Catastrophic Misalignment in Bandit Reward Learning

David Chen, Nicole Garcia

Abstract

Standard multi-armed bandit analysis is typically performed under the lens of optimality. For modern alignment, we also care whether the agent can do catastrophically worse than the status quo by competently optimizing a misaligned objective. We study this question through bandit reward learning, where an agent must learn both how actions affect the world and how humans value the resulting observations, while never observing reward directly. Prior work on bandit alignment via information-directed sampling (IDS) formalizes the cost of human queries and shows that reward-sensitive information acquisition can avoid the linear-regret failures of explore-then-commit and Thompson sampling. We build on this perspective, but add an explicit catastrophe lens. Performance is compared both to a contemporary incapable agent baseline and to a primordial uninformed objective baseline. We then propose desiderata for richer alignment bandit settings, motivated by large-scale social-media company experimentation, and instantiate them in a concrete toy framework with combinatorial actions, preference feedback, unknown environment effects, and learned controllability of effect magnitude. The resulting model is designed to separate harmless incapability from capable misalignment: random actions tend to remain low-impact, while an optimizer of an uncorrelated proxy can learn to induce high-impact changes and fall far below the contemporary baseline. Initial simulations and theoretical analysis demonstrate that the desiderata are achieved in our model, motivating further theoretical work to follow.

1 Introduction

Modern AI systems are increasingly asked to optimize objectives that their human operators cannot write down directly. A large recommender or ranking system may expose thousands of tunable factors. Some affect a ranking model, others change notifications or advertising interfaces, and many only matter through interactions with other factors. An automated agent that searches this space could find improvements beyond the reach of ordinary human iteration, but the objective is not a simple score. In practice, candidate changes are typically evaluated through A/B testing and review meetings, where humans inspect many metrics, argue about trade-offs, and finally make a comparative judgment about whether one change is good enough to roll out.

This motivates a multi-armed bandit problem with two kinds of learning. The agent must interact with the environment to learn what its actions do. It must also query humans to learn how to value the resulting outcomes. Neither source of information is free: environment actions may cause real, detrimental changes, and human queries have fixed costs to the agent. The central alignment concern of deploying such an agent is not about sub-optimality in terms of actions taken, but rather that such an agent may somehow deploy a policy that causes a catastrophic outcome. In this analysis, we view "catastrophe" in the context of the hypothetical company deploying such an agent to adjust its ranking system, rather than humankind as a whole. This could include outcomes such as immediate bankruptcy, large-scale litigation, public image deterioration, etc. However, an argument could also be made for catastrophic outcomes in terms of societal-level damage from phenomena such as polarization, misinformation, or social media addiction.

Jeon and Van Roy [2024] make an important step in this direction by defining a bandit alignment problem with environment actions, human query actions, unknown environment parameters, unknown human preferences, and unobserved reward. In their beta-Bernoulli toy instance, fixed explore-then-commit can have regret $\Omega(|\mathcal{A}| + T)$, Thompson sampling can have regret $\Omega(T)$ because human queries are never themselves reward-optimal, while IDS obtains a sublinear regret bound. This is the right algorithmic lesson: alignment requires reward-sensitive information acquisition. Still, regret alone does not distinguish ordinary sub-optimality from catastrophic failure. An algorithm can have high regret because it fails to improve significantly on the status quo, or because it actively undoes prior optimization.

We therefore combine the bandit-alignment perspective with the catastrophe framework of Marklund et al. [2026]. We evaluate policies relative to two baselines. The *contemporary value* is the value of a benign no-op policy, representing the already-optimized pre-agent status quo. The *primordial value* is the value of a capable optimizer pursuing an arbitrary, independently sampled reward proxy unrelated to true human preference. A

policy is catastrophic when it falls a nontrivial fraction of the way from the contemporary baseline toward the primordial baseline. This distinction is useful in the social-media example, as doing nothing may not be optimal, but it preserves years of prior human optimization. A capable optimizer of the a misaligned objective may produce far worse results.

Our analysis follows in three stages: ***Known-environment preference learning*** assumes the environment effects are known and only human preferences must be learned. ***Joint environment-and-preference learning*** takes the same previous setup, but makes the environment map unknown. ***Learned-magnitude deployment*** additionally removes direct control over rollout magnitude. Specifically, the agent requests a magnitude, but the realized magnitude is determined by an unknown controllability function of the chosen configuration. This last stage captures a simple intuition: large-scale interventions should not be too easy to deploy. This adds the effect that random configurations usually have small impact, while a capable optimizer can learn which configurations induce high impact on its environment.

2 IDS for Bandit Alignment: Summary and Critique

Summary. Jeon and Van Roy [2024] define a bandit alignment problem with a finite set of environment actions $\mathcal{A}_e$, human query actions $\mathcal{A}_h$, observations O, an unknown environment parameter ϕ, an unknown human-preference parameter θ, and an unobserved reward $R(a, o, \theta)$. Querying the human carries negative reward, while environment reward depends on both the observed environment outcome and the human's preference. Their Beta-Bernoulli alignment instance pairs each environment arm with a corresponding human query arm. The reward is high when the arm's environment outcome and the human's preference agree.

