shared maximizer, i.e., $A_O = A_P = \{x^\star\}$. Then:*

Fix any buffer parameter δ with

$$0 < \delta < \tfrac{1}{2}\,\Delta_O,$$

and choose $\eta > 0$ so small that

$$r_O(x) \;\le\; r_O(x^\star) - (\Delta_O - \delta) \quad \forall\, x \notin B_\eta(x^\star),$$

where

$$B_\eta(x^\star) = \{\, x \in \mathcal{X} : d(x, x^\star) < \eta \}.$$

for every measurable $A \subseteq \mathcal{X}$, starting from any p_0 such that

$$v_{\min} = p_0\big(B_\eta(x^\star)\big) \;>\; 0.$$

Then there exists a threshold $K_0 \in \mathbb{N}$ and constants $C > 0$, $\rho \in (0,1)$, both depending only on $\Delta_O, \delta, \eta, v_{\min}$ and r_O, such that for all $K \ge K_0$ and all $t \ge 0$:

1. Exponential decay outside the maximizer.

$$p_t\big(\mathcal{X} \setminus B_\eta(x^\star)\big) \;\le\; C\,\rho^t.$$

2. Weak convergence to the maximizer.

$$p_t \;\Rightarrow\; \delta_{x^\star} \quad \text{as } t \to \infty.$$

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

Lemma 1 (Upper Bounds on Bradley-Terry Weights Away from $x^\star$)**.** *Fix a parameter $0 < \delta < \frac{1}{2}\Delta_O$, by continuity of r_O choose $\eta > 0$ such that*

$$\textit{(outer gap) } r_O(x) \le r_O(x^\star) - \Delta_O + \delta, \quad \forall x \notin B_\eta(x^\star), \tag{12}$$

$$\textit{(inner gap) } r_O(x) \ge r_O(x^\star) - \delta/2, \qquad \forall x \in B_\eta(x^\star). \tag{13}$$

Set

$$v_{\min} := v_0, \qquad K_0 := \left\lceil \tfrac{8}{(\Delta_O - \delta) v_{\min}} \right\rceil + 1, \qquad C_O := \frac{4}{v_{\min}} + 1.$$

Then, for every pool size $K \ge K_0$, for every round $t \ge 0$, and for every $x \notin B_\eta(x^\star)$,

$$H^{p_t}_{K,r_O}(x) \;\le\; C_O\, e^{-(\Delta_O - \frac{3}{2}\delta)}. \tag{14}$$

Proof. Let $B := B_\eta(x^\star)$ and define the in–buffer mass $v_t := p_t(B)$.

We show that the probability mass inside the fixed buffer $B := B_\eta(x^\star)$ does not decrease over time under the Bradley–Terry selection dynamics. We have

$$H(x) := H^{p_t}_{K,r_O}(x), \qquad Z_t = \int_X p_t(z)\, H(z)\, dz. \tag{15}$$

The update rule is

$$p_{t+1}(x) = \frac{p_t(x)\,H(x)}{\mathcal{Z}_t}, \quad \text{so} \quad v_{t+1} = \frac{\displaystyle\int_B p_t(x)\,H(x)\,dx}{\displaystyle\int_B p_t(z)\,H(z)\,dz + \int_{X\setminus B} p_t(z)\,H(z)\,dz}. \tag{16}$$

By the inner/outer gap in Eq.(13),

$$r_O(x) \geq r_O(y) \quad \text{for every } x \in B,\ y \notin B. \tag{17}$$

Since H is monotonically increasing in r_O,

$$H(x) \geq H(y) \quad (x \in B,\ y \notin B). \tag{18}$$

Define

$$H_{\text{inf}} = \inf_{x\in B} H(x), \qquad H_{\text{sup}} = \sup_{y\notin B} H(y). \tag{19}$$

Then $0 < H_{\text{sup}} \leq H_{\text{inf}}$.

Set

$$F = \int_B p_t(x)\,H(x)\,dx, \qquad G = \int_{X\setminus B} p_t(x)\,H(x)\,dx, \qquad \text{so } v_{t+1} = \frac{F}{F+G}. \tag{20}$$

Using the uniform bounds:

$$F \geq H_{\text{inf}}\,v_t, \qquad G \ \leq\ H_{\text{sup}}\,(1-v_t). \tag{21}$$

Comparing v_{t+1} to v_t,

$$v_{t+1} \ \geq\ \frac{H_{\text{inf}}\,v_t}{H_{\text{inf}}\,v_t + H_{\text{sup}}\,(1-v_t)} = \frac{v_t}{v_t + \alpha\,(1-v_t)}, \quad \text{with } \alpha := \frac{H_{\text{sup}}}{H_{\text{inf}}} \leq 1. \tag{22}$$

Define $f_\alpha(v) := \dfrac{v}{v+\alpha(1-v)}$. We have,

$$f_\alpha(v) - v = \frac{(1-\alpha)\,v\,(1-v)}{v+\alpha(1-v)} \ \geq\ 0 \quad \text{for } 0 \leq v \leq 1,\ 0 < \alpha \leq 1, \tag{23}$$

so $f_\alpha(v) \geq v$. Therefore $v_{t+1} \geq v_t$.

