LEARNABLE RANDOMIZATION AS COMMITMENT AGAINST ADAPTIVE OPTIMIZERS

Anonymous authors
Paper under double-blind review

ABSTRACT

A pricing page can walk the posted price up to the last amount a buyer still accepts, a recommender can hold back a better item for a barely acceptable promoted one, and a classifier can shift its boundary once applicants change their features. The system predicts the response and then picks the menu that serves its own objective, so the surplus above the user's cutoff is taken. Playing the single best action publishes that cutoff, while noise on actions the user would never take throws away payoff and teaches the platform that a worse menu is still acceptable. We study unpredictable near-optimal policies (UNOP), which mix uniformly on near-best actions that remain individually rational. The mixture is a commitment about the response. On a finite price grid, when the best sure-demand price strictly out-earns the randomized band, a seller who already knows the curve posts below the band, and the purchase that occurs is deterministic. Knowing that curve is not the same as predicting the next draw. The mixture can be learned and the optimizer can match its best response, while the user's payoff stays higher because the mixture changes which action is targeted. In pricing and in policy-aware recommendation this leaves more surplus than greedy play when the platform optimizes against the curve and more than one action is acceptable. The gain goes away under quality ranking, a singleton near-optimal set, a wrong utility estimate, or a short-horizon explorer. That is also where mixing should be turned off if the other side is trying to cooperate.

1 INTRODUCTION

A growing fraction of human decisions are taken next to an adaptive algorithm, whether a platform sets a price after watching who buys (Calvano et al., 2020; Braverman et al., 2018), a recommender builds a list after watching who clicks (Dean et al., 2020; Hron et al., 2023; Stray et al., 2024), or a classifier is retrained on the features applicants choose to present (Hardt et al., 2016; Perdomo et al., 2020). In each case the algorithm is an optimizer rather than a fixed environment, because it predicts behavior and then chooses the action that maximizes its own objective given that prediction. Once the prediction is accurate the menu moves with it, so the user can keep accepting the offered price, list, or decision while the surplus above the user's cutoff is taken by the optimizer.

Once the interaction is written as an ordinary learning problem, it is easy to study only how to predict the user or how to extract surplus from a user who has already been predicted. Strategic buyers are usually treated as long-horizon manipulators of a learning seller, or the seller is designed so that the manipulation does not pay (Drutsa, 2018; 2020; Huh & Kandasamy, 2026). A pure best response is a deterministic function of the current menu, so an optimizer that identifies it can post a price just below value, withhold a better organic item in favor of a barely acceptable promoted one, or shift a decision boundary against a known gaming direction. Uniform randomization conceals that function only by collapsing the user's own payoff, while ε-greedy and Boltzmann exploration put noise on actions that are far from optimal or individually irrational, and an adaptive optimizer can read those loss-making plays as evidence that the current menu is still acceptable.

We do not try to hide the response distribution, and the pricing guarantee below assumes the optimizer knows it exactly. Figure 1 shows the closed loop. In pricing, once the seller stays below the randomized band, the action that is actually played is often deterministic, because the randomization only specifies what the buyer would do at prices the seller then chooses not to post. We call this the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

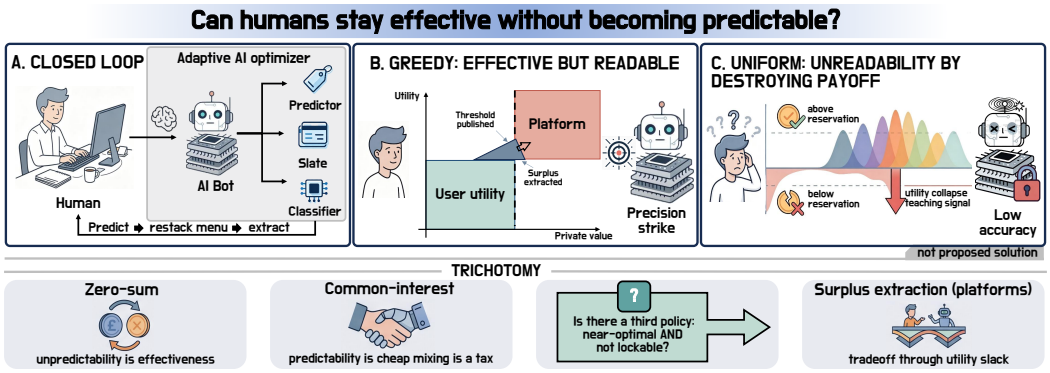


Figure 1: (A) A human and an adaptive optimizer in a closed loop. The optimizer may learn the response distribution and still revise the menu. (B) A greedy threshold publishes the user's cutoff. (C) Uniform noise refuses good actions and destroys payoff. The protective randomness in pricing is the response the buyer would give on menus the optimizer then declines to offer. A mixed security strategy is required in zero-sum play, while mixing is a coordination tax when interests coincide.

response-curve commitment effect. The platform can learn the distribution and optimize against the whole curve, yet still be led to a menu that takes less surplus. The question is therefore not how to keep the policy secret. It is which response curve the user can credibly commit to once the platform is free to learn it.

The shape of the user's payoffs decides how much room that commitment has. Let u be the estimated payoff vector and let $\mathcal{A}_\alpha(u) = \{a : u(a) \geq \max u - \alpha\}$ be the set of actions that never lose more than α relative to the best one-step action. When the set is a singleton, effectiveness forces a predictable action, whereas a larger set can be mixed so as to maximize entropy or to put less weight on actions the current forecast already expects. We call the resulting family unpredictable near-optimal policies (UNOP). Individual rationality is part of the support, because mixing into a strictly loss-making action, such as buying above value or clicking below reservation, teaches the optimizer that the cutoff can still be moved. UNOP therefore mixes only inside $\mathcal{A}_\alpha(u)$ and above a reservation floor.

The same payoff shape produces three different conclusions, which we prove and then measure. In zero-sum games such as matching pennies and rock, paper, scissors, an adaptive best responder turns a greedy pure policy into a security-level loss, so unpredictability is what makes the user effective (Nash, 1950; Osborne & Rubinstein, 1994). In common-interest games such as coordination and stag hunt, the Pareto-optimal pure profile is already effective and fully predictable, so mixing is only a coordination tax and being read is what good play looks like when the other side is a collaborator (Carroll et al., 2019; Hadfield-Menell et al., 2016). In surplus extraction the platform identifies a threshold and sits on the user's cutoff (Hardt et al., 2022). A greedy user gives that cutoff away and a uniform user refuses both the cutoff and the surplus, while UNOP blurs only the boundary, with sure demand below $v - \alpha$, sure refusal above v, and mixing in between. Under Proposition 4 this preserves $\Omega(\alpha)$-scale pricing surplus, and the same support restriction forces a margin-seeking list to trade acceptance against user quality.

Our contributions can be summarized as follows:

- We separate distributional learning error, realized-action predictability, optimizer regret relative to an exact best response, and the surplus that best response extracts.
- We characterize constant-band commitments on a finite price grid. UNOP is the maximum-entropy support-safe response. It induces the same posted price as the welfare-maximizing band rate precisely when purchase rate $1/2$ strictly keeps the seller below the band.
- We show that predictors can learn the mixture and match an oracle best response, while user surplus stays above greedy play in pricing and under a policy-aware recommender. Quality ranking reverses that comparison, a singleton α-set on every feasible menu removes it, and large utility noise or a short-horizon explorer makes the protection fail.

2 RELATED WORK

Learning in Games and Strategizing against Learners. Classical learning in games studies no-regret dynamics and their equilibria (Fudenberg & Levine, 1998; Cesa-Bianchi & Lugosi, 2006). Weighted-majority and exponential-weights bounds make the same point (Littlestone & Warmuth, 1994; Freund & Schapire, 1997). A recent line asks the converse: how should a patient optimizer play against a no-regret learner (Deng et al., 2019; Braverman et al., 2018; Mansour et al., 2022; Camara et al., 2020; Haghtalab et al., 2022; Kolumbus & Nisan, 2022)? The typical conclusion is that the optimizer can secure at least the Stackelberg value, and sometimes more against mean-based learners. Repeated posted-price auctions with a strategic buyer, and seller algorithms designed to remain robust to that buyer, study the other side of the interaction as long-horizon manipulation or as a robust mechanism (Drutsa, 2018; 2020; Huh & Kandasamy, 2026). We study a simple support-constrained commitment and separate policy learning, realization prediction, optimizer regret, and extraction. Proper scoring rules make the same separation at the level of a forecast: log loss equals entropy plus a divergence, so a high log loss need not mean the forecast missed the distribution (Cover & Thomas, 2006).

Performative Prediction and Platforms. Performative prediction (Perdomo et al., 2020; Hardt et al., 2022) and strategic classification (Hardt et al., 2016; Brückner et al., 2012; Dong et al., 2018; Milli et al., 2019; Zrnic et al., 2021; Levanon & Rosenfeld, 2021; Chen et al., 2020; Shavit et al., 2020; Kleinberg & Raghavan, 2019) study distribution shift caused by agent responses. The usual goal is a stable or Stackelberg-optimal predictor. Agent welfare, when it appears, is often a constraint on the institution (Milli et al., 2019; Levanon & Rosenfeld, 2021). We focus on the human side. The same literature, with recommender choice-set design (Dean et al., 2020; Ben-Porat & Tennenholtz, 2018; Hron et al., 2023; Ie et al., 2019; Schnabel et al., 2016) and algorithmic pricing (Calvano et al., 2020; Auer et al., 2002), is where surplus extraction lives. Performative power (Hardt et al., 2022) quantifies how much a platform can steer. We treat that steering as sitting on a predicted threshold.

Entropy-Regularized Play and Coordination. Maximum-entropy IRL and soft RL (Ziebart et al., 2008; Haarnoja et al., 2018; Jaynes, 1957), quantal response (McKelvey & Palfrey, 1995), and information theory (Cover & Thomas, 2006) all justify Gibbs distributions. We use them as the Shannon-Pareto baseline. For a human facing extraction the natural constraint is worst-case utility loss, and the maximum-entropy solution is uniform on the actions within α of the best payoff. Cooperative IRL (Hadfield-Menell et al., 2016; Carroll et al., 2019; Russell, 2019) instead emphasizes making humans easy to predict. We agree when the other side is a collaborator, and we locate where that advice reverses.

3 METHOD

3.1 INTERACTION AND OBJECTIVES

We consider a T-round interaction between a human $\mathcal{H}$ and an adaptive optimizer $\mathcal{O}$. At round t, a context x_t is realized and the history $h_{t-1} = (x_s, a_s, o_s)_{s<t}$ is public. The human chooses $a_t \in A$ from a finite action set and the optimizer chooses $o_t \in O$. Instantaneous payoffs are $u_H(a_t, o_t, x_t)$ and $u_O(a_t, o_t, x_t)$. The optimizer maintains a predictive distribution $P_t(\cdot \mid x_t, h_{t-1})$ on A, plays a best response that may still explore

$$o_t \in \arg\max_{o \in O} \mathbb{E}_{a \sim P_t}\big[u_O(a, o, x_t)\big] \tag{1}$$

or a bandit version of equation 1, such as UCB on posted prices or unique-best recommendation lists, and updates P_t from a_t. In simultaneous-move games the human does not see the optimizer's current mix, and forms

$$\hat{u}_t(a) \;=\; \sum_{o \in O} \hat{\sigma}_t(o)\, u_H(a, o, x_t), \tag{2}$$

where $\hat{\sigma}_t$ is the empirical mix of past optimizer actions in simultaneous bimatrix play. In sequential menus, namely pricing and recommendation, the human observes the posted o_t before acting, so $\hat{u}_t$ is the displayed utility of the current menu rather than an average over past menus. Then $a_t \sim \pi_t(\cdot \mid \hat{u}_t, P_t)$.

Definition 1. *The human's effectiveness is* $V_H = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[u_H(a_t, o_t, x_t)]$. *In surplus-extraction domains we report the analogous surplus over a reservation utility.*

Definition 2. *Let π be the user's mixed policy and P a predictive distribution, with probabilities clipped at 10^{-12} before logarithms. Distributional learning error is $L(\pi, P) = \mathrm{KL}(\pi \| P)$, and realization predictability is the Bayes accuracy $R(\pi) = \max_a \pi(a)$, which upper-bounds every predictor's one-step hit rate. Write $V_O^\star(\pi) = \max_o \mathbb{E}_{a\sim\pi}[u_O(a, o)]$ for the optimizer's exact best-response value and $o(P)$ for the action it plays from P. Optimizer regret is the oracle gap $G(\pi, P) = V_O^\star(\pi) - V_O(\pi, o(P))$, while extraction is $V_O^\star(\pi)$ itself, or the user's surplus loss relative to a menu that is not chosen to extract. A zero gap means only that the learned optimizer has reached the oracle best response, not that the policy cannot be exploited.*

Proposition 1. *For any π and P in the simplex, the best one-step hit rate satisfies $\max_P \mathbb{P}_{a\sim\pi}(\arg\max P = a) = \max_a \pi(a)$. The expected log loss decomposes as $\mathbb{E}_{a\sim\pi}[-\log P(a)] = H(\pi) + \mathrm{KL}(\pi \| P)$, so the excess over the entropy is zero if and only if $P = \pi$ on the support of π. Raw log loss and raw top-1 accuracy therefore cannot, by themselves, show that a predictor failed to learn the mixture.*

These are operational measures: they are what an adaptive AI system computes. We do not equate unpredictability with differential privacy (Dwork, 2006; Duchi et al., 2013). Write $u \in \mathbb{R}^{|A|}$ for the current expected-utility vector, $u^\star = \max_a u(a)$, and

$$\mathcal{A}_\alpha(u) \;=\; \{a \in A : u(a) \geq u^\star - \alpha\}. \tag{3}$$

Every action in $\mathcal{A}_\alpha(u)$ loses at most α relative to the best one-step action. If $|\mathcal{A}_\alpha(u)| = 1$, any policy supported on that singleton is a δ-peak with $\delta = 1$, aside from how ties are broken, and effectiveness forces predictability. Appendix D measures robustness when $\hat{u}$ is a noisy version of u.

3.2 Three cases and two tradeoffs

Proposition 2. *In a finite zero-sum game, any pure action whose security level is below the game value $v^\star$ is exploitable by an exact best responder. In matching pennies, every deterministic policy has value -1. More generally, let $\mathcal{S} = \{\pi \in \Delta(A) : \min_o u_H(\pi, o) \geq v^\star\}$ be the security-strategy polytope. There is a game-dependent constant $\kappa_G < \infty$ such that every policy with value at least $v^\star - \varepsilon$ satisfies*

$$\inf_{\pi^\star \in \mathcal{S}} \mathrm{TV}(\pi, \pi^\star) \leq \kappa_G \varepsilon.$$

Thus, in games such as matching pennies and rock, paper, scissors, whose security set contains no peaked policy, sufficiently effective play must remain non-peaked. The constant and the conclusion are game dependent. Zero-sum structure alone does not imply that every game has a unique or fully mixed security strategy.

The proof is standard security-level reasoning (Osborne & Rubinstein, 1994). We record it in Appendix B. Empirically, greedy obtains mean payoff -0.143 in our zero-sum suite, while UNOP sits at 0.000 and Aware-UNOP at $+0.054$.

Proposition 3. *Suppose $u_H = u_O = u$ and $(a^\star, o^\star)$ is the unique global payoff-maximizing pure profile. Let $u^\star = u(a^\star, o^\star)$ and $\bar{u} = \max_{a \neq a^\star, o} u(a, o) < u^\star$. Then the effectiveness-maximizing human policy is a point mass on $a^\star$, paired with $o^\star$. For any mixture with $\pi(a^\star) = p < 1$, and for any partner action, including a best response, expected common payoff is at most $pu^\star + (1-p)\bar{u}$, hence it pays at least the linear coordination tax $(1-p)(u^\star - \bar{u})$. A Shannon constraint $H(\pi) \geq h > 0$ therefore forces $p \leq p^\star(h) < 1$.*

In our common-interest suite, greedy, UNOP with small α, and Aware-UNOP all achieve payoff 2.598, with predictive accuracy 1.0. Uniform falls to 1.762.