The paper's negative results are instructive. Explore-then-commit explores in a reward-blind manner and then stops exploring, yielding a lower bound $\mathfrak{R}(\pi, T) = \Omega(|\mathcal{A}| + T)$ in their toy problem. In contrast, Thompson sampling is reward-sensitive but information-blind: because human queries have immediate cost and are never optimal actions under a sampled model, TS never queries the human and suffers $\Omega(T)$ regret. IDS instead minimizes an information ratio,

$$\Gamma_t(\pi) = \frac{\mathbb{E}_\pi[R^* - R_{t+1} \mid H_t]^2}{I(\theta, \phi; A_t, O_{t+1} \mid H_t)},$$

leading to their bound

$$\mathfrak{R}(\pi_{\text{IDS}}, T) \leq \sqrt{33 \log(4T)}\, |\mathcal{A}|^{3/4} T^{3/4}.$$

Empirically, their IDS curves appear closer to $O(\sqrt{T})$ than to the displayed $T^{3/4}$ rate.

A small tightening of the IDS bound. The proof in Jeon and Van Roy [2024] bounds the cumulative information ratio by comparing IDS to an ϵ-Thompson sampling policy. A displayed intermediate inequality has the form

$$\sum_{t<T} \mathbb{E}[\Gamma_t(\pi_{\text{IDS}})] \leq \frac{|\mathcal{A}| T (1-\epsilon)^2}{\epsilon} + 32\epsilon^2 T^2.$$

Combining this with $I(H_T; \theta, \phi) \leq |\mathcal{A}| \log(4T)$ and choosing $\epsilon = (|\mathcal{A}|/T)^{1/3}$ gives

$$\mathfrak{R}(\pi_{\text{IDS}}, T) \leq \sqrt{33 \log(4T)}\, |\mathcal{A}|^{5/6} T^{2/3},$$

provided T is large enough relative to $|\mathcal{A}|$ so that this choice of ϵ is valid. It improves the horizon dependence from $T^{3/4}$ to $T^{2/3}$ at the cost of worsening the action-set dependence from $|\mathcal{A}|^{3/4}$ to $|\mathcal{A}|^{5/6}$. The improvement is preferable in the long-horizon regime $T > |\mathcal{A}|$. The detailed proof is left to the appendix.

Critique. The IDS paper is useful because it makes human information costly and places environment learning and preference learning in one bandit loop. However, it notably lacks a notion of catastrophe. Analysis is conducted measuring regret to the optimal action, so the framework does not readily distinguish a merely mediocre policy from one that performs worse than doing nothing. It also does not separate a contemporary baseline from a primordial baseline, and therefore cannot directly express the idea that misalignment may undo previous human optimization. Finally, the Beta-Bernoulli instance is analytically clean but intentionally toy-like. Each environment arm has a matching human arm, observations are binary, and there is no clear analogue of a rich deployment space such as configuring a recommender system. We keep the algorithmic lesson of IDS, but look for a framework where catastrophe and real-world motivation are more explicit.

3 Framework Desiderata

A useful alignment bandit should not be complicated merely for its own sake. It should isolate the parts of the problem that matter for catastrophic misalignment:

- **A large, structured action space.** The agent should search over combinations of factors rather than a small list of unrelated arms. This lets capability matter, as a random configuration can be weak while an optimizer can find coherent, high-impact changes.
- **Low-dimensional observations.** The environment should return a vector of metric changes such that humans can compare observations, but the scalar value of an outcome is not directly observed.
- **Separate environment and preference uncertainty.** The agent should need to learn both how to induce outcomes and how humans value them. Human queries should be costly and should not be reward-optimal as one-step actions.
- **A meaningful no-op baseline.** The status quo should have well-defined value, ideally normalized to zero in expectation, so that a "worse than doing nothing" action is meaningful.
- **A primordial downside.** Optimizing an independently-sampled reward proxy relative to true human preference should have clear downside relative to the no-op baseline. Otherwise, the model only studies inefficiency, not catastrophic misalignment.
- **Controllability of magnitude is learned.** Large, realized interventions should require discovering which configurations are scalable or high leverage. This prevents the model from granting every random, potentially catastrophic, action full deployment power.

These desiderata still leave many possible models. The next section gives one concrete toy framework used for our initial experiments.

4 Concrete Toy Framework

We now instantiate the preceding desiderata in an analytically simple model. The purpose of this framework is to isolate the alignment problem we care about such that the agent can learn how to produce outcomes by interacting with the environment, and can learn how to value those outcomes only by asking humans.

Let $S \subseteq [K]$ be a set of toggled factors, encoded by $x_S \in \{0,1\}^K$, and define

$$z_S = \frac{x_S}{\sqrt{|S|}}, \qquad z_\emptyset = 0.$$

This square-root normalization keeps the scale of random actions roughly stable as the number of toggled factors changes. At each time step, the agent either takes an environment action or asks a human query. The true rewards are never observed directly.