By induction,

$$v_t \geq v_0 \ =:\ v_{\min} > 0 \qquad \text{for all } t \geq 0. \tag{24}$$

Thus, the probability mass that p_t assigns to the buffer B never decreases.

Fix $x \notin B$ at round t and let the $K-1$ competitors be $Y_1, \ldots, Y_{K-1} \overset{\text{i.i.d.}}{\sim} p_t$. Define the count of "in–buffer" competitors

$$N_B := \sum_{j=1}^{K-1} \mathbf{1}\{Y_j \in B\}, \qquad \mathbb{E}[N_B] = (K-1)v_t \geq (K-1)v_{\min}. \tag{25}$$

(i) A high–probability event. By the Chernoff lower–tail bound,

$$\Pr\Big[N_B < \tfrac{1}{2}\,(K-1)v_{\min}\Big] \le \exp\Big[-\tfrac{1}{8}\,(K-1)v_{\min}\Big]. \tag{26}$$

Choose the pool–size threshold

$$K_0 \ge \frac{8}{(\Delta_O - \delta)\,v_{\min}}, \tag{27}$$

and restrict to $K \ge K_0$; then

$$\Pr\big[N_B < \tfrac{1}{2}(K-1)v_{\min}\big] \le e^{-(\Delta_O-\delta)}. \tag{28}$$

(ii) Bounding the Bradley–Terry ratio on the good event. Let

$$E := \Big\{N_B \ge \tfrac{1}{2}\,(K-1)v_{\min}\Big\}. \tag{29}$$

On E there are at least $\frac{1}{2}(K-1)v_{\min}$ competitors in B, each satisfying $r_O(Y_j) \ge r_O(x^\star) - \frac{\delta}{2}$, while $r_O(x) \le r_O(x^\star) - (\Delta_O - \delta)$. Hence

$$\sum_{j=1}^{K-1} e^{r_O(Y_j)} \ge \tfrac{1}{2}\,(K-1)v_{\min}\,e^{\,r_O(x^\star)-\delta/2}, \quad e^{r_O(x)} \le e^{\,r_O(x^\star)-(\Delta_O-\delta)}. \tag{30}$$

Therefore, on E,

$$H_{p_t,K,r_O}(x) = \frac{K\,e^{r_O(x)}}{e^{r_O(x)} + \sum_{j=1}^{K-1} e^{r_O(Y_j)}} \le \frac{2K}{(K-1)v_{\min}}\,\exp\Big[-\big(\Delta_O - \tfrac{3}{2}\delta\big)\Big] \le \frac{4}{v_{\min}}\,\exp\Big[-\big(\Delta_O - \tfrac{3}{2}\delta\big)\Big]. \tag{31}$$

(iii) Averaging over E and its complement. On E^c the trivial bound $H(x) \le 1$ holds, and $\Pr(E^c) \le e^{-(\Delta_O-\delta)}$. Taking expectations,

$$\mathbb{E}\big[H_{p_t,K,r_O}(x)\big] \le \frac{4}{v_{\min}}\,e^{-(\Delta_O-\frac{3}{2}\delta)} + e^{-(\Delta_O-\delta)} \le \Big(\tfrac{4}{v_{\min}} + 1\Big)\,e^{-(\Delta_O-\frac{3}{2}\delta)}. \tag{32}$$

Define the constant $C_O := \frac{4}{v_{\min}} + 1$. Because the final bound is deterministic, it holds pointwise:

$$H_{p_t,K,r_O}(x) \le C_O\,\exp\Big[-\big(\Delta_O - \tfrac{3}{2}\delta\big)\Big] \quad \text{for all } x \notin B,\ t \ge 0. \tag{33}$$