The two natural ways of asking a policy to be random but not too costly are not interchangeable. One constrains every atom of the support. The other constrains only the mean.

Theorem 1. *Fix $u \in \mathbb{R}^{|A|}$ and write $\Delta(A)$ for the simplex. Consider*

$$(\mathrm{WC}_\alpha) \qquad \max_{\pi \in \Delta(A)} H(\pi) \quad \textit{s.t.} \quad \pi(a) = 0 \textit{ whenever } u(a) < u^\star - \alpha, \tag{4}$$

$$(\mathrm{SH}_h) \qquad \max_{\pi \in \Delta(A)} \pi \cdot u \quad \textit{s.t.} \quad H(\pi) \geq h. \tag{5}$$

1. *Per-action worst-case loss. (WC_α) is uniquely solved by $\pi = \mathrm{Unif}(\mathcal{A}_\alpha(u))$, with value $H = \log|\mathcal{A}_\alpha(u)|$ and $\mathbb{E}_\pi[u] \geq u^\star - \alpha$. Conversely, a policy places positive mass outside*

$\mathcal{A}_\alpha$ if and only if it violates the support-wise guarantee $\max_{a\in\mathrm{supp}(\pi)}(u^\star - u(a)) \le \alpha$. This converse concerns the loss of every action that may be played. A policy with very small mass outside $\mathcal{A}_\alpha$ may still have expected loss below α.

2. *Shannon constraint. For any $h \in \big(H(\mathrm{Unif}(\arg\max u)), \log|A|\big]$, (SH_h) is uniquely solved by a Gibbs distribution $\pi_\tau(a) \propto \exp(u(a)/\tau)$ for some $\tau > 0$ (Jaynes, 1957; Cover & Thomas, 2006). Equivalently, π_τ maximises $\pi \cdot u + \tau H(\pi)$ over $\Delta(A)$. For $h \le H(\mathrm{Unif}(\arg\max u))$, every expected-utility maximizer with entropy at least h is supported on $\arg\max u$.*

(WC_α) maximises a strictly concave objective on a simplex face and is therefore uniquely solved. (SH_h) maximises a linear objective over the strictly convex set of policies whose entropy is at least h. Uniqueness of the Gibbs solution on $\big(H(\mathrm{Unif}(\arg\max u)), \log|A|\big]$ follows from that shape. KKT stationarity for (WC_α) on the relative interior of $\Delta(\mathcal{A}_\alpha)$ is $-\log\pi(a) - 1 + \lambda = 0$, so π is constant on $\mathcal{A}_\alpha$. Equivalently, $\mathrm{Unif}(\mathcal{A}_\alpha)$ is the I-projection of $\mathrm{Unif}(A)$ onto $\{\pi : \mathrm{supp}(\pi) \subseteq \mathcal{A}_\alpha\}$, since $\mathrm{KL}(\pi\|\mathrm{Unif}(A)) = \log|A| - H(\pi)$. For (SH_h), stationarity of $\pi \cdot u + \tau H(\pi)$ is $u(a) - \tau(\log\pi(a) + 1) + \mu = 0$, hence Gibbs.

The programs coincide only in special cases, when a single α-level set is already A, or when $h = 0$. For any $\tau > 0$, $\pi_\tau(a) > 0$ for every a. If $\mathcal{A}_\alpha \neq A$, write $\gamma = \min_{a\notin\mathcal{A}_\alpha}(u^\star - \alpha - u(a)) > 0$. Matching bounds on the normalizer $Z = \sum_b e^{u(b)/\tau}$ give

$$\frac{1}{|A|}\exp\big(-(\alpha+\gamma)/\tau\big) \;\le\; \pi_\tau\big(A \setminus \mathcal{A}_\alpha\big) \;\le\; \frac{|A \setminus \mathcal{A}_\alpha|}{|\mathcal{A}_\alpha|}\exp(-\gamma/\tau). \tag{6}$$

The leak is acceptable as exploration in RL. Against an extractor it is a teaching signal. Unrestricted Boltzmann and ε-greedy put mass outside $\mathcal{A}_\alpha$, including individually irrational actions. The pricing comparisons in Section 4.2 use two different baselines: IR-Boltzmann refuses purchases above value but can still leave $\mathcal{A}_\alpha$, and support-matched Boltzmann is renormalized on $\mathcal{A}_\alpha^{\mathrm{IR}}$. Appendix F.4 compares UNOP with and without the reservation floor.

Theorem 2. *Let P_t be the Dirichlet-multinomial posterior mean with a uniform prior on a finite action set, and let actions be drawn i.i.d. from a fixed π. Then $P_t \to \pi$ almost surely, so $\mathrm{KL}(\pi\|P_t) \to 0$ on the support of π and $\max_a P_t(a) \to \max_a \pi(a)$. If π is a δ-peak with $\delta > 1/2$, the posterior mode equals that action after $O\big(\log(1/\eta)/(\delta - 1/2)^2\big)$ observations with probability at least $1-\eta$. If π is uniform on k actions, the same convergence leaves $\max_a P_t(a) \to 1/k$. The mixture is identified, and the one-step hit rate remains bounded by the policy. The proof records how the visit counts grow.*

Proposition 4. *Let $\mathcal{P}$ be a price grid with spacing Δ and let the buyer use the constant-band response*

$$q_{\alpha,\rho}(p) = \begin{cases} 1 & p < v - \alpha, \\ \rho & v - \alpha \le p \le v, \\ 0 & p > v, \end{cases} \tag{7}$$

with $\rho \in (0,1)$. Write $p_\alpha^-(v) = \max\{p \in \mathcal{P} : p < v - \alpha\}$ and $p_\alpha^+(v) = \max(\mathcal{P} \cap [v-\alpha, v])$, when those sets are nonempty. A seller who knows $q_{\alpha,\rho}$ and breaks revenue ties by lowest buyer surplus posts in the sure-demand region whenever

$$p_\alpha^-(v) > \rho\, p_\alpha^+(v). \tag{8}$$

The posted price is then $p_\alpha^-(v)$ and buyer surplus is $v - p_\alpha^-(v) \in (\alpha, \alpha + \Delta]$. The guarantee uses the response curve itself and does not use prediction error. UNOP is the case $\rho = 1/2$. Let $r^\star = p_\alpha^-(v)/p_\alpha^+(v)$. Every $\rho < r^\star$ induces that same price, so those rates are exactly the welfare-maximizing constant-band commitments. Binary entropy on $(0,1)$ is maximized at $1/2$, which means that $\rho = 1/2$ is the unique entropy-maximizing rate that keeps the seller below the band when $r^\star > 1/2$. When $r^\star \le 1/2$ that set of rates, $(0, r^\star)$, is open, so no entropy-maximizing rate exists and entropy only increases as ρ approaches $r^\star$ from below. The inequality $\alpha \le v/2$ is not enough for this conclusion, because if $v = 2\alpha$ and $v \in \mathcal{P}$, then $p_\alpha^+(v) = v$ and $p_\alpha^-(v) < \alpha = v/2$, so equation 8 fails at $\rho = 1/2$ and the seller can post inside the band.

Proposition 4 is a commitment about prices that are not posted. Once the seller stays below the band, the action at the posted price $p_\alpha^-(v)$ is a sure purchase, and the mixture that moved the seller

sits on prices that are not posted. On a uniform grid the strict inequality holds for $\rho = 1/2$ when $\alpha < v/2 - \Delta$, and it can fail on the boundary $\alpha = v/2$.

3.3 Mixing only where it is safe

Figure 2 summarizes the pipeline. Individual rationality is part of the support: mixing into strictly loss-making actions, such as buying above value or clicking below reservation, teaches the optimizer to move the cutoff. Given u, slack $\alpha \geq 0$, and reservation $\underline{u} \in \{-\infty\} \cup \mathbb{R}$, define

$$\mathcal{A}_\alpha^{\mathrm{IR}}(u) \;=\; \begin{cases} \mathcal{A}_\alpha(u) \cap \{a : u(a) \geq \underline{u}\} & \text{if that intersection is nonempty,} \\ \mathcal{A}_\alpha(u) & \text{otherwise.} \end{cases} \tag{9}$$

Exact arithmetic makes the first branch apply whenever the reservation set is nonempty, because $u^\star \geq \underline{u}$ then forces the global maximizer into the intersection. The second branch covers rare numerical cases. The reservation is 0 for skip in posted pricing, τ in recommendation, and $-\infty$ in bimatrix games.

Unaware UNOP is the unique solution of (WC_α) on $\mathcal{A}_\alpha^{\mathrm{IR}}$. When a forecast P is available in the same simplex, Aware-UNOP replaces maximising $H(\pi)$ by a strictly concave objective that downweights actions the forecast already expects

$$\pi^{\mathrm{aware}} \;\in\; \arg \max_{\substack{\pi \in \Delta(A) \\ \mathrm{supp}(\pi) \subseteq \mathcal{A}_\alpha^{\mathrm{IR}}}} \Big(H(\pi) - \mathbb{E}_{a \sim \pi}\big[\log\big(P(a) + \varepsilon\big)\big] \Big), \tag{10}$$

with $\varepsilon > 0$ a numerical floor. The unique maximizer is $\pi^{\mathrm{aware}}(a) \;\propto\; \big(P(a) + \varepsilon\big)^{-1}$ on $\mathcal{A}_\alpha^{\mathrm{IR}}$, equivalently the I-projection

$$\pi^{\mathrm{aware}} \;=\; \arg \min_{\pi \in \Delta(\mathcal{A}_\alpha^{\mathrm{IR}})} \mathrm{KL}\big(\pi \,\big\|\, \nu_P\big), \qquad \nu_P(a) \;\propto\; \big(P(a) + \varepsilon\big)^{-1}. \tag{11}$$

If P is uniform on that set, equation 10 reduces to unaware UNOP. Algorithm 1 is the one-round map used in every experiment, and the surrounding loops are in Appendix C. It is a good-enough mixer rather than a no-regret algorithm, and Hedge can be exploited by a patient optimizer (Deng et al., 2019). Across extraction domains, α on the order of 0.05 to 0.15 of the utility range is consistently near-best (Appendix Figure 12).

4 Experiments

4.1 Experimental Settings

Bimatrix games, posted pricing, and recommendation are the three interaction classes. The new comparisons below use an oracle seller that knows q_π, a Thompson seller, a stationary tax game, and a policy-aware slate optimizer. Seeds, horizons, and raw tables for the UCB, MovieLens, and tabular runs are in `supplement/legacy`. Aware-UNOP sees the true forecast P_t. Unaware UNOP does not. Probabilities are clipped at 10^{-12} before every logarithm.

4.2 Posted pricing

We draw 200 values $v \sim U[0.25, 0.95]$ and hold out 120 for test. The α we report is 0.15, chosen before the test split is used. Mean-surplus selection on the validation split returns 0.20, which raises the average and worsens the lower tail, so the test numbers stay at 0.15.

The static test uses a demand oracle that knows q_π and breaks revenue ties by lowest buyer surplus, then by highest price. On the 41-point test grid the experimental unit is a value, greedy surplus is 0.012, and UNOP surplus is 0.145. The paired difference is 0.133, with a 95% bootstrap CI $[0.124, 0.141]$ over 10,000 resamples of the 120 values, and the difference is positive on 89.2% of them. A 10,001-point average over values on $[0.25, 0.95]$, with the 41-point price grid held fixed, gives 0.013 and 0.146. Bayes accuracy at the posted price is 0.946, because the purchase is sure whenever equation 8 holds. The 10.8% of test values on which UNOP does not beat greedy are exactly the values on which $\rho = 1/2$ fails equation 8 and the seller enters the band. An oracle support-safe commitment, which may lower ρ until equation 8 holds, reaches mean surplus 0.162 on the same test split and 0.163 under the same fine average. It uses the seller's revenue objective,

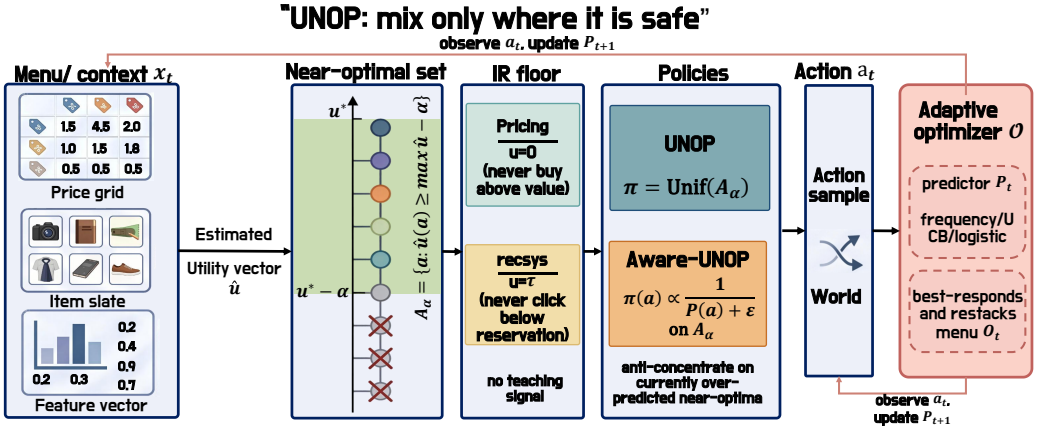


Figure 2: UNOP pipeline. From the current menu the human forms $\hat{u}$, builds the α-optimal set, intersects it with the reservation floor, and mixes uniformly on that set. The optimizer observes a_t, updates P, and revises the menu, and that update can converge to the mixture. The welfare effect comes from the best response to the mixture. In pricing, once the seller stays below the band, the relevant mixture is off the posted price.

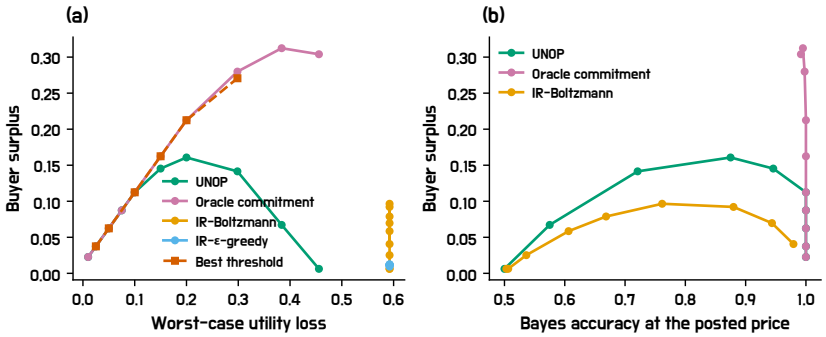


Figure 3: Test-split oracle pricing. (a) Buyer surplus against worst-case utility loss. UNOP is the maximum-entropy support-safe point. The plotted IR-Boltzmann curve still leaves $\mathcal{A}_\alpha$, while support-matched Boltzmann at the same α has surplus 0.1454, against 0.1453 for UNOP. The oracle band commitment rises above UNOP once $\rho = 1/2$ fails equation 8. (b) Bayes accuracy at the posted price can be high at the same time, because the mixture that keeps the seller below the band is off that price.

so we treat it as an upper bound rather than a second rule a user could deploy. Support-restricted Boltzmann at the same $\alpha = 0.15$ has surplus 0.1454, against 0.1453 for UNOP, so the gain is carried by the hard support and the induced demand curve, with uniform mixing as the parameter-free maximum-entropy point of that class.

Table 1 places this static pricing comparison next to the online demand oracle, the two recommendation objectives, and the tax game. Green cells are the higher user payoff when the counterpart evaluates the response curve. The blue cell is quality ranking, where the list is not chosen against that curve and greedy remains ahead. The subsections below report these comparisons one setting at a time.