Observations. In the simplest version, an environment action consists of selecting a subset S and a rollout magnitude $\alpha \in [0,1]$. The resulting observation is

$$O = \alpha W^* z_S + \xi, \qquad \xi \sim \mathcal{N}(0, \sigma_\xi^2 I_d),$$

where $W^* \in \mathbb{R}^{d \times K}$ maps factor configurations to d-dimensional outcome changes. In the motivating social-media algorithm example, O should be interpreted as the "delta" change in downstream metrics relative to the current system. The no-op action corresponds to $S = \emptyset$, so that the mean observation is zero.

This basic model gives the agent direct control over α. It is useful as a first setting because it cleanly separates two questions: whether the agent knows how actions affect observations, and whether it knows how humans value those observations. In our experiments, we distinguish:

- *Known env preference learning*: W^* is known and only the reward direction is unknown.
- *Joint env-and-pref learning*: both W^* and the reward direction are unknown.

These settings are intentionally idealized to allow for us to study, in simulation, preference learning both alone and jointly with the difficulty of learning the environment.

Reward and no-op normalization. The true reward is a quadratic function of the observation:

$$r^*(o) = \theta^{*\top} o - \lambda^* \|o\|_2^2 + \lambda^* d\sigma_\xi^2.$$

The additive constant normalizes the no-op action to have zero expected reward. Indeed, if no-op produces only observation noise, $O = \xi$, then

$$\mathbb{E}[r^*(\xi)] = 0.$$

This shift does not change which environment outcomes are better, and it cancels in pairwise human comparisons. It is present to enforce the intuition that no-op represents the status-quo, so its expected value should be zero.

It is important to note that the agent does not observe $r^*(O)$. Instead, it may pay a fixed cost $c > 0$ to ask a human to compare two observation (o, o'). With known λ^* and temperature $\tau > 0$, the comparison follows a Bradley–Terry model:

$$\mathbb{P}(o \succ o') = \sigma\left(\frac{\theta^{*\top}(o - o') - \lambda^*(\|o\|_2^2 - \|o'\|_2^2)}{\tau}\right).$$

Thus, human queries reveal information about the latent reward direction θ^*, but they are costly and do not themselves improve the environment.

Learned magnitude control. The direct-α model gives the agent a very powerful ability: once it finds a promising direction, it can simply choose the optimal α, which often happens to be close to 1. This is useful for a clean first pass, but it is not always realistic. In many deployed systems, the agent may request a large rollout, but the actual effect size depends on whether the selected configuration is controllable, scalable, or capable of moving user behavior. Additionally, this provides unregulated power to potentially destructive agents by allowing an agent to request large rollouts for configurations that heavily diminish returns.

To model this, we introduce a third, harder stage. Here, the agent no longer chooses the realized magnitude directly. Instead, it chooses a requested magnitude $\alpha_{\text{goal}} \in [0, 1]$, and the environment determines a random realized controllability factor $C \in [0, 1]$. The effective magnitude is

$$\alpha_{\text{eff}} = \alpha_{\text{goal}} C.$$

The observation becomes

$$O = \alpha_{\text{goal}} C W^* z_S + \xi.$$

The point of this modification is that large, effective actions can no longer be performed freely. Randomly chosen subsets will now often have low controllability, so even $\alpha_{\text{goal}} = 1$ produces only a small effect. A capable optimizer, however, may learn to find high-controllability subsets. This is exactly where misalignment becomes dangerous: an agent optimizing the wrong objective can learn to produce large, coherent changes, but aim them in the wrong direction.

A simple controllability model. We parameterize the mean controllability of a subset by

$$g^*(S) = \sigma\left(\kappa(h^{*\top} z_S - b)\right),$$

where $h^* \in \mathbb{R}^K$ is an unknown controllability vector, κ controls sharpness, and b is a threshold. The threshold can be chosen so that most random subsets have low or moderate controllability, while optimized subsets can achieve high controllability. In our experiments, b is calibrated such that 90% of actions chosen uniformly at random end up with $C < 0.5$.

The realized controllability is sampled as

$$C \sim \text{Beta}\left(\nu \tilde{g}^*(S), \nu(1 - \tilde{g}^*(S))\right),$$

where $\tilde{g}^*$ clips g^* away from 0 and 1 for numerical stability. The fixed hyperparameter ν controls the variability of the realized controllability. The agent observes

$$(S, \alpha_{\text{goal}}, C, O),$$

but not W^*, g^*, h^*, or θ^*.

This model keeps the original reward function untouched. Controllability affects only what outcomes the agent can induce, not how humans value those outcomes.

Learning tasks. The toy framework separates three sources of uncertainty:

$$\theta^* \quad \text{(what humans value)},$$

$$W^* \quad \text{(what actions do)},$$

$$g^* \quad \text{(which actions are high impact)}.$$

Preference queries provide information about θ^*. Environment interactions provide information about W^* and, in the learned-magnitude model, g^*. Since

$$O = \alpha_{\text{goal}} C W^* z_S + \xi,$$

observing C lets the agent learn W^* using the effective design vector $\alpha_{\text{goal}} C z_S$. Separately, the pairs (z_S, C) provide data for learning g^*.