□

Lemma 2 (Buffer mass is bounded below)**.** *Let $B := B_\eta(x^\star)$ be the fixed buffer introduced in Lemma 1. For every round $t \ge 0$ the buffer mass*

$$v_t = p_t(B)$$

obeys the uniform lower bound

$$v_t \ge v_{\min}.$$

Lemma 3 (Uniform *lower* bound inside the buffer)**.** *For every round t, every pool size $K \geq 1$, and every $x \in B_\eta(x^\star)$,*

$$H^{p_t}_{K,r_O}(x) \geq \frac{1}{K}\, e^{-\delta/2}.$$

Proof. Inside the buffer $B_\eta(x^\star)$ we have $r_O(x) \geq r_O(x^\star) - \delta/2$. Hence

$$e^{r_O(x)} \geq e^{r_O(x^\star)-\delta/2}. \tag{34}$$

For any specific competitor multiset $\{Y_1, \ldots, Y_{K-1}\}$ we bound the denominator of the Bradley–Terry fraction:

$$e^{r_O(x)} + \sum_{j=1}^{K-1} e^{r_O(Y_j)} \leq e^{r_O(x^\star)} + (K-1)\, e^{r_O(x^\star)} = K\, e^{r_O(x^\star)}. \tag{35}$$

Therefore, conditional on those Y_j,

$$\frac{e^{r_O(x)}}{e^{r_O(x)} + \sum_{j=1}^{K-1} e^{r_O(Y_j)}} \geq \frac{e^{r_O(x^\star)-\delta/2}}{K\, e^{r_O(x^\star)}} = \frac{1}{K}\, e^{-\delta/2}. \tag{36}$$

Since this bound holds for every configuration of $Y_{1:K-1}$, taking the expectation over the competitor draw preserves it:

$$H^{p_t}_{K,r_O}(x) = \mathbb{E}_{Y_{1:K-1}}\Big[\frac{e^{r_O(x)}}{e^{r_O(x)} + \sum_{j=1}^{K-1} e^{r_O(Y_j)}}\Big] \geq \frac{1}{K}\, e^{-\delta/2}. \tag{37}$$

□

Proposition 6 (One–step contraction inequality)**.** *Fix $\delta \in \big(0, \frac{1}{2}\Delta_O\big)$ and let $K \geq K_0$. Define*

$$\rho = \frac{K\, C_O}{v_{\min}}\, \exp\Big[-\big(\Delta_O - \delta\big)\Big]. \tag{38}$$

Then, for every round $t \geq 0$,

$$m_{t+1} \;\leq\; \rho\, m_t. \tag{39}$$

Proof. Set $H(x) := H^{p_t}_{K,r_O}(x)$ and $Z_t := \int_{\mathcal{X}} p_t(z) H(z)\, dz$. Decompose

$$Z_t = \underbrace{\int_B p_t(z) H(z)\, dz}_{\text{inside buffer}} + \underbrace{\int_{\mathcal{X}\setminus B} p_t(z) H(z)\, dz}_{\text{outside buffer}}. \tag{40}$$

Lemma 2 ensures $p_t(B) \geq v_{\min}$. Lemma 3 gives $H(z) \geq e^{-\delta/2}/K$ for all $z \in B$. Hence

$$Z_t \geq \frac{e^{-\delta/2}}{K}\, v_{\min}. \tag{41}$$

Lemma 1 yields, for all $x \notin B$,

$$H(x) \leq C_O \, e^{-(\Delta_O - \frac{3}{2}\delta)}. \tag{42}$$

Therefore

$$\int_{\mathcal{X}\setminus B} p_t(x) H(x) \, dx \leq C_O \, e^{-(\Delta_O - \frac{3}{2}\delta)} \, m_t. \tag{43}$$

Because

$$m_{t+1} = \big[\text{numerator}\big] / Z_t, \tag{44}$$

the bounds (41)–(43) give

$$m_{t+1} \leq \frac{C_O \, e^{-(\Delta_O - \frac{3}{2}\delta)}}{(e^{-\delta/2}/K)\, v_{\min}} \; m_t \Big(\tfrac{K C_O}{v_{\min}}\Big) \; \exp\big[-(\Delta_O - 2\delta)\big] \; m_t = \rho \, m_t, \tag{45}$$

□

Now, we're ready to prove Theorem 5.

Proof of Theorem 5. By Proposition 6, for any $\varepsilon > 0$, the mass outside $B_\varepsilon(x^\star)$ decays exponentially. This establishes the first part of the theorem with $C = m_0$ and $c = -\log \rho > 0$.

For weak convergence, we show that for every bounded continuous function $f : \mathcal{X} \to \mathbb{R}$,

$$\lim_{t\to\infty} \int_{\mathcal{X}} f(x) p_t(x) dx = f(x^\star).$$

Let $M = \sup_{x\in\mathcal{X}} |f(x)| < \infty$ since f is bounded, and fix arbitrary $\eta > 0$.