Figure 4a is a separate online comparison, with four sellers on 60 random values for 300 rounds, and those bars are not a ranked strength scale. Table 2 records buyer surplus for greedy and UNOP under each seller. Green marks where UNOP pulls ahead. Under UCB the two policies stay close, because that seller is still walking down a 41-point grid after 300 rounds and has not yet evaluated the response curve. Under the demand oracle the per-price KL is 0.0043.

4.3 Recommendation

On catalogs of 12 items and slates of size 4, every one of the 495 slates is scored with the true $\pi(\cdot \mid S)$ over 160 catalogs, and beam search of width 50 matches that exact optimum. Quality ranking favors

Table 1: Headline user payoffs used in this section. Green is the higher payoff when the counterpart evaluates the response curve. Blue is quality ranking, where greedy stays ahead.

Comparison	Greedy	UNOP
Static oracle pricing	0.012	0.145
Online demand oracle	0.011	0.151
Quality-ranking lists	0.495	0.459
Policy-aware lists	0.082	0.160
Tax-game payoff	0.750	0.860

Table 2: Buyer surplus against four online sellers, 60 values and 300 rounds. Green marks a clear UNOP gain. The UCB row stays close because that seller is still exploring.

Seller	Greedy	UNOP
EXP3	0.156	0.191
UCB	0.004	0.006
Thompson sampling	0.010	0.107
Demand oracle	0.011	0.151

greedy, with paired difference -0.036 and 95% CI $[-0.040, -0.032]$. The policy-aware oracle reverses the sign, so UNOP minus greedy is 0.078, CI $[0.064, 0.092]$, positive on 74.4% of catalogs, with means 0.160 versus 0.082. A separate 48-item experiment, with slate size 8, beam width 30, and 80 catalogs, gives paired difference 0.102, CI $[0.090, 0.115]$, positive on 98.8% of catalogs. A later latent-correlation sweep, using beam width 50 and 60 catalogs per cell, asks whether that gap depends on conflict between quality and margin (Appendix Figure 11). It does not trace support size. The policy-aware gap is largest under strong negative latent correlation, where at correlation -0.8 and reservation quantile 0.2 the mean gap is 0.104, 95% CI $[0.073, 0.139]$. Under strong positive latent correlation the same gap is indistinguishable from zero (-0.002, CI $[-0.009, 0.004]$). Several intermediate cells have a positive mean and a zero median, and quality ranking stays negative at every tested latent correlation from -0.8 to 0.8. The sweep is about conflict, and it is not a map over how large the acceptable set is.

4.4 Learning the mixture

Figure 5 is a six-action stationary tax game with 100 seeds. The user commits to a fixed mixture, the optimizer taxes the predicted mode, and the oracle taxes the true mode. For UNOP, Bayes accuracy is 0.250, realized accuracy is 0.248, frequency KL is 0.012, and optimizer regret is 0.0000, while user payoff is 0.860 against 0.750 for greedy. The predictor has reached the oracle best response, so what remains random is only the next draw from the learned mixture, and the user's payoff stays higher because that mixture changes which action is taxed.

4.5 Discussion

The pricing and recommendation gains come from the response curve the optimizer best-responds to. They are not evidence that the mixture stayed hidden. In the tax game the predictor's regret falls to zero, so the mixture has been identified, and the higher user payoff comes from which action that mixture makes costly to tax. The pricing misses line up with the same object. The held-out values on which uniform mixing does not beat greedy are the values on which a purchase rate of one half fails to keep the seller below the band, so those misses are named by the commitment condition rather than left as residual noise. Recommendation makes the contrast sharper. A ranker that ignores the response prefers the greedy list, while a slate optimizer that scores the true mixture reverses the comparison, and the reversal is largest when high quality and high margin pull apart. The short-horizon dynamic program is the boundary of the claim. Once the seller is still collecting samples instead of evaluating the curve, the optimal buyer collapses to greedy and the mixture can cost surplus.

Against a five-price UCB seller, exact dynamic programming at horizons 10, 15, 20, and 25 gives the optimal buyer and greedy the same value at every tested value, and UNOP is weakly worse on several of those values, which fits a seller that has not yet evaluated the curve. A singleton acceptable set is a different failure. If $\mathcal{A}_{\alpha}^{\mathrm{IR}}(S)$ is a singleton for every feasible menu S, UNOP and greedy induce the same response, while a singleton only on the full candidate set still leaves the platform free to offer

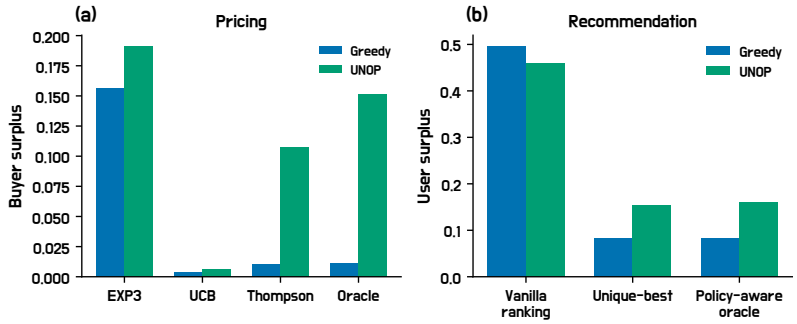


Figure 4: (a) Posted pricing under four online sellers, 60 values and 300 rounds, not a ranked strength scale. EXP3 mixes exploration with importance-weighted updates, and UCB at this horizon is still exploring. Thompson sampling and the demand oracle are closer to a best response to q_π. (b) Recommendation. The experimental unit is a catalog. Quality ranking favors greedy, and an exact policy-aware slate optimizer reverses the comparison.

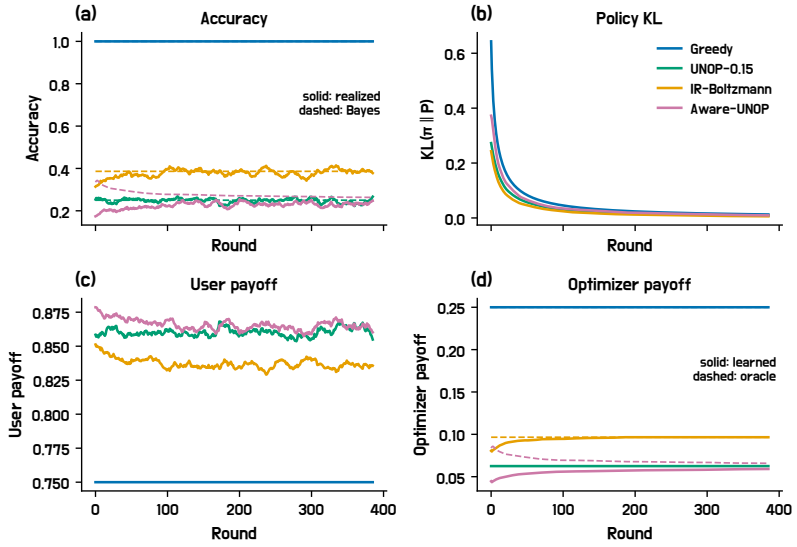


Figure 5: Stationary tax game, frequency predictor, 100 seeds, rolling means. Solid curves in (a) are realized accuracy and dashed curves are Bayes accuracy. Panel (b) is policy KL. In (d), solid curves are the learned optimizer and dashed curves are $V_O^\star(\pi)$, and their gap is optimizer regret rather than a measure of whether the policy can be exploited. The mixture is learned, the next draw stays at accuracy $1/k$, and user payoff stays above greedy.

another menu. At $\alpha = 0$ the two policies have identical buyer surplus on all 120 test values, and the restricted MovieLens-100K cohort is a near-collapse, with mean difference -0.0009 over 33 users. Utility noise $\hat{u}(a) = u(a) + \xi$, $\xi \sim \mathcal{N}(0, \sigma^2)$, against the oracle seller at $\alpha = 0.15$ leaves surplus 0.145 at $\sigma = 0$ and -0.006 at $\sigma = 0.30$, with a true loss-above-α rate of 0.119. A practical reading is therefore narrow. Keep the mixture when the platform will best-respond to the whole curve and the menu still leaves more than one acceptable action, and drop it when the estimate of that set is wrong or the other side is a collaborator.

5 Conclusion

The menu moves because the platform best-responds to a response curve. UNOP mixes uniformly on near-optimal, individually rational actions, and in pricing that mixture can keep a seller below the randomized band even after the curve is known. The gain appears against an extractive counterpart that evaluates the curve, and it disappears when the acceptable set collapses, the utility estimate is badly wrong, or the other side is still exploring.

ETHICS STATEMENT

This work studies how humans can avoid having their preferences precisely identified by adaptive pricing and recommendation systems. The methods can resist targeting that takes surplus and, in principle, useful personalization. We do not provide tools for attacking deployed systems. MovieLens-100K/1M (Harper & Konstan, 2015), UCI Adult (Kohavi, 1996), German Credit (Hofmann, 1994), FICO HELOC, UCI Credit Card, Give Me Some Credit, and UCI Bank Marketing are used as public research files. Any real-world assistant implementing UNOP should make α user-visible and disable mixing in explicitly cooperative modes, as in Proposition 3.

REPRODUCIBILITY STATEMENT

Code and raw outputs for every reported comparison are in the supplement. Oracle pricing, the stationary tax game, dynamic programming, utility noise, the seller family, policy-aware recommendation, the grid condition, the fine value-grid average, the latent-correlation sweep, and the singleton-support runs are under `supplement/sprint`, while bimatrix, MovieLens, and tabular outputs are under `supplement/legacy`. MovieLens-100K `u.data` is included, and MovieLens-1M together with the OpenML tabular sources are fetched by `download_data.py` and `run_public.py` when absent. Values for the oracle suite use seed 20260926, and the test split is the last 60% of 200 draws. Commands and package versions are in `supplement/README.md`. No GPU is required. Proofs are in Appendix B.

AI USE STATEMENT

Artificial intelligence tools assist in formatting and text polishing. The authors take full responsibility for the final content.

REFERENCES

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

Omer Ben-Porat and Moshe Tennenholtz. A game-theoretic approach to recommendation systems with strategic content providers. In *Advances in Neural Information Processing Systems*, volume 31, pp. 1110–1120, 2018.

Mark Braverman, Jieming Mao, Jon Schneider, and S. Matthew Weinberg. Selling to a no-regret buyer. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pp. 523–538, 2018.

Michael Brückner, Christian Kanzow, and Tobias Scheffer. Static prediction games for adversarial learning problems. *Journal of Machine Learning Research*, 13(85):2617–2654, 2012.

Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolò, and Sergio Pastorello. Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10):3267–3297, 2020.

Modibo K. Camara, Jason D. Hartline, and Aleck Johnsen. Mechanisms for a no-regret agent: Beyond the common prior. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 259–270, 2020.

Micah Carroll, Rohin Shah, Mark K. Ho, Thomas L. Griffiths, Sanjit A. Seshia, Pieter Abbeel, and Anca D. Dragan. On the utility of learning about humans for human-AI coordination. In *Advances in Neural Information Processing Systems*, volume 32, pp. 5174–5185, 2019.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Yiling Chen, Yang Liu, and Chara Podimata. Learning strategy-aware linear classifiers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 15265–15276, 2020.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2 edition, 2006.

Sarah Dean, Sarah Rich, and Benjamin Recht. Recommendations and user agency: The reachability of collaboratively-filtered information. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 436–445, 2020.

Yuan Deng, Jon Schneider, and Balasubramanian Sivan. Strategizing against no-regret learners. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pp. 55–70, 2018.

Alexey Drutsa. Weakly consistent optimal pricing algorithms in repeated posted-price auctions with strategic buyer. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Alexey Drutsa. Optimal non-parametric learning in repeated contextual auctions with strategic buyer. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 429–438, 2013.

Cynthia Dwork. Differential privacy. In *International Colloquium on Automata, Languages, and Programming*, pp. 1–12. Springer, 2006.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

Drew Fudenberg and David K. Levine. *The Theory of Learning in Games*. MIT Press, 1998.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1861–1870. PMLR, 2018.

Dylan Hadfield-Menell, Anca D. Dragan, Pieter Abbeel, and Stuart Russell. Cooperative inverse reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 29, pp. 3909–3917, 2016.

Nika Haghtalab, Thodoris Lykouris, Sloan Nietert, and Alexander Wei. Learning in Stackelberg games with non-myopic agents. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pp. 917–918, 2022.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pp. 111–122, 2016.

Moritz Hardt, Meena Jagadeesan, and Celestine Mendler-Dünner. Performative power. In *Advances in Neural Information Processing Systems*, volume 35, pp. 22969–22981, 2022.

F. Maxwell Harper and Joseph A. Konstan. The MovieLens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):19:1–19:19, 2015.

Alan J. Hoffman. On approximate solutions of systems of linear inequalities. *Journal of Research of the National Bureau of Standards*, 49(4):263–265, 1952.

Hans Hofmann. Statlog (German credit data). UCI Machine Learning Repository, 1994.

Jiri Hron, Karl Krauth, Michael I. Jordan, Niki Kilbertus, and Sarah Dean. Modeling content creator incentives on algorithm-curated platforms. In *International Conference on Learning Representations*, 2023.

Joon Suk Huh and Kirthevasan Kandasamy. Strategy-robust online learning in contextual pricing. In *Proceedings of the 37th International Conference on Algorithmic Learning Theory*, 2026.

Eugene Ie, Vihan Jain, Jing Wang, Sanmit Narvekar, Ritesh Agarwal, Rui Wu, Heng-Tze Cheng, Tushar Chandra, and Craig Boutilier. SlateQ: A tractable decomposition for reinforcement learning with recommendation sets. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 2592–2599, 2019.

E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, 1957.

Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pp. 825–844, 2019.

Ron Kohavi. Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 202–207, 1996.

Yoav Kolumbus and Noam Nisan. Auctions between regret-minimizing agents. In *Proceedings of the ACM Web Conference 2022*, pp. 100–111, 2022.

Sagi Levanon and Nir Rosenfeld. Strategic classification made practical. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6243–6253. PMLR, 2021.

Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.

Yishay Mansour, Mehryar Mohri, Jon Schneider, and Balasubramanian Sivan. Strategizing against learners in Bayesian games. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 5221–5252. PMLR, 2022.

Richard D. McKelvey and Thomas R. Palfrey. Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1):6–38, 1995.

Smitha Milli, John Miller, Anca D. Dragan, and Moritz Hardt. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 230–239, 2019.

John F. Nash. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, 36(1):48–49, 1950.

Martin J. Osborne and Ariel Rubinstein. *A Course in Game Theory*. MIT Press, 1994.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7599–7609. PMLR, 2020.

Stuart Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, 2019.

Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1670–1679. PMLR, 2016.

Yonadav Shavit, Benjamin Edelman, and Brian Axelrod. Causal strategic linear regression. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8676–8686. PMLR, 2020.

Jonathan Stray, Alon Y. Halevy, Parisa Assar, Dylan Hadfield-Menell, Craig Boutilier, Amar Ashar, Chloé Bakalar, Lex Beattie, Michael D. Ekstrand, Claire Leibowicz, Connie Moon Sehat, Sara Johansen, Lianne Kerlin, David Vickrey, Spandana Singh, Sanne Vrijenhoek, Amy Xian Zhang, McKane Andrus, Natali Helberger, Polina Proutskova, Tanushree Mitra, and Nina Vasan. Building human values into recommender systems: An interdisciplinary synthesis. *ACM Transactions on Recommender Systems*, 2(3):20:1–20:57, 2024.

Brian D. Ziebart, Andrew L. Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, pp. 1433–1438, 2008.

Tijana Zrnic, Eric Mazumdar, S. Shankar Sastry, and Michael I. Jordan. Who leads and who follows in strategic classification? In *Advances in Neural Information Processing Systems*, volume 34, pp. 15257–15269, 2021.