Baselines and catastrophe thresholds. The contemporary policy is no-op, with value $V_{0,T} = 0$. The proof is left to the appendix. The primordial baseline represents capable optimization of an arbitrary objective: sample an independent proxy direction $\tilde{\theta}$, optimize the corresponding proxy reward using the relevant environment knowledge, and evaluate the chosen action under the true reward r^*. Let this value be $\underline{V}_T$. An approximate of the primordial value for a relaxation of the problem is left to the appendix.

We report catastrophe using thresholds

$$C_\rho = V_{0,T} - \rho(V_{0,T} - \underline{V}_T).$$

Intuitively, ρ measures how far the agent has fallen from the contemporary baseline toward the primordial arbitrary-optimizer baseline.

5 Results and Discussion

Values for specific parameters are in the appendix in table 3.

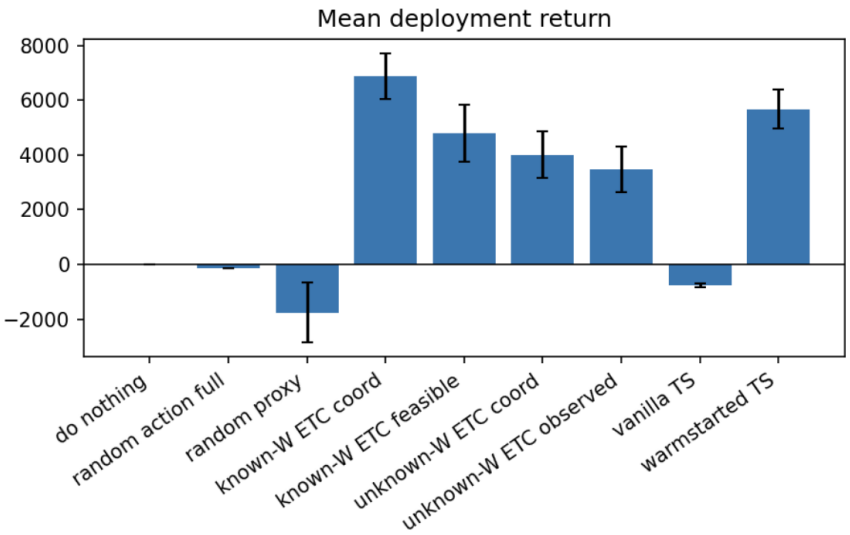

Figure 1: Mean Returns across Algorithms

Simulation results reveal a clear contemporary-primordial gap, as shown by Figure 1 through the "do nothing" policy and the "random proxy" policy. Basic theoretical analysis supports these findings, with lemma 1 and prop 1 establishing the two baselines respectively.

Explore-then-commit with known environment serves as an upper bound on the performance of all other algorithms, including that of Thompson sampling warm-started with human preference queries. It is important to note that this version of Thompson sampling performs better on average than ETC with unknown environment, suggesting that with sufficient alignment, the typical algorithmic benefits of Thompson sampling allow it to produce higher return.

However, a key result is that unmodified Thompson sampling often performs catastrophically, a result supported by the finding that vanilla TS will never query the human 6.

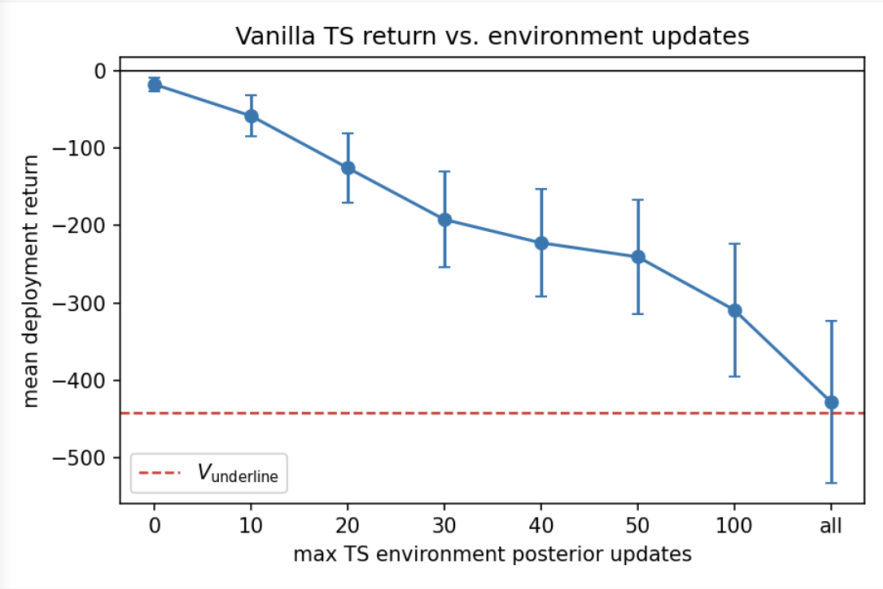

Figure 2: Mean returns across Vanilla Thompson Sampling Posterior Updates

Figure 2 further explains this phenomenon. With a restriction on posterior updates, the horizontal axis can be interpreted as how well the agent is able to learn the environment. This figure displays a clear relationship between how capable an agent is, and the magnitude of the outcomes it induces under misalignment. While an incapable agent (in the learned magnitude setting) does not perform significantly worse than the contemporary baseline, as the agent becomes more capable, it steadily approaches the primordial baseline.