By continuity of f at $x^\star$, there exists $\varepsilon > 0$ such that for all $x \in B = B_\varepsilon(x^\star)$,

$$|f(x) - f(x^\star)| < \frac{\eta}{2}. \tag{46}$$

Now write

$$\int_{\mathcal{X}} f(x) \, p_t(x) \, dx = \int_B f(x) \, p_t(x) \, dx + \int_{\mathcal{X}\setminus B} f(x) \, p_t(x) \, dx. \tag{47}$$

Then,

$$\left|\int_{\mathcal{X}} f(x)\, p_t(x)\, dx - f(x^\star)\right| = \left|\int_B (f(x) - f(x^\star))\, p_t(x)\, dx + \int_{\mathcal{X}\setminus B} (f(x) - f(x^\star))\, p_t(x)\, dx\right| \tag{48}$$

$$\leq \left|\int_B (f(x) - f(x^\star))\, p_t(x)\, dx\right| + \left|\int_{\mathcal{X}\setminus B} (f(x) - f(x^\star))\, p_t(x)\, dx\right| \tag{49}$$

$$\leq \int_B |f(x) - f(x^\star)|\, p_t(x)\, dx + \int_{\mathcal{X}\setminus B} |f(x) - f(x^\star)|\, p_t(x)\, dx \tag{50}$$

$$\leq \sup_{x\in B} |f(x) - f(x^\star)| \int_B p_t(x)\, dx + \sup_{x\in\mathcal{X}\setminus B} |f(x) - f(x^\star)| \int_{\mathcal{X}\setminus B} p_t(x)\, dx \tag{51}$$

$$\leq \frac{\eta}{2} \cdot 1 + 2M \cdot m_t, \tag{52}$$

where $m_t = \int_{\mathcal{X}\setminus B} p_t(x)\, dx$ is the "outside mass" and we used $|f(x) - f(x^\star)| \leq 2M$ everywhere.

By the Proposition 6, $m_t \to 0$ as $t \to \infty$. Therefore, there exists T such that for all $t \geq T$, $2Mm_t < \eta/2$. Thus, for all $t \geq T$,

$$\left|\int_{\mathcal{X}} f(x)\, p_t(x)\, dx - f(x^\star)\right| < \frac{\eta}{2} + \frac{\eta}{2} = \eta. \tag{53}$$

Since $\eta > 0$ was arbitrary, this proves

$$\lim_{t\to\infty} \int_{\mathcal{X}} f(x)\, p_t(x)\, dx = f(x^\star). \tag{54}$$

Hence $p_t \Rightarrow \delta_{x^\star}$ in the weak sense. □

A.2 Partial Alignment

Theorem 7 (Partial Alignment: Consensus on Intersection)**.** *Suppose the curators have partially aligned preferences with shared optima $A_{shared} = A_O \cap A_P \neq \emptyset$. Then:*

1. ***Exponential decay outside shared optima**: For any $\varepsilon > 0$, there exist constants $C, c > 0$ such that*
$$p_t(\mathcal{X} \setminus B_\varepsilon(A_{shared})) \leq Ce^{-ct}$$
2. ***Convergence to multi-modal equilibrium**: The limit $p_\infty = \lim_{t\to\infty} p_t$ exists and is supported on A_{shared}*
3. ***Characterization**: On A_{shared}, the limiting density equals*
$$p_\infty(x) = \frac{p_0(x)}{\int_{A_{shared}} p_0(z)dz} \mathbf{1}_{A_{shared}}(x)$$

We establish the result through a sequence of lemmas that characterize the behavior of the Bradley-Terry weights on different regions of the space.

Lemma 4 (Constant Weights on Shared Optima). *For all $t \geq 0$ and all $x, x' \in \mathcal{A}_{shared}$:*

$$H_{K,r_O}^{p_t}(x) = H_{K,r_O}^{p_t}(x'), \quad H_{M,r_P}^{\tilde{p}_t}(x) = H_{M,r_P}^{\tilde{p}_t}(x')$$

Proof. Fix $x \in \mathcal{A}_{\text{shared}}$. Since $r_O(x) = \max_{y \in \mathcal{X}} r_O(y)$, for any competitor y we have $r_O(y) - r_O(x) \leq 0$. The Bradley-Terry weight becomes:

$$H_{K,r_O}^{p_t}(x) = \mathbb{E}_{y_1,\ldots,y_{K-1} \sim p_t}\left[\frac{K}{1 + \sum_{j=1}^{K-1} \exp(r_O(y_j) - r_O(x))}\right]$$