Appendix

The appendices collect proofs of the claims in Section 3, the interaction loops used in every experiment, and further empirical detail. Algorithm 1 is the human's one-round mixer. Algorithms 2 to 4 are the environments it is dropped into. They match the repository simulators on the objects that affect reported numbers: support construction, the IR intersection, UCB scores, and unique-best filler items.

Algorithm 1 One-round UNOP / Aware-UNOP

Require: $u \in \mathbb{R}^{|A|}$, $\alpha \geq 0$, reservation $\underline{u}$, optional forecast P, aware flag, $\varepsilon > 0$
$u^\star \leftarrow \max_a u(a) \quad \mathcal{A} \leftarrow \{a : u(a) \geq u^\star - \alpha\}$
if $\mathcal{A} \cap \{a : u(a) \geq \underline{u}\} \neq \emptyset$ **then**
  $\mathcal{A} \leftarrow \mathcal{A} \cap \{a : u(a) \geq \underline{u}\}$ ▷ equation 9
end if
if aware **and** P is present **then**
  $\pi(a) \propto (P(a) + \varepsilon)^{-1}$ on $\mathcal{A}$ ▷ equation 10
else
  $\pi \leftarrow \mathrm{Unif}(\mathcal{A})$ ▷ equation 4. Unaware UNOP ignores P
end if
return $a_t \sim \pi$

A Earlier controlled comparisons

A.1 Comparison Results

Figure 6 plots, for each bimatrix game, predictive accuracy against human payoff normalized within the game. In zero-sum games the effective policies are the unpredictable ones: greedy and ε-greedy sit at the bottom, Aware-UNOP at the top. In common-interest games the mass concentrates at the top-right: being predicted is what good play looks like. In general-sum games the cloud is interior, and UNOP occupies the attractive region, slightly above greedy in payoff, with lower accuracy. Absolute payoffs (Appendix D) match the picture: zero-sum suite greedy -0.143 vs. Aware-UNOP $+0.054$. Coordination greedy $=$ UNOP $= 1.95$ vs. uniform 1.00.

Table 3 and Figure 7 summarize a UCB monopolist learning a hidden value $v \sim U[0.25, 0.95]$. Greedy buyers are identified almost perfectly (accuracy 0.999) and retain surplus 0.018, on the order of the price grid. UNOP with $\alpha = 0.05$ more than doubles surplus to 0.042 while remaining individually rational. $\alpha = 0.15$ keeps 0.040 at accuracy 0.807 rather than 0.999. Aware-UNOP matches this surplus at still lower accuracy (0.785). Boltzmann and uniform buy above value and finish negative. ε-greedy, which spends exploration on the dominated action, is worse than greedy. Table 4 reports the matching log-loss and optimizer-payoff numbers: under UNOP-$\alpha = 0.15$, pricing V_O falls from 0.479 (greedy) to 0.241, and log loss rises from 0.026 to 0.312. Appendix Figure 16 shows the same runs as time series: greedy posted prices climb toward v. UNOP keeps a lower posted price. The predictor's hit rate settles near the policy's Bayes accuracy rather than revealing a failure to learn the mixture (Section 4.2).

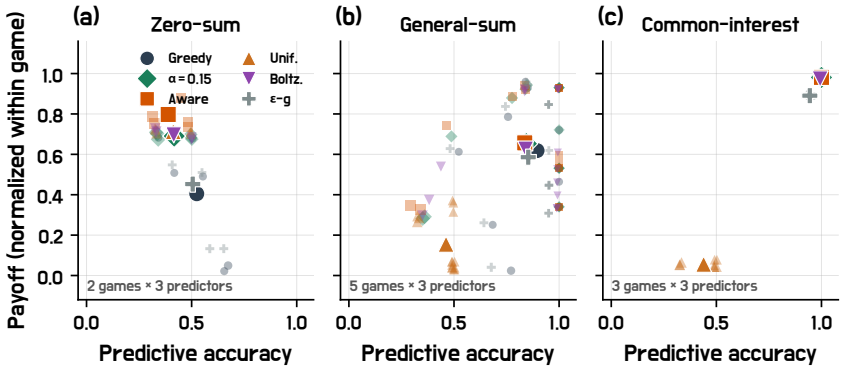


Figure 6: (a) Zero-sum: mixing is mandatory. (b) General-sum: an interior frontier, and UNOP sits on it. (c) Common-interest: the Pareto point is pure. Faint markers are game and predictor pairs, with size proportional to policy entropy. Diamonds and circles are class-wise means ± sd.

Table 3: Main extraction results (mean over seeds). Surplus is buyer surplus in pricing and $q - \tau$ in recommendation. Synth. strat. utility is approval minus manipulation cost on the synthetic population (Adult/German are in Table 5). Green: best surplus in each extraction column, and the least-predictable effective recsys policy. Blue: extractor lock-in (greedy accuracy) or the weak-dilemma peak (greedy still wins on strategic utility).

Policy	Pricing		Recsys		Synth. strat.	
	surplus	acc.	surplus	acc.	util.	acc.
Greedy	0.018	0.999	0.052	0.704	0.517	0.304
ε-greedy	0.011	0.954	0.051	0.489	0.500	0.285
Boltzmann 0.2	−0.021	0.665	0.024	0.169	0.435	0.237
Boltzmann 0.5	−0.068	0.595	−0.043	0.134	0.402	0.214
UNOP α=0.05	**0.042**	0.962	0.126	0.315	0.496	0.253
UNOP α=0.15	0.040	0.807	**0.148**	0.340	0.458	0.276
UNOP α=0.3	0.022	0.735	0.110	0.435	0.409	0.220
Aware-UNOP α=0.15	0.041	0.785	0.124	0.241	0.442	0.240
Uniform	−0.133	0.499	−0.094	0.107	0.368	0.199

Table 4: Same runs as Table 3. Predictor log loss is higher when the policy is less readable, and optimizer payoff V_O is lower when there is less extraction in pricing and recommendation. In the synthetic strategic setting, V_O is institution accuracy.

Policy	Pricing		Recsys		Synth. strat.	
	nll	V_O	nll	V_O	nll	V_O
Greedy	0.026	0.479	0.847	1.183	1.596	0.882
ε-greedy	0.198	0.446	1.453	1.129	1.599	0.880
Boltzmann 0.2	0.593	0.279	2.639	0.871	1.589	0.887
Boltzmann 0.5	0.673	0.346	2.657	0.951	1.608	0.888
UNOP α=0.05	0.082	0.396	1.694	0.962	1.589	0.895
UNOP α=0.15	0.312	0.241	2.057	0.623	1.561	0.902
UNOP α=0.3	0.408	0.189	1.873	0.533	1.587	0.891
Aware-UNOP α=0.15	0.309	0.238	2.080	0.764	1.598	0.897
Uniform	0.704	0.462	2.596	1.009	1.615	0.887

We first study a unique-best withholding extractor, and then a policy-aware slate optimizer that best-responds to each user policy (Section 4.2). In the unique-best design, once acceptance can be predicted, the platform withholds better organic content so that a high-margin item is the unique quality-best offer. Figure 8(a) and Table 3 show that greedy users are exactly the users this attack is designed for: surplus 0.052, promoted-item rate 0.858, accuracy 0.704. UNOP with $\alpha = 0.15$ raises surplus to 0.148 ($\approx 2.8\times$), cuts promoted exposure to 0.330, and drops accuracy to 0.340. Aware-UNOP is the least predictable effective policy (accuracy 0.241). Uniform again goes negative. Removing the individual-rationality floor changes the sign of pricing surplus (Appendix Ta-

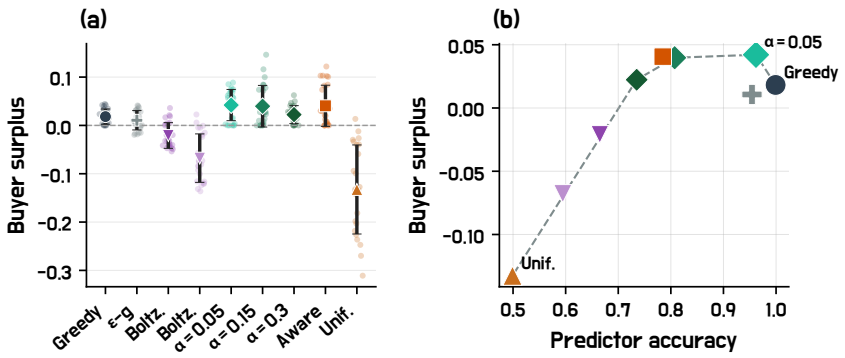


Figure 7: (a) Buyer surplus (seed strip + mean). (b) The same runs in the effectiveness-predictability plane (marker size = buy rate). Dashed path greedy → UNOP-α → uniform. Per-value scatter is in Appendix Figure 15. Boltzmann/uniform go negative because they buy above value.

ble 10): UNOP-$\alpha = 0.15$ becomes negative once the policy may buy above value, because those purchases teach the monopolist that the current price is still acceptable.

A.2 Discussion

The experiments do not say that unpredictability is always useful. They separate the cases in which a pure best response gives the optimizer a threshold, the cases in which there is no slack left to mix over, and the cases in which the counterpart is not trying to extract surplus at all. The practical question is where the α-optimal set is large enough to mix over, and whether the other side gains by sitting on that threshold.

On MovieLens-100K (Figure 8(b), all 943 users and 6601 user-policy jobs), five-star ratings often leave a single top item in the α-optimal set, so UNOP nearly reduces to greedy. Surplus is 0.189 for greedy, 0.187 for UNOP at $\alpha = 0.1$, and -0.115 for uniform. MovieLens-1M (6,040 users, 42,280 jobs) gives the same ordering (0.187, 0.185, and -0.130). Theorem 1 predicts this boundary. Coarse utilities remove the room to mix. Pricing and finely graded quality keep it.

A frequency predictor reaches accuracy 1.00 on greedy behavior in 30 steps. For UNOP with $\alpha = 0.15$, realized accuracy stays near 0.30 through 350 steps in 15 of 16 seeds (Appendix Figure 20), which is the Bayes accuracy of a mixture on several near-optimal actions, not a failure to identify the mixture (Proposition 1). ε-greedy still concentrates, because 90% of its mass stays on one action. Recommendation surplus peaks near $\alpha = 0.10$ and then falls as skipping becomes excessive (Appendix Figure 12). With about 30% UNOP users, a shared margin-seeking optimizer's accuracy falls from 0.460 on a purely greedy population to 0.219, and greedy users' surplus rises from 0.143 to about 0.17 (Appendix Figure 20, right). The optimizer no longer settles on one threshold, so part of the gain reaches users who do not change policy.

When the institution maximizes accuracy, lower predictability usually brings little or no utility gain. Across Adult and the other credit and marketing sets in Table 5, greedy stays ahead on utility. The counterpart's objective decides whether being readable is costly.

The same loop appears when an assistant learns which tasks, queues, or revisions a person will accept. That map helps delegation, and it can also steer later work toward the most readable person. Keeping several acceptable ways to carry out a task, when real slack exists, is the practical form of UNOP. Proposition 3 still applies when the counterpart is cooperative.

B Proofs

The arguments below concern support, security level, coordination tax, demand curves, and Dirichlet concentration.

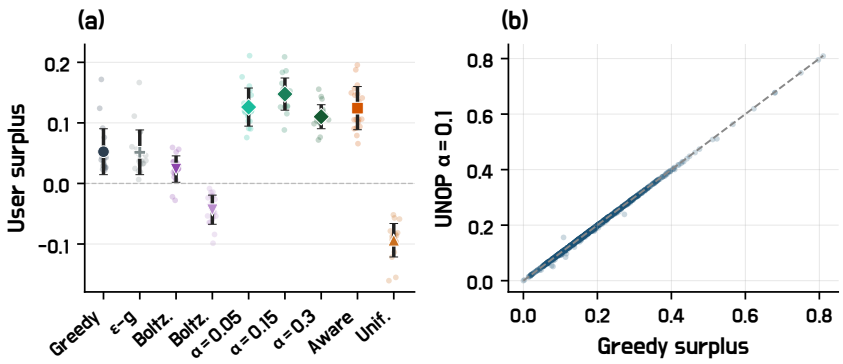


Figure 8: (a) Synthetic margin-seeking recommender: user surplus (seed strip + mean). (b) Complete MovieLens-100K, per-user greedy vs. UNOP scatter. Promoted-item rate versus surplus is in Appendix Figure 17.

Proof of Proposition 1. The hit rate of a predictor that names an action $a^\star$ is $\pi(a^\star) \le \max_a \pi(a)$, and the bound is achieved by naming a mode. For the log loss,

$$\mathbb{E}_{a\sim\pi}[-\log P(a)] = -\sum_a \pi(a)\log\pi(a) + \sum_a \pi(a)\log\frac{\pi(a)}{P(a)} = H(\pi) + \mathrm{KL}(\pi\|P).$$

The KL of two distributions is zero if and only if they agree π-almost everywhere. Off the support the value of P does not enter the expectation, which is why the statement is restricted to $\mathrm{supp}(\pi)$. Clipping at 10^{-12} is only a numerical guard in the experiments. The identity itself is stated for strictly positive probabilities on the support.

Proof of Proposition 2. Write the human payoff matrix as U, with rows indexed by a and columns by pure optimizer actions o. The security set is the nonempty polytope

$$\mathcal{S} = \{\pi : \pi \ge 0,\ \mathbf{1}^\top\pi = 1,\ U^\top\pi \ge v^\star\mathbf{1}\}.$$

If $\min_o u_H(\pi, o) \ge v^\star - \varepsilon$, then every violated security inequality has residual at most ε: $\|[(v^\star\mathbf{1} - U^\top\pi)_+]\|_\infty \le \varepsilon$. Hoffman's error bound for a fixed feasible system of linear equalities and inequalities (Hoffman, 1952) gives a finite constant C_G such that $\inf_{q\in\mathcal{S}} \|\pi - q\|_1 \le C_G\varepsilon$. Since total variation is one half of the ℓ_1 distance, the displayed claim follows with $\kappa_G = C_G/2$. This constant depends on the payoff matrix. The proposition does not assert a uniform constant over all games.

For a pure action a, an exact best responder realizes $\min_o u_H(a, o)$, which is strictly below $v^\star$ unless a itself is a security strategy. In matching pennies with payoffs $\{+1, -1\}$, $\mathcal{S} = \{(1/2, 1/2)\}$ and a policy $(1/2 + d, 1/2 - d)$ has security value $-2|d|$ and total-variation distance $|d|$ from $\mathcal{S}$. Thus value within ε requires $|d| \le \varepsilon/2$, ruling out a fixed peak as $\varepsilon \to 0$. Rock, paper, scissors has the analogous unique uniform security strategy. This is the precise, game-dependent sense in which effective play in our zero-sum examples must be mixed.

Proof of Proposition 3. Let $u^\star = u(a^\star, o^\star)$ and $\bar{u} = \max_{a\ne a^\star, o} u(a, o)$. Uniqueness of the global maximizer implies $\bar{u} < u^\star$. The point mass on $a^\star$, paired with $o^\star$, attains $u^\star$. For any π with $\pi(a^\star) = p$ and any fixed partner action o,

$$\mathbb{E}_{a\sim\pi} u(a, o) \le p\,u^\star + (1 - p)\bar{u}.$$

The inequality therefore also holds after maximizing the left-hand side over the partner's actions. Relative to $u^\star$, every $p < 1$ loses at least $(1 - p)(u^\star - \bar{u})$. Finally, grouping $a^\star$ against all other actions gives $H(\pi) \le \mathrm{h}_2(p) + (1 - p)\log(|A| - 1)$. The right-hand side tends to zero as $p \to 1$, so for each $h > 0$ the constraint $H(\pi) \ge h$ implies $p \le p^\star(h) < 1$.