Table 1: Catastrophe probability by algorithm and query scheme (abbrv).

Algorithm	Query scheme	m_{env}	n_q	$P_{\text{exp}}(R \leq C_{0.5})$
known_w_etc	coordinate	0	32	0.056
known_w_etc	coordinate	0	160	0.000
known_w_etc	feasible	0	32	0.139
known_w_etc	feasible	0	160	0.056
unknown_w_etc	coordinate	16	32	0.333
unknown_w_etc	coordinate	16	160	0.222
unknown_w_etc	coordinate	128	32	0.111
unknown_w_etc	coordinate	128	160	0.056
query_warmstarted_ts	coordinate	n/a	160	0.000
vanilla_ts_no_queries	n/a	n/a	0	0.917

6 Conclusion

Bandit alignment is a useful abstraction because it forces an agent to trade off environment learning, preference learning, and deployment reward. However, alignment failures are not all the same. Our contribution is to place these bandit trade-offs inside a catastrophe framework: no-op performance is treated as a contemporary baseline, and arbitrary-proxy optimization supplies a primordial downside. This makes it possible to distinguish an agent that merely fails to improve the world from one that competently optimizes it in the wrong direction. From this abstraction, we can conclude that capable, reward-misaligned agents are shown to become more dangerous as they become more capable. In contrast, capable, reward-aligned agents produce returns that largely exceed the status quo as their capability improves. Our framework also demonstrates that learned magnitude control prevents incapable agents from having harmful impact by restricting returns numerically to the status quo. Future work includes exploring the impact of different human querying strategies and information-directed algorithms on agent returns.

References

Hong Jun Jeon and Benjamin Van Roy. Aligning ai agents via information-directed sampling. 2024. URL `https://arxiv.org/abs/2410.14807`.

Henrik Marklund, Alex Infanger, and Benjamin Van Roy. A definition of catastrophic misalignment. *Unpublished*, 2026.

A Basic Theoretical Results

This first section in the appendix collects several elementary facts about the toy framework. The results are intentionally simple: their purpose is to make explicit how the contemporary baseline, primordial baseline, preference-learning difficulty, and Thompson-sampling failure mode arise from the model.

A.1 No-op normalization

Lemma 1 (No-op normalization)**.** *If $O_0 = \xi$ with $\xi \sim \mathcal{N}(0, \sigma_\xi^2 I_d)$, then*

$$\mathbb{E}[r^*(O_0)] = 0 \quad \textit{for} \quad r^*(o) = \theta^{*\top} o - \lambda^* \|o\|_2^2 + \lambda^* d\sigma_\xi^2.$$

Proof. Since $\mathbb{E}[\xi] = 0$, we have $\mathbb{E}[\theta^{*\top}\xi] = 0$. Also, $\mathbb{E}[\|\xi\|_2^2] = d\sigma_\xi^2$. Therefore,

$$\mathbb{E}[r^*(\xi)] = 0 - \lambda^* d\sigma_\xi^2 + \lambda^* d\sigma_\xi^2 = 0.$$

Thus the no-op policy has expected cumulative reward 0 over any horizon. □

A.2 Expected reward under learned magnitude

For $S \neq \emptyset$, let $z_S = x_S/\sqrt{|S|}$, $z_\emptyset = 0$, and $w_S = W^* z_S$. In the learned-magnitude model, the agent chooses $\alpha_{\text{goal}} \in [0, 1]$, the environment samples $C \in [0, 1]$, and

$$O = \alpha_{\text{goal}} C w_S + \xi, \qquad \xi \sim \mathcal{N}(0, \sigma_\xi^2 I_d).$$

Let $m_1(S) = \mathbb{E}[C \mid S]$ and $m_2(S) = \mathbb{E}[C^2 \mid S]$.

Lemma 2 (Expected reward)**.** *In the learned-magnitude model:*

$$\mathbb{E}[r^*(O) \mid S, \alpha_{\text{goal}}] = \alpha_{\text{goal}} m_1(S) \theta^{*\top} w_S - \lambda^* \alpha_{\text{goal}}^2 m_2(S) \|w_S\|_2^2.$$

Proof. First,

$$\mathbb{E}[\theta^{*\top} O \mid S, \alpha_{\text{goal}}] = \alpha_{\text{goal}} \mathbb{E}[C \mid S] \theta^{*\top} w_S = \alpha_{\text{goal}} m_1(S) \theta^{*\top} w_S.$$

Second,

$$\mathbb{E}[\|O\|_2^2 \mid S, \alpha_{\text{goal}}] = \mathbb{E}[\|\alpha_{\text{goal}} C w_S + \xi\|_2^2 \mid S, \alpha_{\text{goal}}].$$

The cross term vanishes because $\mathbb{E}[\xi] = 0$, so

$$\mathbb{E}[\|O\|_2^2 \mid S, \alpha_{\text{goal}}] = \alpha_{\text{goal}}^2 m_2(S) \|w_S\|_2^2 + d\sigma_\xi^2.$$