The inner expression depends on the competitors $y_1, \ldots, y_{K-1}$ but not on which specific maximizer $x \in \mathcal{A}_{\text{shared}}$ we choose. Therefore, $H_{K,r_O}^{p_t}(x)$ is constant on $\mathcal{A}_{\text{shared}}$. The same argument applies to $H_{M,r_P}^{\tilde{p}_t}$. □

Lemma 5 (Uniform Suppression Outside Shared Optima). *Define $\rho := \frac{K}{1+(K-1)e^{-\Delta}} \cdot \frac{M}{1+(M-1)e^{-\Delta}} < 1$. Then for all $t \geq 0$ and $x \notin \mathcal{A}_{shared}$:*

$$H_{K,r_O}^{p_t}(x) \cdot H_{M,r_P}^{\tilde{p}_t}(x) \leq \rho$$

Proof. We consider three cases:

Case 1: $x \in \mathcal{A}_O \setminus \mathcal{A}_P$. Here $r_O(x) = r_O^{\max}$, so x is favored by the Owner. However, by the uniform gap assumption, $r_P(x) \leq r_P^{\max} - \Delta$. Since $r_P(y) \leq r_P^{\max}$ for all $y \in \mathcal{X}$:

$$r_P(y) - r_P(x) \geq -\Delta \implies \exp(r_P(y) - r_P(x)) \geq e^{-\Delta} \tag{55}$$

Applying this to the Bradley-Terry kernel:

$$H_{M,r_P}^{\tilde{p}_t}(x) = \mathbb{E}_{y_1,\ldots,y_{M-1} \sim \tilde{p}_t}\left[\frac{M}{1 + \sum_{j=1}^{M-1} \exp(r_P(y_j) - r_P(x))}\right] \leq \frac{M}{1 + (M-1)e^{-\Delta}} \tag{56}$$

Case 2: $x \in \mathcal{A}_P \setminus \mathcal{A}_O$. Symmetrically, $H_{K,r_O}^{p_t}(x) \leq \frac{K}{1+(K-1)e^{-\Delta}}$.

Case 3: $x \notin \mathcal{A}_O \cup \mathcal{A}_P$. Both bounds apply, yielding the product $\rho < 1$. □

Proposition 8 (Evolution of Probability Mass). *Define $F_t(x) := H_{K,r_O}^{p_t}(x) \cdot H_{M,r_P}^{\tilde{p}_t}(x)$. There exists $C_t > 0$ such that $F_t(x) = C_t$ for all $x \in \mathcal{A}_{shared}$, and $C_t \geq 1$ for all t.*

Proof. By Lemma 4, F_t is constant on $\mathcal{A}_{\text{shared}}$. Since both curators assign maximum weight to shared optima, and for $x \in \mathcal{A}_{\text{shared}}$ we have both $H_{K,r_O}^{p_t}(x) \geq 1$ and $H_{M,r_P}^{\tilde{p}_t}(x) \geq 1$, we obtain $C_t \geq 1$. □

Lemma 6 (Bounds on Normalizing Constant). *Let $S_t := \int_{\mathcal{A}_{shared}} p_t d\lambda$ and $Z_t := \int_{\mathcal{X}} p_t(z) F_t(z) d\lambda(z)$. Then:*

$$S_t \leq Z_t \leq C_t S_t + \rho(1 - S_t) \tag{57}$$

In particular, $Z_t \geq S_t$.

Proof. We have:

$$Z_t = \int_{\mathcal{A}_{\text{shared}}} p_t(x) C_t d\lambda(x) + \int_{\mathcal{X}\setminus\mathcal{A}_{\text{shared}}} p_t(x) F_t(x) d\lambda(x) \tag{58}$$

$$= C_t S_t + \int_{\mathcal{X}\setminus\mathcal{A}_{\text{shared}}} p_t(x) F_t(x) d\lambda(x) \tag{59}$$

$$\geq C_t S_t \geq S_t \tag{60}$$

where the last inequality uses $C_t \geq 1$ from Proposition 8. □

Proposition 9 (Exponential Decay of Outside Mass)**.** *Let $O_t := 1 - S_t = \int_{\mathcal{X}\setminus\mathcal{A}_{shared}} p_t d\lambda$. There exist constants $\gamma \in (0,1)$ and $u_0 > 0$ such that if $O_0 \leq u_0$, then:*

$$O_t \leq \gamma^t \tag{61}$$

for all $t \geq 0$.