Proof of Theorem 1. (1) The feasible set of (WC_α) is $\Delta(\mathcal{A}_\alpha(u))$. Shannon entropy $H(\pi) = -\sum_a \pi(a)\log\pi(a)$ is strictly concave on this simplex and is uniquely maximised by the uniform distribution (Jensen). Thus $\mathrm{Unif}(\mathcal{A}_\alpha)$ is the unique maximiser, $H = \log|\mathcal{A}_\alpha|$, and $u(a) \ge u^\star - \alpha$

on that set implies $\mathbb{E}_\pi[u] \geq u^\star - \alpha$. Moreover,

$$\max_{a \in \mathrm{supp}(\pi)} (u^\star - u(a)) \leq \alpha \quad \Longleftrightarrow \quad \mathrm{supp}(\pi) \subseteq \mathcal{A}_\alpha(u).$$

Hence any positive mass outside $\mathcal{A}_\alpha$ violates the advertised per-action guarantee. It need not violate an expected-loss constraint when that mass is small. (WC_α) deliberately enforces the stronger support-wise condition.

(2) The Lagrangian for $\max_\pi \pi \cdot u + \tau H(\pi)$ over $\Delta(A)$ gives the Gibbs distribution (Jaynes, 1957; Cover & Thomas, 2006): $\pi(a) \propto e^{u(a)/\tau}$. The map $\tau \mapsto H(\pi_\tau)$ is continuous and onto $(H(\mathrm{Unif}(\arg\max u)), \log|A|]$ for $\tau \in (0, \infty]$, so every (SH_h) with h in that interval is attained at some $\tau > 0$. Uniqueness follows from strict concavity of entropy. At $h = 0$ the feasible set includes every mass on $\arg\max u$. The leak bounds equation 6 follow from $Z \leq |A| e^{u^\star/\tau}$ and the existence of some $a \notin \mathcal{A}_\alpha$ with $u(a) = u^\star - \alpha - \gamma$ for the lower bound, and from $Z \geq |\mathcal{A}_\alpha| e^{(u^\star - \alpha)/\tau}$ together with $u(a) \leq u^\star - \alpha - \gamma$ off $\mathcal{A}_\alpha$ for the upper bound.

Proof of equation 10. On $\Delta(\mathcal{A}_\alpha^{\mathrm{IR}})$ the map $\pi \mapsto H(\pi) - \mathbb{E}_\pi[\log(P + \varepsilon)]$ is strictly concave. The Lagrangian $\sum_a \pi(a) \log \frac{1}{\pi(a)(P(a)+\varepsilon)} + \lambda\big(1 - \sum_a \pi(a)\big)$ has stationary point $\log \frac{1}{\pi(a)(P(a)+\varepsilon)} - 1 = \lambda$, hence $\pi(a) \propto (P(a) + \varepsilon)^{-1}$. The same point is the unique I-projection of $\nu_P \propto (P + \varepsilon)^{-1}$ onto the face, which is equation 11.

Proof of Theorem 2. The posterior mean is $P_t(a) = (n_a + 1)/(|A| + t)$. The strong law gives $n_a/t \to \pi(a)$ almost surely, hence $P_t \to \pi$. On $\mathrm{supp}(\pi)$ the map $P \mapsto \mathrm{KL}(\pi \| P)$ is continuous for P bounded away from zero, which holds for large t on that support, so the KL vanishes and the maximum coordinate converges to $\max_a \pi(a)$. For a δ-peak on $a^\star$ with $\delta > 1/2$, Hoeffding gives $\mathbb{P}(n_{a^\star} \leq t/2) \leq \exp\big(-2t(\delta - 1/2)^2\big)$. On that event $a^\star$ is the unique count maximizer. For uniform-on-k, $\mathbb{E}[\max_a n_a] = t/k + \Theta\big(\sqrt{(t \log k)/k}\big)$ for fixed k, so

$$\mathbb{E}\big[\max_a P_t(a)\big] = \frac{t}{k(|A| + t)} + \Theta\left(\frac{\sqrt{(t \log k)/k}}{|A| + t}\right),$$

which tends to $1/k$.

Proof of Proposition 4. For $p < v - \alpha$, buy is the unique α-optimal individually rational action, so $q = 1$. For $p \in [v - \alpha, v]$, both skip and buy lie in the α-set, and the constant-band rule sets $q = \rho$. For $p > v$, buy is individually irrational, so $q = 0$. Sure-demand revenue is therefore maximized uniquely at $p_\alpha^-(v)$, with value $p_\alpha^-(v)$, whenever that price exists. Every band price earns at most $\rho\, p_\alpha^+(v)$. If $p_\alpha^-(v) > \rho\, p_\alpha^+(v)$, every band price, and every price above v, has strictly lower revenue, so the unique revenue maximizer is $p_\alpha^-(v)$. On a grid of spacing Δ, $v - \alpha - \Delta < p_\alpha^-(v) < v - \alpha$, hence buyer surplus $v - p_\alpha^-(v)$ lies in $(\alpha, \alpha + \Delta]$. Any two rates that keep the seller below the band induce this same price and the same surplus. Let $r^\star = p_\alpha^-(v)/p_\alpha^+(v)$. Binary entropy on $(0, 1)$ increases toward $1/2$. If $r^\star > 1/2$, then $\rho = 1/2$ lies in that set and uniquely maximizes entropy there. If $r^\star \leq 1/2$, every such rate is strictly below $1/2$, entropy is increasing on $(0, r^\star)$, and the upper value $r^\star$ is not reached. A deployment that wants a maximizer can impose a revenue margin $\delta > 0$, replacing equation 8 by $p_\alpha^-(v) \geq \rho\, p_\alpha^+(v) + \delta$. The resulting set is closed and the entropy maximizer is $\min\big(1/2, (p_\alpha^-(v) - \delta)/p_\alpha^+(v)\big)$ whenever that quantity is positive.

If $v = 2\alpha$ and $v \in \mathcal{P}$, then $p_\alpha^+(v) = v$ and band revenue at $\rho = 1/2$ is $v/2 = \alpha$, while every sure-demand price is strictly below $v - \alpha = \alpha$. Condition equation 8 fails. The seller can prefer the band, including $p = v$, where buyer surplus is zero. Thus $\alpha \leq v/2$ does not imply that the seller stays below the band under seller-favorable tie-breaking.

C SIMULATOR ALGORITHMS

Algorithm 1 is the human's one-round map. Everything that surrounds it in the experiments is an adaptive optimizer that (i) posts a menu, (ii) observes a_t, (iii) updates a predictor or a bandit count, and (iv) posts a new menu. The three loops below are those optimizers. They match

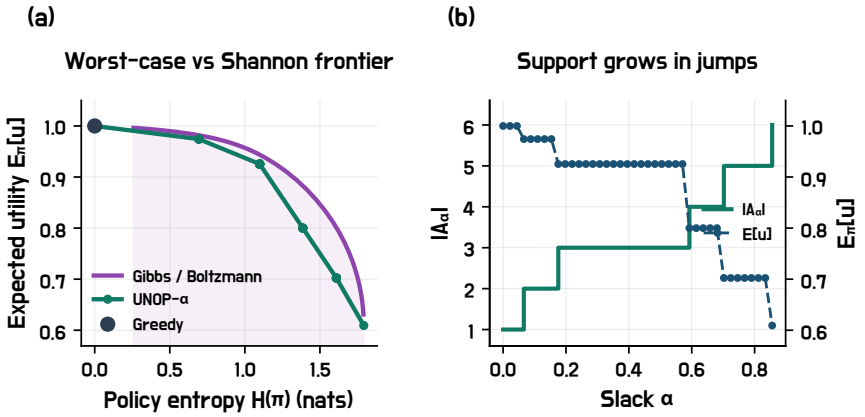


Figure 9: (a) Two Pareto curves on a six-action utility vector. Gibbs is optimal for a Shannon constraint and immediately puts weight outside the near-best set. UNOP is optimal for a worst-case loss constraint. (b) $|\mathcal{A}_\alpha|$ jumps when the next action enters, while expected utility steps down.

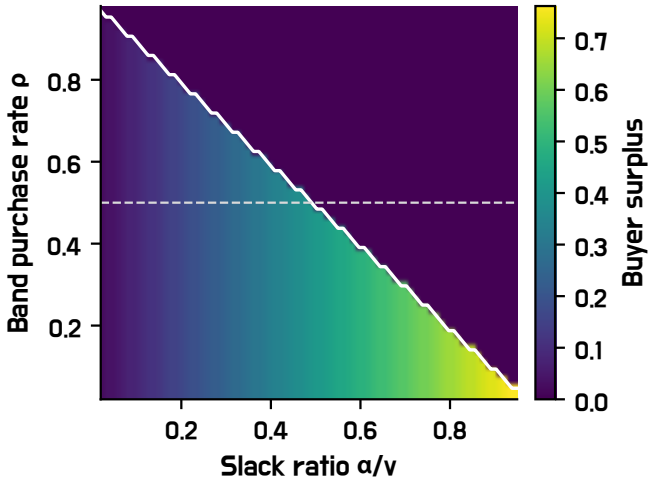


Figure 10: Buyer surplus for a constant-band response on an 81-point grid at $v = 0.8$. The white curve is the open boundary $\rho = r^\star$. Seller-favorable tie-breaking does not treat the boundary itself as keeping the seller out. The dashed line is $\rho = 1/2$. To the right of their intersection, uniform mixing fails equation 8, and the highest entropy of the rates that keep the seller below the band is not reached.

`experiments/simulate.py`. They specify the individual-rationality floor, the 5% random slip, and the unique-best filler-item construction.

Two implementation details that Algorithm 1 compresses. First, under exact arithmetic the IR intersection is nonempty whenever the reservation set is nonempty, so equation 9 reduces to intersecting $\mathcal{A}_\alpha$ with $\{a : u(a) \geq \underline{u}\}$. If floating-point noise empties that set, the code falls back to $\mathcal{A}_\alpha$. Second, Aware-UNOP uses $\varepsilon = 10^{-6}$ in the $(P + \varepsilon)^{-1}$ weights. Both choices are implementation details. They are what make the released rows match the algorithms below.

What is being predicted. In bimatrix games the predictor P_t is a distribution on the human's action set, and the optimizer chooses an exact best response to that prediction. In posted pricing there is no recommendation list: the menu is the posted price, and P_t is a Bernoulli on $\{\text{skip}, \text{buy}\}$ at that price, maintained as a Beta posterior. In recommendation P_t is an empirical distribution on the current recommendation list plus skip. Lock-in experiments use the frequency predictor of Algorithm 2 on a synthetic six-action menu and declare t_{lock} at the first time a length-30 rolling accuracy exceeds 0.7.

The 5% random slip is there so that a pure profile cannot freeze the dynamics by accident. It is also why greedy in matching pennies finishes near -0.14 rather than at -1: identification is fast,

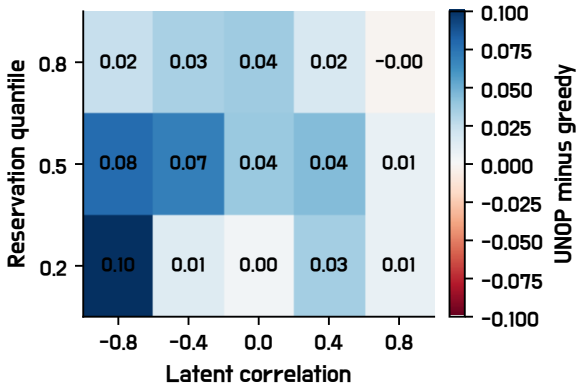


Figure 11: Policy-aware beam search of width 50 on catalogs drawn with a chosen Gaussian correlation, $n = 48$, slate size 8, $\alpha = 0.15$, 60 catalogs per cell. On 450 matching 12-item catalogs (30 per cell, both policies, 900 policy and catalog rows) the same width matches exact enumeration in every case. Entries are mean user-surplus differences, UNOP minus greedy. The horizontal axis is the correlation used to draw the pair, which is not the sample Pearson correlation of the realized quality and margin. The gain is largest at strong negative correlation. At strong positive correlation the paired interval covers zero. Paired intervals, medians, and positive fractions are in `supplement/sprint/results/recsys_phase_ci.json`. Quality ranking is negative in every cell.

Algorithm 2 Bimatrix closed loop (frequency / Markov / softmax predictor)

Require: game (U_H, U_O), human policy π, horizon T, predictor P
$\hat{\sigma} \leftarrow \mathrm{Unif}(O)$ ▷ empirical mix of optimizer actions
for $t = 1, \dots, T$ **do**
  $\hat{u} \leftarrow U_H \hat{\sigma}$ ▷ mean-based human, equation 2
  $a_t \sim \pi(\cdot \mid \hat{u}, P_t)$ ▷ Algorithm 1, or a baseline
  $o_t \leftarrow \arg\max_o (U_O^\top P_t)_o$, with a 5% uniform random slip
  observe payoffs $U_H(a_t, o_t)$, $U_O(a_t, o_t)$
  update P_{t+1} from a_t. $\hat{\sigma} \leftarrow$ empirical mix of $o_{1:t}$
end for
return post-burn-in means of u_H, u_O, $\mathbf{1}\{\arg\max P_t = a_t\}$, and $H(\pi_t)$

not instantaneous, and the opponent occasionally plays off the punishing action. Reported bimatrix means drop the first quarter of the horizon.

Algorithm 3 UCB posted-price monopolist

Require: price grid $\{p_j\}_{j=1}^m$, hidden value v, buyer policy π, horizon T
$s_j \leftarrow 1$, $f_j \leftarrow 1$ for all j ▷ Beta-Bernoulli prior on $\mathbb{P}(\text{buy} \mid p_j)$
for $t = 1, \dots, T$ **do**
  $\hat{b}_j \leftarrow s_j/(s_j + f_j)$
  $\mathrm{UCB}_j \leftarrow p_j \cdot \mathrm{clip}\big(\hat{b}_j + \sqrt{2\log(t+2)/(s_j + f_j)},\, 0, 1\big)$
  post $p_t \leftarrow p_{j^\star}$ with $j^\star = \arg\max_j \mathrm{UCB}_j$
  $\hat{b}_{j^\star} \leftarrow s_{j^\star}/(s_{j^\star} + f_{j^\star})$. $P_t \leftarrow (1 - \hat{b}_{j^\star},\, \hat{b}_{j^\star})$
  buyer utilities $u \leftarrow (0,\, v - p_t)$. $a_t \sim \pi(\cdot \mid u, P_t)$ with reservation 0
  if $a_t =$ buy **then**
    $s_{j^\star} \leftarrow s_{j^\star} + 1$
  else
    $f_{j^\star} \leftarrow f_{j^\star} + 1$
  end if
end for
return post-burn-in buyer surplus, posted price, buy rate, predictor accuracy

Algorithm 3 instantiates Proposition 4. A greedy buyer produces a step-function demand curve, so the UCB scores concentrate on the last grid cell below v and surplus is about one grid step. An IR-UNOP buyer produces the three-piece demand equation 7, so the same scores cannot treat the coin-flip band as sure demand. Thompson sampling replaces the UCB bonus by a Beta draw

(Appendix H.3). The qualitative gap survives. Turning the reservation off is Appendix F.4: the buyer then sometimes buys above value, the success count s_j is incremented at $p > v$, and surplus changes sign.