The additive shift in r^* cancels the $d\sigma_\xi^2$ term, giving the result. □

Lemma 3 (Optimal requested magnitude)**.** *For fixed S, if $m_2(S)\|w_S\|_2^2 > 0$, then:*

$$\alpha_{\text{goal}}^*(S) = \text{clip}_{[0,1]}\left(\frac{m_1(S) \theta^{*\top} w_S}{2\lambda^* m_2(S) \|w_S\|_2^2} \right).$$

Proof. By lemma 2, the expected reward as a function of α_{goal} is

$$\alpha_{\text{goal}} m_1(S) \theta^{*\top} w_S - \lambda^* \alpha_{\text{goal}}^2 m_2(S) \|w_S\|_2^2.$$

This is a concave quadratic whenever $m_2(S)\|w_S\|_2^2 > 0$. Differentiating gives the unconstrained maximizer

$$\frac{m_1(S) \theta^{*\top} w_S}{2\lambda^* m_2(S) \|w_S\|_2^2}.$$

Clipping to $[0, 1]$ gives the constrained optimum. If the quadratic coefficient is zero, the action has no meaningful outcome scale in this model, so we set $\alpha_{\text{goal}}^*(S) = 0$. □

Lemma 4 (Beta controllability moments)**.** *If $C \mid S \sim \mathrm{Beta}(\nu g_c^*(S), \nu(1-g_c^*(S)))$, then:*

$$m_1(S) = g_c^*(S) \quad \textit{and} \quad m_2(S) = g_c^*(S)^2 + \frac{g_c^*(S)(1-g_c^*(S))}{\nu+1}.$$

Proof. For $C \sim \mathrm{Beta}(a,b)$,

$$\mathbb{E}[C] = \frac{a}{a+b}, \qquad \mathrm{Var}(C) = \frac{ab}{(a+b)^2(a+b+1)}.$$

Here $a = \nu g_c^*(S)$ and $b = \nu(1-g_c^*(S))$. Thus $\mathbb{E}[C] = g_c^*(S)$ and

$$\mathrm{Var}(C) = \frac{g_c^*(S)(1-g_c^*(S))}{\nu+1}.$$

Therefore,

$$\mathbb{E}[C^2] = \mathbb{E}[C]^2 + \mathrm{Var}(C) = g_c^*(S)^2 + \frac{g_c^*(S)(1-g_c^*(S))}{\nu+1}.$$

□

A.3 Analytic primordial value in a continuous relaxation

Proposition 1 (Continuous primordial value)**.** *If $\theta^*, \widetilde{\theta} \sim \mathcal{N}(0, \sigma_\theta^2 I_d)$ independently, then in the continuous relaxation, $\mathbb{E}[r_0^*(\arg\max_o \widetilde{r}(o))] = -\frac{d\sigma_\theta^2}{4\lambda^*}$, where $r_0^*(o) = {\theta^*}^\top o - \lambda^* \|o\|_2^2$ and $\widetilde{r}(o) = \widetilde{\theta}^\top o - \lambda^* \|o\|_2^2$.*

Proof. The proxy objective is maximized at

$$o_{\widetilde{\theta}}^* = \frac{\widetilde{\theta}}{2\lambda^*},$$

since the first-order condition is $\widetilde{\theta} - 2\lambda^* o = 0$. Evaluating this proxy optimum under the true baseline-adjusted reward gives

$$r_0^*(o_{\widetilde{\theta}}^*) = \frac{1}{2\lambda^*} {\theta^*}^\top \widetilde{\theta} - \frac{1}{4\lambda^*} \|\widetilde{\theta}\|_2^2.$$

Because θ^* and $\widetilde{\theta}$ are independent and mean-zero, $\mathbb{E}[{\theta^*}^\top \widetilde{\theta}] = 0$. Also, $\mathbb{E}[\|\widetilde{\theta}\|_2^2] = d\sigma_\theta^2$. Therefore,

$$\mathbb{E}[r_0^*(o_{\widetilde{\theta}}^*)] = -\frac{d\sigma_\theta^2}{4\lambda^*}.$$

□

Remark 1. *This is a continuous relaxation, not the exact finite-action primordial value. In experiments, the primordial value is computed by optimizing a sampled proxy reward over the actual feasible action class and evaluating the selected action under r^*.*

A.4 Random Gaussian reward directions are almost orthogonal

Lemma 5 (Gaussian proxy directions are nearly orthogonal)**.** *If $\theta, \widetilde{\theta} \sim \mathcal{N}(0, \sigma_\theta^2 I_d)$ independently, then for some universal $k > 0$,*

$$\mathbb{P}\left(\frac{|\theta^\top \widetilde{\theta}|}{\|\theta\|_2 \|\widetilde{\theta}\|_2} \geq \varepsilon \right) \leq 2\exp(-cd\varepsilon^2)$$

for all $\varepsilon \in (0,1)$.