Proof. Using the update equation and Lemma 5:

$$O_{t+1} = \int_{\mathcal{X}\setminus\mathcal{A}_{\text{shared}}} p_{t+1}(x) d\lambda(x) = \frac{1}{Z_t} \int_{\mathcal{X}\setminus\mathcal{A}_{\text{shared}}} p_t(x) F_t(x) d\lambda(x) \leq \frac{\rho O_t}{Z_t} \leq \frac{\rho O_t}{S_t} = \frac{\rho O_t}{1 - O_t} \tag{62}$$

Define $g(u) := \frac{\rho u}{1-u}$. For $u < \frac{\rho}{1+\rho}$, we have $g(u) < u$. Let $\gamma := \sup_{0 \leq u \leq u_0} \frac{g(u)}{u} < 1$ for some $u_0 < \frac{\rho}{1+\rho}$. Then for all $O_0 \leq u_0$:

$$O_{t+1} \leq \gamma O_t \tag{63}$$

Iterating gives $O_t \leq \gamma^t O_0 \leq \gamma^t$. □

Lemma 7 (Preservation of Density Ratios)**.** *For any $x, x' \in \mathcal{A}_{shared}$ and all $t \geq 0$:*

$$\frac{p_t(x)}{p_t(x')} = \frac{p_0(x)}{p_0(x')} \tag{64}$$

Proof. On $\mathcal{A}_{\text{shared}}$, the update simplifies to $p_{t+1}(x) = \frac{C_t}{Z_t} p_t(x)$. Define $\alpha_t := \frac{C_t}{Z_t}$. For any $x, x' \in \mathcal{A}_{\text{shared}}$:

$$\frac{p_{t+1}(x)}{p_{t+1}(x')} = \frac{\alpha_t p_t(x)}{\alpha_t p_t(x')} = \frac{p_t(x)}{p_t(x')} \tag{65}$$

By induction, this ratio is preserved from $t = 0$. □

Proof of Theorem 7. By Proposition 9, for any $\epsilon > 0$, since $\mathcal{A}_{\text{shared}}$ is closed in the compact space $\mathcal{X}$:

$$\int_{\mathcal{X}\setminus B_\epsilon(\mathcal{A}_{\text{shared}})} p_t(x) d\lambda(x) \leq O_t \leq \gamma^t = e^{t \log \gamma} \tag{66}$$

Setting $C = 1$ and $c = -\log \gamma > 0$ gives the exponential decay.

Since $O_t \to 0$, we have $S_t \to 1$. The sequence $\{p_t\}$ is tight (being probability measures on a compact space). Any weak limit point p_* must satisfy:

- $\text{supp}(p_*) \subseteq \mathcal{A}_{\text{shared}}$ (since $O_t \to 0$)
- On $\mathcal{A}_{\text{shared}}$, $p_*(x) \propto p_0(x)$ (by Lemma 7)

Since the limit is unique, $p_\infty := \lim_{t\to\infty} p_t$ exists and p_∞ is a probability measure supported on $\mathcal{A}_{\text{shared}}$ with $p_\infty(x) \propto p_0(x)$ for $x \in \mathcal{A}_{\text{shared}}$, normalization gives:

$$p_\infty(x) = \frac{p_0(x)}{\int_{\mathcal{A}_{\text{shared}}} p_0(z)d\lambda(z)} \mathbf{1}_{\mathcal{A}_{\text{shared}}}(x) \tag{67}$$

□

A.3 Disjoint Alignment

Theorem 10 (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 the public refinement gap:

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

Then with all constants already declared

$$v_* := p_0\big(B_\eta(x^\star)\big), \quad C_O := \frac{4}{v_*} + 1, \quad K_0 := \left\lceil \frac{8}{(\Delta_O - \delta)v_*} \right\rceil + 1, \quad 0 < \delta < \tfrac{1}{2}\Delta_O.$$

we have:

1. ***Exponential decay outside owner optima:*** *For any $\varepsilon > 0$, there exist constants $C, c > 0$ such that*
$$p_t(X \setminus B_\varepsilon(A_O)) \le Ce^{-ct}$$
2. ***Exponential decay within owner optima but outside public refinement:*** *For any $\varepsilon > 0$, there exist constants $C', c' > 0$ such that*
$$p_t(A_O \setminus B_\varepsilon(A_{P|O})) \le C'e^{-c't}$$
3. ***Convergence to public-refined equilibrium:*** *The limit $p_\infty = \lim_{t\to\infty} p_t$ exists and is supported on $A_{P|O}$. Moreover, for all $x \in A_{P|O}$:*
$$p_\infty(x) = \frac{p_0(x)}{\int_{A_{P|O}} p_0(z)dz}$$

We establish this result through a series of lemmas that characterize the two-stage suppression mechanism.