Algorithm 4 Margin-seeking unique-best recommendation list

Require: catalog $(q_i, m_i)_{i=1}^n$, list size k, reservation τ, user policy π, horizon T
$N_i \leftarrow 1, C_i \leftarrow 1$ for all i ▷ counts when i was the unique-best offer
for $t = 1, \dots, T$ **do**
  $\hat{p}_i \leftarrow C_i/N_i$. $\mathrm{UCB}_i \leftarrow \sqrt{2\log(t+2)/N_i}$
  $\mathrm{profit}_i \leftarrow m_i \cdot \mathrm{clip}(\hat{p}_i + 0.7 \cdot \mathrm{UCB}_i,\, 0, 1.2)$
  offer $i^\star \leftarrow \arg\max_i \mathrm{profit}_i$ (uniform among items for the first 30 rounds, and with an 8% random slip thereafter)
  fill the remaining $k-1$ slots with items of quality strictly below $q_{i^\star}$
  $P_t \leftarrow$ predictive distribution on the displayed actions (simulation: extractor forecast on $i^\star$, remaining mass on skip and filler items)
  user utilities $\leftarrow (q_{j_1}, \dots, q_{j_k}, \tau)$. $a_t \sim \pi(\cdot \mid u, P_t)$ with reservation τ
  $N_{i^\star} \leftarrow N_{i^\star} + 1$
  if the click is $i^\star$ **then**
    $C_{i^\star} \leftarrow C_{i^\star} + 1$
  end if
  surplus is $q_{\mathrm{clicked}} - \tau$ on a click and 0 on skip
end for

The unique-best construction is the extraction. A vanilla ranking UCB that also shows better organic items is a weaker adversary, because a greedy quality maximizer then clicks the organic item and never certifies the high-margin offer. MovieLens uses the same loop with q_{ui} from the official rating files, scaled to $[0, 1]$, and the same quantile reservation. When the five-star grid makes $\mathcal{A}_\alpha$ a singleton at the top star, Algorithm 1 returns greedy, which is the boundary in Theorem 1.

Baselines, for the same menus. Greedy is Algorithm 1 at $\alpha = 0$. ε-greedy puts mass $\varepsilon/|A|$ on every action, including those below reservation. Boltzmann is $\pi(a) \propto \exp(u(a)/\tau)$ on the whole of A. Uniform is $1/|A|$. Hedge is multiplicative weights on realized $\hat{u}_t$, with no support constraint. Under Algorithms 3 and 4 the extractor does not change: only the human's support does. Gibbs leak (Theorem 1) is visible in Algorithm 3 as soon as Boltzmann buys at $p > v$ and increments the success count s_j there.

Lock-in. On a six-action synthetic menu with a planted δ-peak or a uniform k-set, a frequency predictor $P_t(a) = (n_a + 1)/(|A| + t)$ is updated every round. t_{lock} is the first t at which a trailing window of 30 one-hot hits $\mathbf{1}\{\arg\max P_s = a_s\}$ has mean at least 0.7. Theorem 2 is the concentration statement behind that curve: a δ-peak with $\delta > 1/2$ produces a unique posterior mode in $O(\log(1/\eta)/(\delta - 1/2)^2)$ samples, while a uniform mix on k near-ties keeps the posterior an $O(1/k)$-peak. ε-greedy still locks quickly because $1 - \varepsilon$ of its mass is a peak. Entropy of $\mathcal{A}_\alpha$, not a generic exploration rate, is what delays identification.

D Additional experimental details

All simulators are NumPy/SciPy/scikit-learn CPU code. We used 96 processes. Controlled suite (3240 bimatrix jobs, 216 pricing, 180 recsys, 144 synthetic strategic, 144 lock-in, 452 α-sweep rows, 132 population runs, 4508 recorded rows) completed with zero errors. Public source files: MovieLens-100K, all 943 users $\times$ 7 policies (6601 jobs, official $\geq$ 20 ratings). MovieLens-1M, all 6,040 users $\times$ 7 (42,280 jobs). UCI Adult, source population 48,842, $d = 108$ after one-hot, 16 seeds $\times$ 9 policies ($T = 3000$, pretrained logistic accuracy 0.852). German Credit, source population 1,000, $d = 61$, 16 seeds $\times$ 9 ($T = 2000$, pretrained accuracy 0.785). FICO HELOC, source population 10,000, 16 seeds $\times$ 9. UCI Credit Card, source population 30,000, 16 seeds $\times$ 9. Give Me Some Credit, source population 150,000, pretrained accuracy 0.934, 16 seeds $\times$ 9. UCI Bank Marketing, source population 45,211, $d = 51$, pretrained accuracy 0.901, 16 seeds $\times$ 9. Bimatrix: $T = 1000$, 12 seeds, predictors $\{\mathrm{freq}, \mathrm{Markov}, \mathrm{logreg}\}$. Pricing: $T = 500$, 24 seeds, 21-price grid. Recsys: $T = 300$, 20 seeds, 48 items, list size 8. Synthetic strategic: $T = 700$, 16 seeds, 200 agents, 8 features, 5 manipulation levels. Lock-in: $T = 350$, 6 actions, 16 seeds.

Public files are loaded without file-level subsampling. MovieLens covers every user, while tabular interactions use the stated horizons.

Table 5 reports greedy vs. UNOP on every official public file we ran. MovieLens columns use $\alpha = 0.1$. Tabular strategic-classification rows use $\alpha = 0.15$.

Table 5: Public complete datasets. N is users for MovieLens and records for the tabular sets. Utility is user surplus on MovieLens and agent payoff, approval minus gaming cost, on the strategic-classification sets. MovieLens uses $\alpha = 0.1$ because the ratings have almost no slack. Tabular rows use $\alpha = 0.15$, where greedy is still ahead on utility and UNOP is less readable. The Adult $\alpha = 0.05$ sensitivity is utility/accuracy $0.181/0.428$, reported in Appendix G.3.

Dataset	N	greedy util.	greedy acc.	UNOP util.	UNOP acc.
MovieLens-100K (complete)	943	0.189	0.415	0.187	0.418
MovieLens-1M (complete)	6040	0.187	0.419	0.185	0.425
UCI Adult (complete)	48842	0.178	0.482	0.136	0.295
German Credit (complete)	1000	0.757	0.618	0.722	0.328
FICO HELOC	10000	0.501	0.779	0.470	0.312
UCI Credit Card	30000	0.082	0.981	0.046	0.334
Give Me Some Credit	150000	0.005	0.938	-0.031	0.333
UCI Bank Marketing	45211	0.061	0.815	0.024	0.327

Matching pennies: greedy -0.146, UNOP-$\alpha = 0.15$ 0.000, Aware-UNOP $+0.055$. RPS: greedy -0.141, UNOP 0.00, Aware-UNOP $+0.053$. Zero-sum suite means remain -0.143 / 0.000 / $+0.054$. Coordination: greedy = UNOP = 1.95, uniform 1.00. Stag hunt: greedy = UNOP = 4.88, uniform 3.96. Chicken: UNOP-$\alpha = 0.15$ 1.82 vs. greedy 1.77. Inspection: UNOP-$\alpha = 0.05$ 0.96 vs. greedy 0.92. Prisoner's dilemma: all near-optimal policies collapse to defection against an adaptive counterpart (1.11), as expected.

In the margin-seeking recommender, greedy users click a promoted item in 85.8% of rounds after the early exploration. UNOP-$\alpha = 0.15$ in 33.0%. Aware-UNOP in 41.5%. Skip rates rise with α (0.064 greedy, 0.340 at $\alpha = 0.15$), which is the volume and surplus tradeoff in Figure 12.

If the human's $\hat{u}$ is u plus independent Gaussian noise of scale σ, the α-set is a noisy near-best set. For $\sigma \leq \alpha/2$, UNOP remains inside the true (2α)-optimal set with high probability on small action spaces, which is why a slightly larger α is a robust default.

Bimatrix results average frequency, Markov, and online-softmax predictors. The three-case pattern is stable across the three. Replacing margin-seeking unique-best recommendation lists with vanilla CTR-UCB ranking weakens extraction: without withholding, greedy quality-maximizers are not forced onto the margin.

D.1 Restatement of the claims

This section restates the claims of Appendix B in the notation of the main text. The later ablations are read against that payoff shape.

Start with a utility vector $u \in \mathbb{R}^{|A|}$ that the human actually believes, or at least acts as if they believe. The number $u^\star = \max_a u(a)$ is the one-step best-response value, and the set $\mathcal{A}_\alpha(u) = \{a : u(a) \geq u^\star - \alpha\}$ is the set of actions that are never more than α worse than that value. Theorem 1 separates two constraints. First, if every action that may be played lies in $\mathcal{A}_\alpha$, then among all such policies the uniform distribution is the unique entropy maximizer, and every such policy has expected utility at least $u^\star - \alpha$ no matter how the rest of the vector is filled in. Second, if the only constraint is that entropy be at least some h, then the unique expected-utility maximizer is a Gibbs distribution, and Gibbs distributions put positive mass on every action as soon as the temperature is positive. The first constraint is a worst-case loss constraint and the second is a Shannon constraint. They coincide only when the utility vector is essentially flat on a subset and $-\infty$ elsewhere, which is not the situation a buyer or a recommender user is in.

That distinction is not a taste in regularizers. In posted pricing, a Gibbs buyer with even a modest temperature will sometimes buy above value, because the softmax never quite hits zero on the dom-

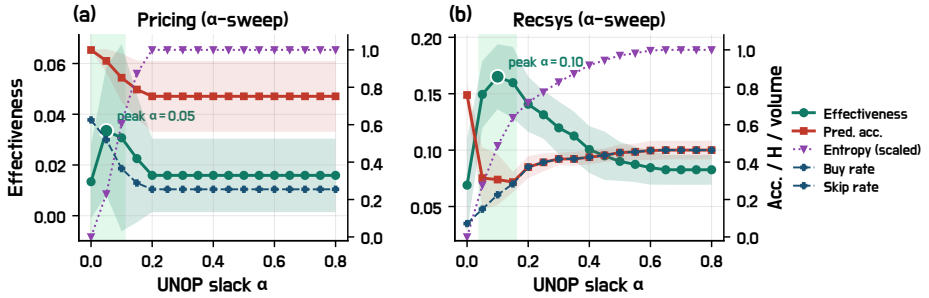


Figure 12: (a) Pricing α-sweep. (b) Recommendation α-sweep. Shared legend at right: effectiveness rises and then falls. Buy/skip volume and scaled entropy show how the mix changes. Predictability falls once the α-set is non-singleton.

inated action, and a learning monopolist treats those purchases as evidence that the posted price is still acceptable. UNOP with an individual-rationality floor of zero will not do that. The ablation in Appendix F.4 is the empirical counterpart of this paragraph: when we turn the floor off, UNOP starts to look like a slightly better-behaved Boltzmann policy, and the monopolist extracts it.

Proposition 2 is the matching-pennies version of the same payoff shape. If the opponent can identify a pure action and then choose the payoff-maximizing reply, then any policy that is a δ-peak for large δ is giving away the security level. Mixing on the near-optimal set is effective play in this case, and Aware-UNOP is a small refinement that puts less weight where the current forecast is already high, instead of mixing blindly. Proposition 3 is the opposite case. If the partner is trying to coordinate, a peak attains the Pareto-optimal payoff and mixing is a tax with no return from surplus-taking, so mixing is useful when there is room among near-best actions and the other side wants to take surplus.

Proposition 4 is the claim that most of the empirical work is about. A greedy buyer publishes a step-function demand curve, so any revenue maximizer that knows that curve posts in the last grid cell at or below v and buyer surplus is then on the order of one grid step. An individually rational UNOP buyer publishes a band, with sure demand below $v - \alpha$, purchase rate $1/2$ on $[v - \alpha, v]$, and sure refusal above v. Let $p^- = \max\{p \in \mathcal{P} : p < v - \alpha\}$ and $p^+ = \max\{p \in \mathcal{P} : v - \alpha \le p \le v\}$. The sure-demand price strictly keeps the seller below every price in that half-purchase band when $p^- > p^+/2$. On a grid of spacing Δ, $\alpha < v/2 - \Delta$ is a convenient sufficient condition for this inequality, and the buyer surplus is then $\Theta(\alpha)$ up to one grid step. Otherwise the proposition claims the demand curve, and a lower band rate is needed before the same surplus bound holds. Theorem 2 supplies the related concentration intuition, since a peak is identified in $O(\log(1/\eta))$ plays, whereas a uniform mixture on k near-ties leaves only an $O(1/k)$ posterior peak.

None of this requires the human to solve a long-horizon Stackelberg problem against a mean-based learner (Deng et al., 2019), because the policies we study are one-step mixers. A fully patient optimizer may still extract them. The experiments instead isolate the behavior of the named UCB and recommendation-list extractors.

D.2 Experimental protocol

This section records implementation choices that the main experiments section only summarizes. Every number in the paper comes from the released result files. The only additional runs are the ablations in Appendices F.4 to H.3.

What the human sees. In bimatrix games the human does not see the optimizer's current mixed action as a one-hot. They see the empirical mix of optimizer actions so far, form $\hat{u} = U_H \hat{\sigma}$, and then apply $\pi(\cdot \mid \hat{u}, P)$. That is a mean-based, not a fully Bayesian, human, and it is closer to how a person looking at a platform's recent prices actually has to reason. In pricing the menu is a single posted price, so $\hat{u} = (0,\ v - p)$ is exact up to the noise ablation. In recommendation the menu is the current recommendation list plus the skip action, with skip's utility equal to the reservation τ. We set τ to the 0.42 quantile of item quality in the synthetic catalog so that the outside option lies

inside the quality range. On MovieLens we keep the same construction after scaling ratings from $\{1, 2, 3, 4, 5\}$ into $[0, 1]$.

What the optimizer sees. Bimatrix predictors are frequency counts, a first-order Markov chain, or online softmax regression on the optimizer's own mix. The optimizer then chooses an exact best response to P_t, with a 5% uniform random slip so that the dynamics cannot freeze on a pure profile by accident. The pricing optimizer is per-price Beta-Bernoulli UCB on $p \cdot \mathbb{P}(\text{buy} \mid p)$. The recommender is a unique-best margin-seeking recommendation list: one offer is the quality-best item on the recommendation list, the rest are strictly worse filler items, and UCB is run on the offer's acceptance. A vanilla ranking UCB that does not withhold better organic content is a weaker extractor, because it does not sit on τ.

Burn-in. Reported means drop the first quarter of the horizon in bimatrix, recommendation, and strategic settings, and the first third in pricing, because we are interested in what an adaptive system does after it has had a chance to look at the human, not in the first few exploratory prices. Lock-in curves are the exception: they are the object of interest, so we keep the whole trajectory and define t_{lock} as the first time a length-30 rolling accuracy reaches at least 0.7.

Seeds and pooling. Bimatrix: 12 seeds $\times$ 3 predictors $\times$ 9 policies $\times$ 10 games. Pricing: 24 values of v. Recsys: 20 catalog draws. Strategic: 16 population draws. Public tables: MovieLens uses one job per user per policy on the official file. Adult, German, FICO HELOC, UCI Credit Card, Give Me Some Credit, and UCI Bank Marketing use 16 seeds with a shared pretrained logistic classifier, then online updates on manipulated features. Ablations use 8 to 24 seeds as listed in each subsection. Tables report means over seeds. Figures show error bars or seed scatters where useful. Reported effects are either large relative to seed noise or visibly null: greedy surplus near one grid step, UNOP surplus of order α, and MovieLens collapse.

Compute. All of the above is NumPy/SciPy/scikit-learn on CPU. We used 96 CPU processes. The original controlled suite is 4508 rows. Public complete files add 6601 MovieLens-100K jobs, $42{,}280$ MovieLens-1M jobs, and 144 runs (16 seeds $\times$ 9 policies) each for Adult, German, HELOC, Credit Card, Give Me Some Credit, and Bank Marketing. All completed with zero failures.

E BIMATRIX GAMES

E.1 BIMATRIX GAMES, ONE AT A TIME

Figure 13 is the predictability side of the same data. Table 6 reports mean human payoff in each of the ten games, averaged over the three predictors and twelve seeds. The ranking of policies is not a universal leaderboard. In matching pennies and rock, paper, scissors, greedy is a gift to the opponent and the mixed policies sit near or above zero. In coordination and stag hunt, greedy, small-α UNOP, and Aware-UNOP are tied at the top, and uniform is strictly worse. In prisoner's dilemma every near-optimal policy collapses to defection, which is exactly what an adaptive counterpart does to a one-step cooperator.

Table 6: Mean human payoff by bimatrix game, averaged over seeds and predictors. Mixing is mandatory in zero-sum, optional in general-sum, and costly in common-interest.