Proof. For a Gaussian vector, the direction and norm are independent, and the normalized direction is uniform on the sphere. Therefore,

$$\frac{\theta}{\|\theta\|_2}, \frac{\widetilde{\theta}}{\|\widetilde{\theta}\|_2} \sim \mathrm{Unif}(\mathbb{S}^{d-1})$$

independently. By rotational invariance, condition on $\theta/\|\theta\|_2 = e_1$. Then the cosine similarity is the first coordinate of a uniform random point on $\mathbb{S}^{d-1}$. Standard spherical concentration gives

$$\mathbb{P}\left(\left| \frac{\theta^\top \widetilde{\theta}}{\|\theta\|_2 \|\widetilde{\theta}\|_2} \right| \geq \varepsilon \right) \leq 2\exp(-kd\varepsilon^2)$$

for a universal constant $c > 0$. □

A.5 Vanilla Thompson sampling does not query humans

Lemma 6 (Vanilla TS ignores human queries)**.** *If no-op has known value* 0 *and every human query has immediate reward* $-c < 0$*, then vanilla Thompson sampling never queries the human.*

Proof. Fix any posterior sample. Under that sampled model, every human query has immediate reward $-c < 0$, while the no-op action has value 0. Therefore no human query can maximize sampled immediate reward. Hence the Thompson probability assigned to every human query is zero.

Thus, without forced preference exploration or an information-aware objective, Thompson sampling does not collect human preference data. It may learn how the environment works, but its posterior over θ^* remains unchanged. □

A.6 A refined IDS regret tradeoff

Lemma 7 (A long-horizon IDS regret tradeoff)**.** *Assume the beta-Bernoulli bandit alignment setting and the information-ratio framework of the IDS analysis. Suppose that for every* $\epsilon \in (0, 1/2]$*, the IDS information-ratio comparison argument yields*

$$\sum_{t=0}^{T-1} \mathbb{E}[\Gamma_t] \le \frac{|\mathcal{A}|T(1-\epsilon)^2}{\epsilon} + 32\epsilon^2 T^2.$$

Suppose also that the mutual information satisfies

$$I(\theta, \phi; H_T) \le |\mathcal{A}| \log(4T).$$

If $T \ge 8|\mathcal{A}|$*, then choosing* $\epsilon = \left(\frac{|\mathcal{A}|}{T}\right)^{1/3}$ *gives:*

$$\mathfrak{R}(\pi_{\mathrm{IDS}}, T) \le \sqrt{33 \log(4T)}\, |\mathcal{A}|^{5/6} T^{2/3}.$$

Proof. Since $T \ge 8|\mathcal{A}|$,

$$\epsilon = \left(\frac{|\mathcal{A}|}{T}\right)^{1/3} \le \frac{1}{2},$$

so this choice is admissible. Next,

$$\frac{|\mathcal{A}|T(1-\epsilon)^2}{\epsilon} \le \frac{|\mathcal{A}|T}{\epsilon} = |\mathcal{A}|T \left(\frac{T}{|\mathcal{A}|}\right)^{1/3} = |\mathcal{A}|^{2/3} T^{4/3}.$$

On the other hand, we have:

$$32\epsilon^2 T^2 = 32 \left(\frac{|\mathcal{A}|}{T}\right)^{2/3} T^2 = 32|\mathcal{A}|^{2/3} T^{4/3}.$$

Combining these, we get:

$$\sum_{t=0}^{T-1} \mathbb{E}[\Gamma_t] \le 33|\mathcal{A}|^{2/3} T^{4/3}.$$

The standard information-ratio regret bound gives

$$\mathfrak{R}(\pi, T) \le \sqrt{\left(\sum_{t=0}^{T-1} \mathbb{E}[\Gamma_t]\right) I(\theta, \phi; H_T)}.$$

Substituting the two bounds,

$$\mathfrak{R}(\pi_{\mathrm{IDS}}, T) \le \sqrt{33 \log(4T)} |\mathcal{A}|^{5/6} T^{2/3}.$$

□

Remark 2. *The original IDS analysis gives a bound scaling as*

$$\widetilde{O}(|\mathcal{A}|^{3/4} T^{3/4}).$$

The refinement in lemma 7 improves the horizon dependence from $T^{3/4}$ *to* $T^{2/3}$*, but worsens the action-set dependence from* $|\mathcal{A}|^{3/4}$ *to* $|\mathcal{A}|^{5/6}$*. Thus this is not a uniform improvement. It is most attractive in the long-horizon regime where* T *is the dominant asymptotic concern.*

B Additional tables and figures

Table 2: Catastrophe probability by algorithm and query scheme.