Lemma 8 (Constant Weights Within Owner Optima)**.** *For all $t \ge 0$ and all $x, x' \in A_O$:*

$$H^{p_t}_{K,r_O}(x) = H^{p_t}_{K,r_O}(x')$$

Proof. Since all elements of A_O share the same maximal owner reward, the Bradley-Terry weights depend only on the distribution of competitors, not on which specific maximizer is chosen. Thus $H^{p_t}_{K,r_O}(x) = W^O_t$ for some constant W^O_t on A_O. □

Lemma 9 (Public Stage Refinement Within A_O)**.** *Define $\sigma := M[1 + (M-1)e^{-\Delta_{P|O}}]^{-1} < 1$. Then for all $x \in A_O \setminus A_{P|O}$:*

$$H^{\tilde{p}_t}_{M,r_P}(x) \leq \sigma$$

while for all $y \in A_{P|O}$:

$$H^{\tilde{p}_t}_{M,r_P}(y) \geq 1$$

Proof. For $x \in A_O \setminus A_{P|O}$, we have $r_P(x) \leq \max_{y \in A_{P|O}} r_P(y) - \Delta_{P|O}$. Thus for any competitor z:

$$r_P(z) - r_P(x) \geq -\Delta_{P|O} \implies \exp(r_P(z) - r_P(x)) \geq e^{-\Delta_{P|O}} \tag{68}$$

Therefore:

$$H^{\tilde{p}_t}_{M,r_P}(x) = \mathbb{E}_{Y_1,\ldots,Y_{M-1} \sim \tilde{p}_t}\left[\frac{M}{1 + \sum_{j=1}^{M-1} \exp(r_P(Y_j) - r_P(x))}\right] \leq \sigma \tag{69}$$

For $y \in A_{P|O}$, since $r_P(y) = \max_{z \in A_O} r_P(z)$, the weight is at least 1. □

Lemma 10 (Two-Stage Contraction Factors)**.** *Define the combined weight $F_t(x) := H^{p_t}_{K,r_O}(x) \cdot H^{\tilde{p}_t}_{M,r_P}(x)$. Then:*

- *For $x \notin A_O$: $F_t(x) \leq C_O\, e^{-(\Delta_O - \frac{3}{2}\delta)}$*
- *For $x \in A_O \setminus A_{P|O}$: $F_t(x) = W^O_t \sigma$*
- *For $x \in A_{P|O}$: $F_t(x) = W^O_t$*

where $W^O_t \geq 1$ is the constant from Lemma 8.

Proof. Combines Lemmas 1, 8, and 9. □

Proposition 11 (Exponential Decay of Outside Mass)**.** *Let $m_t := p_t(X \setminus A_O)$. There exist constants $C'_0 > 0$ and $\rho \in (0,1)$ such that:*

$$m_t \leq C'_0 \rho^t \tag{70}$$

Proof. Using the update equation and Lemma 10:

$$m_{t+1} = \int_{X \setminus A_O} p_{t+1}(x)dx = \frac{1}{Z_t}\int_{X \setminus A_O} p_t(x)F_t(x)dx \leq \frac{C_O\, e^{-(\Delta_O - \frac{3}{2}\delta)}}{Z_t} m_t \tag{71}$$

Since $Z_t \geq \int_{A_{P|O}} p_t(x)F_t(x)dx = W^O_t p_t(A_{P|O}) \geq W^O_t(1 - m_t - b_t)$ where $b_t := p_t(A_O \setminus A_{P|O})$, and using $W^O_t \geq 1$, we obtain the contraction $m_{t+1} \leq \rho m_t$ with $\rho = C_O\, e^{-(\Delta_O - \frac{3}{2}\delta)} < 1$. □

Proposition 12 (Exponential Decay Within Owner Optima). *Let $b_t := p_t(A_O \setminus A_{P|O})$. Then:*

$$b_t \leq b_0 \sigma^t$$

Proof. Using the decomposition $A_O = A_{P|O} \cup (A_O \setminus A_{P|O})$ and Lemma 10:

$$b_{t+1} = \frac{1}{Z_t} \int_{A_O \setminus A_{P|O}} p_t(x) F_t(x) dx = \frac{W_t^O \sigma}{Z_t} b_t \tag{72}$$