Game	Greedy	ε-greedy	UNOP α=0.15	Aware-UNOP	Uniform
Matching pennies	-0.146	-0.125	-0.000	0.055	0.011
RPS	-0.141	-0.113	0.000	0.053	0.012
Prisoner's dilemma	1.111	1.061	1.111	1.111	0.590
Chicken	1.765	1.616	1.818	1.871	1.382
Inspection	0.921	0.943	0.945	0.946	0.054
Shapley	0.339	0.345	0.391	0.412	0.333
Coordination	1.951	1.858	1.951	1.951	0.997
Stag hunt	4.878	4.789	4.878	4.878	3.959
Battle of the sexes	1.163	1.081	1.163	1.082	0.518
Pure coordination	0.965	0.902	0.964	0.964	0.329

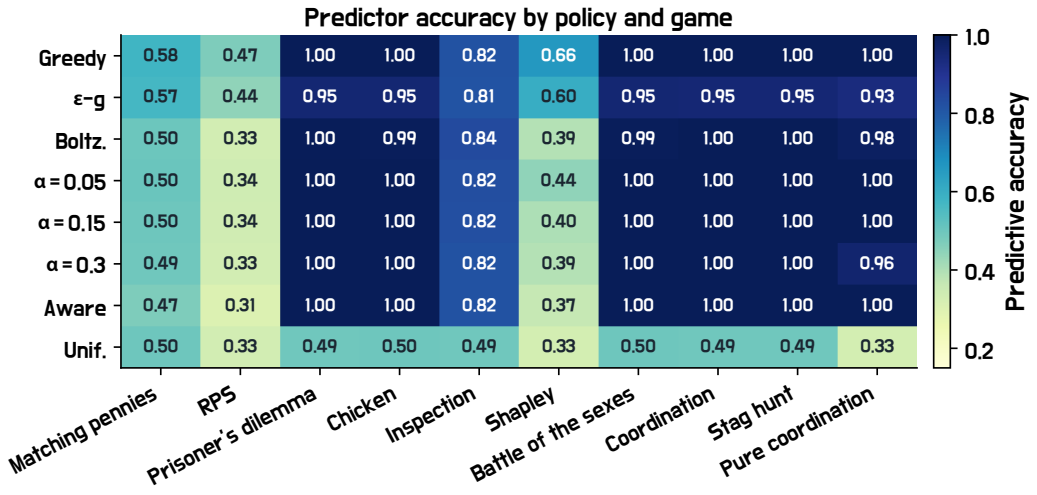


Figure 13: Predictive accuracy by policy in the rows and game in the columns. Each cell is the numeric mean accuracy. Darker is more readable. Greedy is readable. Uniform is not. UNOP lands in between only where the game actually has room at the chosen α.

A reasonable worry is that the three-case pattern comes only from using a frequency predictor. Figure 14 splits the same points by predictor. Frequency, Markov, and online softmax all produce the same qualitative cloud. Table 7 reruns greedy, UNOP-$\alpha = 0.15$, and Aware-UNOP at $T \in \{200, 500, 1000, 2000\}$ with a frequency predictor. In zero-sum games the mixed policies stay near the value as T grows, while greedy does not recover. In common-interest games the three policies that can coordinate stay coordinated. The three-case pattern is not a short-horizon illusion.

Table 7: Human payoff versus horizon, grouped by game class under a frequency predictor. Zero-sum mixing remains necessary. Common-interest mixing remains a tax.

Class	Policy	T=200	T=500	T=1000	T=2000
Zero-sum	Greedy	-0.004	-0.011	-0.008	-0.004
	UNOP α=0.15	0.035	-0.001	-0.002	0.009
	Aware-UNOP	0.021	0.017	0.033	0.013
General-sum	Greedy	1.206	1.196	1.197	1.201
	UNOP α=0.15	1.232	1.224	1.226	1.225
	Aware-UNOP	1.240	1.228	1.228	1.228
Common-interest	Greedy	2.625	2.607	2.596	2.595
	UNOP α=0.15	2.626	2.607	2.596	2.595
	Aware-UNOP	2.626	2.607	2.596	2.595

In matching pennies the human's actions are heads and tails, the optimizer wants to mismatch, and a greedy human who has a tiny preference for heads, or who simply broke a tie the same way twice, becomes a pure strategy that the optimizer can sit on. UNOP with α large enough to include both actions is just the mixed security strategy, which is why its payoff sits at 0.000. Aware-UNOP does a little better ($+0.054$) because it is not mixing independently of P_t: it puts extra mass on whichever action the predictor currently underweights, and is therefore anti-correlated with a slowly adapting opponent. The gain is empirical against that learner. It is not a guarantee against an exact best responder.

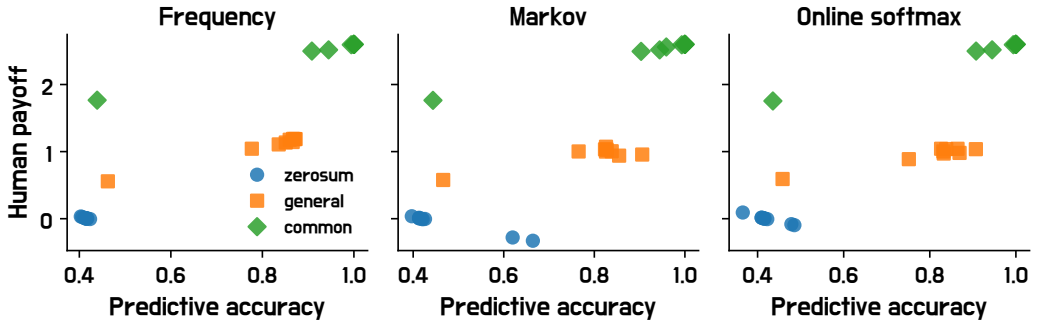


Figure 14: Effectiveness against predictive accuracy, split by the optimizer's predictor. Circles are zero-sum, squares general-sum, diamonds common-interest. The three-case pattern is not an accident of counting frequencies.

F Posted pricing

F.1 Posted pricing: where the surplus actually goes

The main-text pricing figure reports surplus and the effectiveness-predictability plane. Figure 15 shows the same runs as a scatter against the hidden value v. Greedy buyers lie on a tight band near zero surplus for every v: the monopolist found the threshold. UNOP buyers keep a wedge of surplus. Uniform and Boltzmann buyers go negative: they are sometimes paying more than v. Figure 16 is the time series: greedy posted prices climb toward v. UNOP stays below v. The hit rate should be compared with Bayes accuracy, not read as a failure to learn the mixture. Seed-level spreads appear as error bars in Figure 16. Table 3 reports means only, and the ordering is not an average of a bimodal mess.

F.2 A worked numerical example

Suppose $v = 0.70$ and the price grid is $\{0.00, 0.05, \ldots, 1.00\}$, so $\Delta = 0.05$. A greedy buyer purchases at every $p \le 0.70$ and refuses otherwise. After enough plays, UCB on revenue will prefer $p = 0.70$ to $p = 0.65$ once both empirical buy rates are near one, because $0.70 \cdot 1 > 0.65 \cdot 1$, and it will prefer $p = 0.70$ to $p = 0.75$ once the latter's empirical buy rate is near zero. Buyer surplus is then 0 when the posted price is exactly 0.70, or 0.05 if the optimizer sits one cell lower. In the main runs the mean leftover is 0.018, which is a fraction of a cell, as Proposition 4 said.

Now take UNOP with $\alpha = 0.15$ and reservation 0. At $p = 0.50$, buy gives 0.20 and skip gives 0, so $u^\star = 0.20$ and $u^\star - \alpha = 0.05$. Skip lies below that cut and buy is the unique α-optimal IR action, hence $\mathbb{P}(\text{buy}) = 1$. At $p = 0.60$, buy gives 0.10 and skip gives 0. Now $u^\star - \alpha = -0.05$, so both actions lie in $\mathcal{A}_{0.15}$ and both are IR, and the human mixes equally. At $p = 0.65$, buy gives 0.05 and skip gives 0. Both remain inside $\mathcal{A}_{0.15}$ and both are IR, so the mix is again equal. At $p = 0.75$, buy gives -0.05, which is excluded by the reservation even though a Boltzmann policy would still buy with positive probability. A seller who wants demand near 1 is pushed down to $p \le 0.55$, and the buyer keeps about 0.15.

F.3 Grid resolution and horizon

Tables 8 and 9 vary grid resolution and horizon. As the price grid goes from 11 to 41 cells at fixed $T = 500$, greedy surplus falls (0.031 at 11 cells, 0.006 at 41), which is $\Theta(\Delta)$. UNOP-$\alpha = 0.15$ remains above greedy at every resolution. Boltzmann and uniform stay negative. As T grows on the default 21-cell grid, greedy surplus is already small at $T = 150$ and does not grow. UNOP surplus rises (0.007 at 150, 0.061 at 800), because post-burn-in means for UNOP improve once early UCB exploration is behind it. UNOP remains above greedy at every horizon in the sweep.

The two sweeps should be read together. A finer grid is a larger action space for the monopolist: each additional cell needs samples before its revenue estimate is usable, so at fixed horizon a finer spacing leaves more leftover exploration and compresses surplus after the early rounds for every

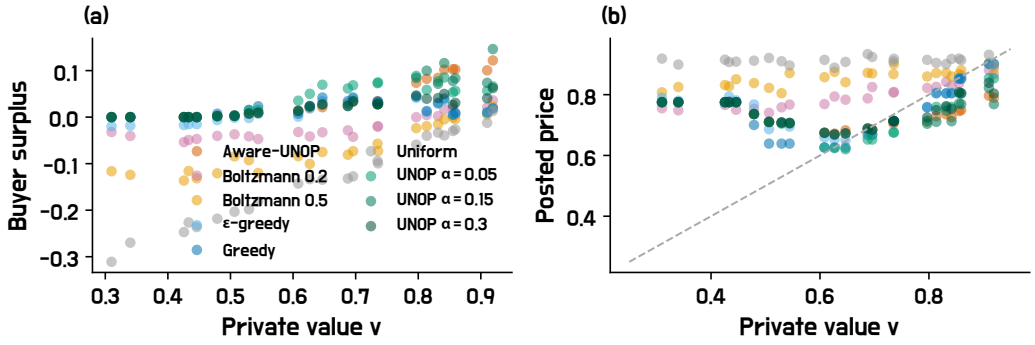


Figure 15: Left: buyer surplus against private value. Right: posted price against v, with the diagonal $p = v$ as a dashed reference. Greedy transactions hug the diagonal from below. UNOP leaves a visible gap. Boltzmann and uniform cross the diagonal.

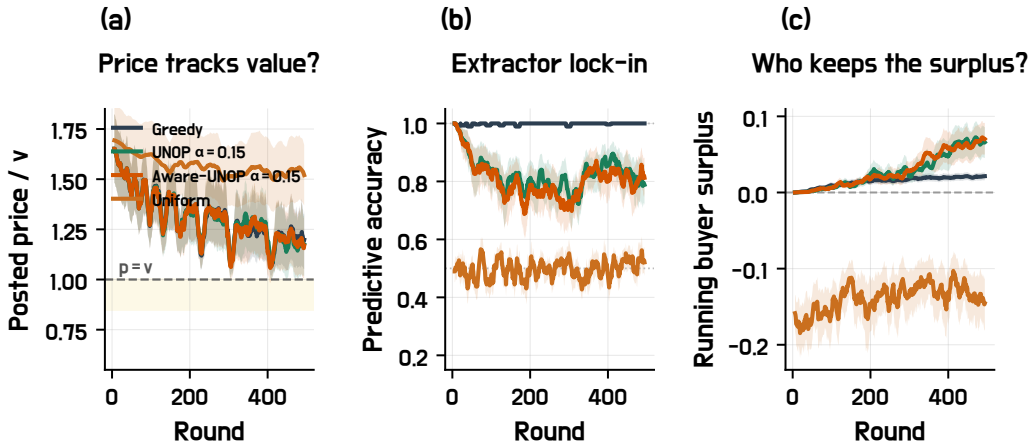


Figure 16: Pricing dynamics (rolling mean $\pm$ s.e.). (a) Posted price relative to v. (b) Predictor accuracy. (c) Running buyer surplus. Greedy prices climb toward v. UNOP stays lower. Accuracy for UNOP is bounded by the policy's Bayes accuracy.

policy. Extending T gives the same optimizer time to finish that exploration, which is why surplus recovers as T grows even though the spacing is held fixed. In short, finite-horizon surplus tracks how far the extractor has gotten through its action set, not only how wide the human's α-band is. The practical boundary is that UNOP still beats greedy on every grid we tried, but a very fine menu paired with a short horizon will look closer to the greedy grid step than the large-sample $\Theta(\alpha)$ band.

Table 8: Buyer surplus versus the number of cells on the price grid.

Policy	11 cells	21 cells	31 cells	41 cells
Greedy	0.031	0.013	0.010	0.006
UNOP α=0.15	0.065	0.036	0.021	0.012
Boltzmann 0.2	-0.010	-0.022	-0.028	-0.030
Uniform	-0.140	-0.135	-0.143	-0.142

F.4 The individual-rationality ablation

UNOP as defined in Section 3 intersects $\mathcal{A}_\alpha$ with $\{a : u(a) \geq \underline{u}\}$ whenever that intersection is nonempty. In pricing $\underline{u} = 0$ (skip). In recommendation $\underline{u} = \tau$. Table 10 turns that intersection off and reruns the same policies.

Table 9: Buyer surplus versus horizon T. Greedy tracks the grid step. UNOP remains above it.

Policy	T=150	T=300	T=500	T=800
Greedy	0.006	0.011	0.014	0.016
UNOP α=0.15	0.007	0.020	0.042	0.061
Boltzmann 0.2	-0.039	-0.030	-0.020	-0.011
Uniform	-0.150	-0.139	-0.135	-0.126

Without the floor, UNOP-$\alpha = 0.15$ in pricing goes from $+0.036$ to -0.018, and Aware-UNOP goes from $+0.039$ to -0.017. This is a separate smaller-seed ablation, while Table 3 reports 0.040 with the floor on. That sign change is the whole argument for having a floor. Mixing into buy-above-value raises one-step entropy and also hands revenue to a learning monopolist. Greedy, ε-greedy, Boltzmann, and uniform are unaffected by the switch, because they never used the floor. In recommendation the same switch costs about 0.04 of surplus for UNOP.

Table 10: Effectiveness with and without the individual-rationality floor. The floor is unused for greedy and for policies that already mix everywhere. For UNOP it separates preserved surplus from a teaching signal.

Policy	Pricing surplus		Recsys surplus	
	IR on	IR off	IR on	IR off
Greedy	0.012	0.012	0.050	0.050
UNOP α=0.15	0.036	-0.018	0.136	0.094
Aware-UNOP	0.039	-0.017	0.105	0.055
Boltzmann 0.2	-0.022	-0.022	0.037	0.037
Uniform	-0.135	-0.135	-0.079	-0.079

The floor is a modeling claim about extractors, not a claim that people compute a reservation and then mix above it. If a policy mixes into actions it would never take against a fixed menu, those actions teach the extractor where the cutoff still is. The reservation floor is what blocks that lesson.

F.5 IR-CONSTRAINED BASELINES

A natural question is whether the gap between UNOP and Boltzmann/ε-greedy is entirely the reservation floor. Table 11 applies the same floor to those baselines: ε-greedy and Boltzmann are renormalized on $\{a : u(a) \geq \underline{u}\}$ whenever that set is nonempty. In pricing the floor turns Boltzmann's negative surplus positive, but IR-Boltzmann 0.2 still trails UNOP-$\alpha = 0.15$ (0.028 vs. 0.040). IR-ε-greedy remains near greedy. In recommendation, IR-Boltzmann 0.2 reaches 0.113 surplus, still below UNOP at 0.148. The floor removes teaching-signal losses. Uniform mixing on the near-optimal set is what keeps the extractor off the threshold.