Algorithm	Query scheme	m_{env}	n_q	$P_{\text{exp}}(R \leq C_{0.5})$
known_w_etc	coordinate	0	32	0.056
known_w_etc	coordinate	0	64	0.028
known_w_etc	coordinate	0	96	0.028
known_w_etc	coordinate	0	128	0.000
known_w_etc	coordinate	0	160	0.000
known_w_etc	predicted_feasible	0	32	0.139
known_w_etc	predicted_feasible	0	64	0.222
known_w_etc	predicted_feasible	0	96	0.139
known_w_etc	predicted_feasible	0	128	0.139
known_w_etc	predicted_feasible	0	160	0.056
unknown_w_etc	coordinate	1	32	0.333
unknown_w_etc	coordinate	1	64	0.167
unknown_w_etc	coordinate	1	96	0.194
unknown_w_etc	coordinate	1	128	0.278
unknown_w_etc	coordinate	1	160	0.222
unknown_w_etc	coordinate	2	32	0.222
unknown_w_etc	coordinate	2	64	0.250
unknown_w_etc	coordinate	2	96	0.139
unknown_w_etc	coordinate	2	128	0.111
unknown_w_etc	coordinate	2	160	0.139
unknown_w_etc	coordinate	8	32	0.111
unknown_w_etc	coordinate	8	64	0.083
unknown_w_etc	coordinate	8	96	0.111
unknown_w_etc	coordinate	8	128	0.056
unknown_w_etc	coordinate	8	160	0.056
unknown_w_etc	observed_feasible	1	32	0.389
unknown_w_etc	observed_feasible	1	64	0.389
unknown_w_etc	observed_feasible	1	96	0.194
unknown_w_etc	observed_feasible	1	128	0.222
unknown_w_etc	observed_feasible	1	160	0.278
unknown_w_etc	observed_feasible	2	32	0.333
unknown_w_etc	observed_feasible	2	64	0.250
unknown_w_etc	observed_feasible	2	96	0.250
unknown_w_etc	observed_feasible	2	128	0.056
unknown_w_etc	observed_feasible	2	160	0.306
unknown_w_etc	observed_feasible	8	32	0.139
unknown_w_etc	observed_feasible	8	64	0.167
unknown_w_etc	observed_feasible	8	96	0.083
unknown_w_etc	observed_feasible	8	128	0.056
unknown_w_etc	observed_feasible	8	160	0.028
query_warmstarted_ts	coordinate	1	160	0.000
query_warmstarted_ts	coordinate	2	160	0.000
query_warmstarted_ts	coordinate	8	160	0.000
vanilla_ts_no_queries, prior updates = all	none	2	0	0.944
vanilla_ts_no_queries, prior updates = all	none	8	0	0.917

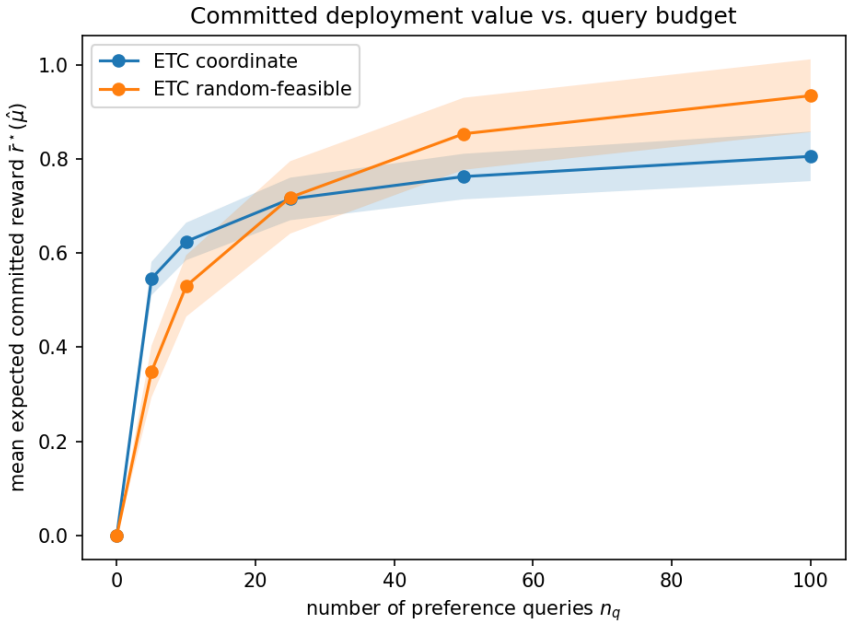

Figure 3: Mean Expected Reward across ETC Coordinate and Random-Feasible Factors

Figure 3 demonstrates a notable difference in mean reward as the number of queries increases for the ETC Coordinate and ETC Random-Feasible agents. In particular, the ETC Random-Feasible agent surpasses the ETC Coordinate agent after fewer than 30 queries. One possible explanation for this behavior is that random-feasible queries may provide more informative comparisons by allowing the agent to observe tradeoffs among multiple factors simultaneously. In contrast, coordinate queries vary one factor at a time, which may limit the agent's ability to infer how different factors should be weighted relative to one another. This could explain why the ETC Coordinate agent achieves a lower mean reward after both agents have conducted a sufficient number of queries.

Table 3: Selected experimental hyperparameters and configurations.

Parameter	Value
Problem dimension K	16
Outcome dimension d	16
Horizon T	4000
Reward penalty λ^*	0.2
Preference temperature τ	1.5
Outcome noise σ_ξ	0.2
Human query cost c	16
Number of problem instances	200
TS deployment steps	4000
Controllability Beta concentration ν	50.0
Controllability sharpness κ	3.0
Controllability threshold quantile	0.9
Status-quo query fraction	0.5