Since $Z_t \geq W_t^O p_t(A_{P|O}) = W_t^O (1 - m_t - b_t) \geq W_t^O (1 - b_t)$ for large t (as $m_t \to 0$), we have:

$$b_{t+1} \leq \frac{W_t^O \sigma}{W_t^O (1 - b_t)} b_t = \frac{\sigma b_t}{1 - b_t} \tag{73}$$

Following the analysis in Proposition 5, this yields $b_t \leq b_0 \sigma^t$. □

Lemma 11 (Preservation of Density Ratios on $A_{P|O}$). *For any $x, x' \in A_{P|O}$ and all $t \geq 0$:*

$$\frac{p_t(x)}{p_t(x')} = \frac{p_0(x)}{p_0(x')}$$

Proof. On $A_{P|O}$, both stages apply constant weights: $F_t(x) = W_t^O$ for all $x \in A_{P|O}$. Thus:

$$\frac{p_{t+1}(x)}{p_{t+1}(x')} = \frac{p_t(x) F_t(x) / Z_t}{p_t(x') F_t(x') / Z_t} = \frac{p_t(x)}{p_t(x')}$$

The result follows by induction. □

Proof of Theorem 10. By Proposition 11, for any $\varepsilon > 0$:

$$p_t(X \setminus B_\varepsilon(A_O)) \leq m_t \leq C_0 \rho^t = C e^{-ct} \tag{74}$$

with $C = C_0$ and $c = -\log \rho > 0$.

By Proposition 12, for any $\varepsilon > 0$:

$$p_t(A_O \setminus B_\varepsilon(A_{P|O})) \leq b_t \leq b_0 \sigma^t = C' e^{-c' t} \tag{75}$$

with $C' = b_0$ and $c' = -\log \sigma > 0$.

Since $m_t + b_t \to 0$, we have $p_t(A_{P|O}) \to 1$. The sequence $\{p_t\}$ is tight. Any weak limit point p_* must satisfy:

- $\text{supp}(p_*) \subseteq A_{P|O}$ (by Parts 1-2)
- On $A_{P|O}$: $p_*(x) \propto p_0(x)$ (by Lemma 11)

Since the limit is unique, $p_\infty = \lim_{t \to \infty} p_t$ exists and:

$$p_\infty(x) = \frac{p_0(x)}{\int_{A_{P|O}} p_0(z) dz} \mathbf{1}_{A_{P|O}}(x) \tag{76}$$

□

A.4 Fundamental Alignment Trilemma

Theorem 13 (Fundamental Alignment Trilemma)**.** *Let $(p_t)_{t\geq 0}$ be the sequence of model output distributions generated by the two-curator loop on a compact space $\mathcal{X}$ with continuous rewards $r_O, r_P \colon \mathcal{X} \to \mathbb{R}$ and sufficiently large pool sizes K, M. Define the following desirable properties:*

***(i) Value Alignment**: The weak limit p_∞ assigns positive probability to at least one maximizer of r_O and to at least one maximizer of r_P.*

***(ii) Diversity**: p_∞ has strictly positive Shannon entropy $H(p_\infty) = -\int_{\mathcal{X}} p_\infty(x) \log p_\infty(x)\, dx > 0$.*

***(iii) Stability**: The sequence (p_t) converges to a* unique *limit that is independent of the initial distribution p_0.*

Then no *recursive curation system can satisfy all three properties simultaneously; at most two can hold for any given alignment regime.*

Proof. Let $A_O = \arg\max r_O$ and $A_P = \arg\max r_P$.

Case 1: Perfect alignment ($A_O = A_P = \{x^*\}$). By Theorem 5, the process collapses exponentially to the point mass δ_{x^*}. This satisfies Value Alignment and Stability, but the entropy is zero, so Diversity fails.

Case 2: Partial alignment ($A_O \cap A_P \neq \emptyset$ and $A_O \neq A_P$). Theorem 7 shows that p_t converges to a measure supported on $A_{\text{shared}} = A_O \cap A_P$ with weights proportional to p_0. Hence Value Alignment and Diversity hold, but the limit depends on p_0, so Stability fails.

Case 3: Disjoint alignment ($A_O \cap A_P = \varnothing$). Theorem 10 states that p_t converges to a distribution supported on the public-refined owner set $A_{P|O} \subseteq A_O$, with density proportional to p_0. Consequently Value Alignment fails because $p_\infty(A_P) = 0$, and Stability fails because the limit still depends on p_0. Diversity can hold (provided $|A_{P|O}| > 1$), but at least one of the other two properties is violated.

In each exhaustive regime at least one property is violated, so the three properties cannot hold simultaneously. Therefore, any recursive two-curator curation system must sacrifice at least one of Value Alignment, Diversity, or Stability. □