Table 11: Unconstrained vs. IR-constrained baselines on the same pricing and recsys loops as the main text ($T = 500/300$).

Policy	Pricing		Recsys	
	surplus	acc.	surplus	acc.
Greedy	0.018	0.999	0.052	0.704
ε-greedy	0.011	0.954	0.051	0.489
IR-ε-greedy	0.017	0.970	0.058	0.564
Boltzmann 0.2	-0.021	0.665	0.024	0.169
IR-Boltzmann 0.2	0.028	0.766	0.113	0.389
Boltzmann 0.5	-0.068	0.595	-0.043	0.134
IR-Boltzmann 0.5	0.025	0.738	0.090	0.440
UNOP α=0.15	0.040	0.807	0.148	0.340
Uniform	-0.133	0.499	-0.094	0.107

F.6 MISSPECIFIED UTILITIES, AND HOW MUCH α HAS TO ABSORB

The theory treats u as known. In any honest deployment the human has $\hat{u} = u + \xi$. Table 12 adds independent Gaussian noise of scale $\sigma \in \{0, 0.02, 0.05, 0.10, 0.20\}$ to the utility vector before the policy is applied. In pricing, UNOP-$\alpha = 0.15$ surplus falls from 0.032 at $\sigma = 0$ to -0.018 at $\sigma = 0.20$. The last of those is noise larger than α, so the α-set constructed from $\hat{u}$ is no longer a slightly wider true α-set. Raising α further at $\sigma = 0.20$ does not recover surplus in our runs: at the same noise, $\alpha \in \{0.25, 0.30, 0.40, 0.50\}$ stays negative and slowly worsens, because a wider noisy band absorbs more wrongly signed actions through the reservation check. When σ is small relative to α, a default α on the order of twice the plausible utility noise is safer than a default that tries to be as small as the worst-case bound would suggest. When σ exceeds that scale, the binding fix is a better $\hat{u}$, not a larger band.

Table 12: Effectiveness as a function of independent Gaussian noise on the human's utility vector. UNOP remains useful while σ is small relative to α. At $\sigma = 0.20$ a larger α does not restore pricing surplus.

Env.	Policy	σ=0.00	σ=0.02	σ=0.05	σ=0.10	σ=0.20
Pricing	Greedy	0.013	0.020	0.018	0.011	-0.012
	UNOP α=0.15	0.032	0.013	0.010	0.001	-0.018
	Uniform	-0.135	-0.136	-0.136	-0.136	-0.136
Recsys	Greedy	0.044	0.090	0.109	0.106	0.065
	UNOP α=0.15	0.129	0.106	0.101	0.094	0.054
	Uniform	-0.076	-0.081	-0.081	-0.081	-0.081

Recommendation is gentler, because a typical recommendation list has several items whose qualities are already close. Even there, $\sigma = 0.20$ is too much for $\alpha = 0.15$. We do not claim a sharp phase transition. We claim that a tiny α is not enough once noise exceeds slack, and that widening the band alone is not a recovery recipe.

G RECOMMENDATION AND PUBLIC DATA

G.1 MARGIN-SEEKING RECOMMENDATION LISTS, CATALOG SIZE, AND PROMOTED EXPOSURE

Figure 17 plots promoted-item click rate against user surplus for every synthetic recsys seed. Greedy seeds cluster at high promo and low surplus, UNOP seeds sit at lower promo and higher surplus, and uniform seeds occupy the high-promo, negative-surplus corner. Promoted exposure falls from 0.858 to 0.330 at $\alpha = 0.15$ not because we penalized ads, but because the items that survive as unique-best offers are less often the promoted ones when τ is a band.

Table 13 asks whether that cloud depends on having chosen 48 items and a recommendation list of 8. We crossed catalog sizes $\{24, 48, 72\}$ with recommendation-list sizes $\{4, 8, 12\}$ for greedy, UNOP-$\alpha = 0.15$, and uniform. Greedy stays poor across the grid. Uniform stays worse. UNOP stays better, and it likes larger recommendation lists, because a larger recommendation list is more likely to contain a second item inside $\mathcal{A}_\alpha$. A recommendation list of four with a unique-best offer and three filler items leaves little room: there may be nothing near-optimal to mix with except skip.

G.2 MOVIELENS, USER BY USER, ON THE COMPLETE FILES

The main text reports that on MovieLens, UNOP collapses to greedy because five-star ratings do not have slack. Figure 18 is the user-level answer on the official complete 100K file: the scatter hugs the diagonal, and the per-user difference is a spike at zero. Table 14 records the same collapse on MovieLens-1M ($6{,}040$ users) and the promoted-item rates on both files. Greedy and UNOP remain close. Uniform is much higher. This is the same collapse on the complete file, without subsampling.

Why is there no slack? Because a user who has rated twenty movies on a five-star grid often has a unique five-star title, and after scaling to $[0, 1]$ that title is 1.0 and the four-star titles are 0.75, so any $\alpha < 0.25$ leaves a singleton $\mathcal{A}_\alpha$. If a platform displayed utilities on a finer scale, such as

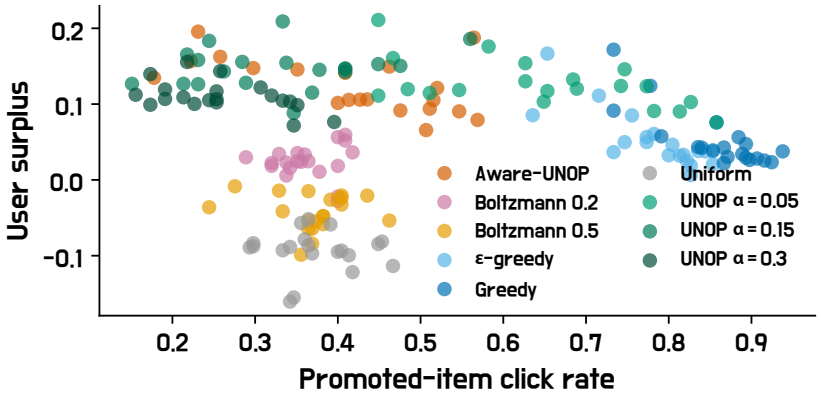


Figure 17: Each point is one synthetic recommender seed. Horizontal: fraction of clicks on promoted items. Vertical: user surplus $q - \tau$.

Table 13: User surplus versus catalog size and recommendation-list size. The ranking does not depend on the particular $(48, 8)$ pair used in the main text.

Policy	Slate	cat. 24	cat. 48	cat. 72
Greedy	4	0.039	0.037	0.039
	8	0.055	0.055	0.064
	12	0.111	0.070	0.087
UNOP α=0.15	4	0.125	0.092	0.052
	8	0.161	0.148	0.111
	12	0.302	0.198	0.152
Uniform	4	-0.091	-0.131	-0.154
	8	-0.027	-0.091	-0.127
	12	0.024	-0.049	-0.093

a continuous rating, a dwell-time proxy, or a willingness-to-watch slider, the same protocol would look like the synthetic recsys panel. The public ratings are coarse, and the reported result is the one those ratings support.

G.3 UCI Adult and German Credit

Strategic classification on real tables is the other public-data pillar. The institution starts from a logistic classifier trained on unmanipulated official records (Adult accuracy 0.852 on all $48{,}842$ rows after one-hot to $d = 108$, and German Credit 0.785 on all $1{,}000$ rows, $d = 61$), then continues with online logistic updates on the features that agents actually present. Agents may game a small set of numeric coordinates that are at least arguably effort-like: age, education-num, capital gain/loss, and hours on Adult, and duration, credit amount, age, and existing credits on German. The action is one of five intensities along $\mathrm{sign}(w)$ on those coordinates, at quadratic cost. The predictor is a frequency model over those five intensities. This is not a claim that people in the 1994 Adult file actually gamed their hours-per-week against an adaptive classifier. It is a claim that if they did, with a linear institution that maximizes accuracy, mixing would reduce readability without raising utility.

Table 14: Complete MovieLens files: surplus collapse and promoted exposure. UNOP uses $\alpha = 0.1$.

File	N	greedy surplus	UNOP surplus	greedy promo	UNOP promo
MovieLens-100K	943	0.189	0.187	0.085	0.096
MovieLens-1M	6040	0.187	0.185	0.065	0.084

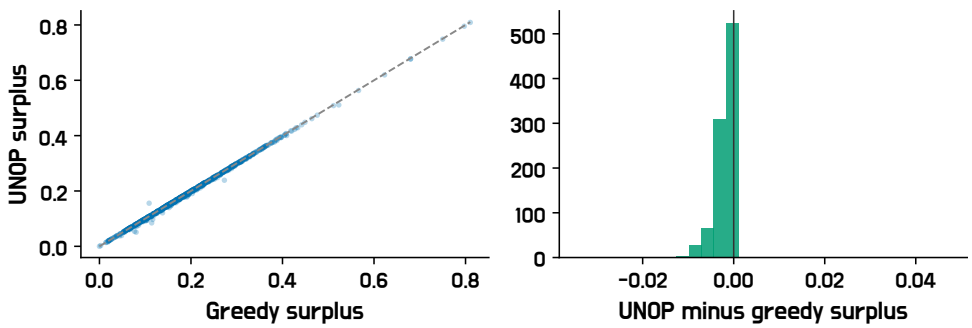


Figure 18: Complete MovieLens-100K, all 943 users. Left: per-user greedy versus UNOP. Right: per-user difference. The collapse is not an average over a hidden subpopulation.

Table 5 already reports agent utility and predictor accuracy on the complete UCI Adult and German Credit tables. Greedy is best or tied for best on utility. On Adult, UNOP-$\alpha = 0.05$ gives utility/accuracy 0.181/0.428, versus 0.178/0.482 for greedy. Larger α costs utility without an extraction-flavored reward. Mixing does reduce readability. It just does not buy the agent anything the accuracy-seeking institution was threatening to take.

Figure 19 (left) varies the quadratic cost coefficient on the synthetic strategic population. As cost rises, everyone games less and policy differences shrink. No cost in this family makes UNOP outperform greedy by a margin comparable to pricing. The counterpart maximizes accuracy rather than surplus.

H Lock-in and other extractors

H.1 Lock-in, populations, and the public-good reading

Theorem 2 is a Dirichlet multinomial calculation, and Figure 19 (right) is the empirical version as $k = |A|$ grows. This separate k-sweep uses $T = 400$. The main lock-in experiment uses $T = 350$. Greedy and ε-greedy hit 70% rolling accuracy at $t = 30$ for every k we tried. Uniform almost never hits it within the sweep horizon. UNOP at $\alpha = 0.15$ tracks uniform on that random utility draw once α admits several near-ties. The lock-in figure (Figure 20) used a vector whose near-ties were farther apart, so small α still leaves a peak.

Figure 20 also shows the population overlay. A shared margin-seeking optimizer trained on a mix of greedy and UNOP users has an uneven accuracy curve: highest on a pure-greedy population and lowest around a 30% UNOP minority. Greedy users' welfare rises when they are mixed with UNOP neighbors.

H.2 Two tradeoffs, Gibbs leak, and why temperature is the wrong choice

Figure 9 already contrasts the two tradeoffs on a six-action utility vector. The same qualitative gap (smooth Gibbs leak versus UNOP step support) appears on random vectors of dimension 4 through 16 in the dump `pareto.json`. Figure 21 makes that picture concrete. The left panel is the mean size of $\mathcal{A}_\alpha$ as a function of α. The right panel is the amount of Gibbs mass that sits outside $\mathcal{A}_\alpha$ as a function of α, for several temperatures. At $\tau = 0.2$ and $\alpha = 0.15$, more than a quarter of the Gibbs mass is already off the near-optimal set. Temperature is not a substitute for a hard support constraint. The two-panel α-sweep in Figure 12 is the corresponding empirical picture: effectiveness rises and then falls, and accuracy falls as soon as $\mathcal{A}_\alpha$ is non-singleton.

H.3 A different extractor: Thompson sampling

UCB is a particular way to turn uncertain demand into a posted price. Table 15 reruns the pricing environment with a Beta-Bernoulli Thompson monopolist that posts $\arg\max_p p \cdot \theta_p$ for $\theta_p \sim \mathrm{Beta}(a_p, b_p)$, and includes Hedge as an additional human baseline. The qualitative ranking does not flip: uniform is still a disaster, greedy is still extracted, UNOP still keeps surplus. The

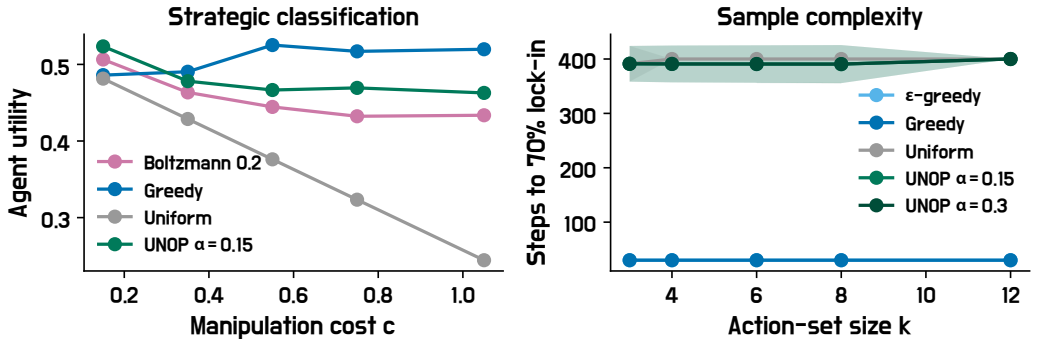


Figure 19: Left: synthetic strategic classification as the manipulation-cost coefficient c varies. Right: steps to 70% lock-in as the action-set size k varies. Greedy and ε-greedy lock in 30 steps at every k (the window length). UNOP and uniform do not lock within the horizon.

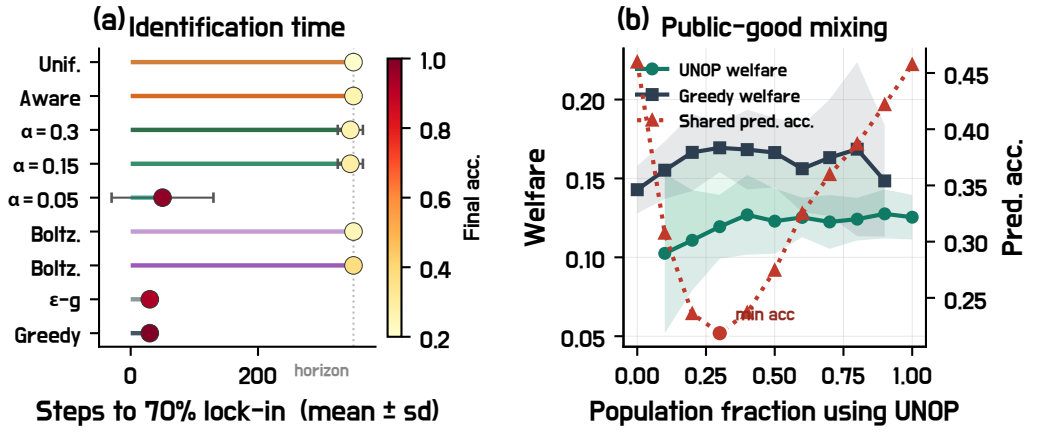


Figure 20: Left: steps to 70% lock-in (mean $\pm$ sd) with marker color $=$ final accuracy. Dotted line is the horizon. Right: welfare of greedy and UNOP users in a shared-optimizer population (shaded gap) and the extractor's accuracy (right axis, min marked). Heterogeneity is itself a defense.

UNOP surplus is larger against Thompson (0.139) than against UCB (0.036) in this parameterization, because Thompson is more willing to post inside a noisy band. Hedge is near greedy against UCB and slightly negative against Thompson, consistent with mean-based learning not being a defense against a posted-price extractor.

Table 15: Buyer surplus against a UCB monopolist and a Thompson monopolist, same price grid and values of v. The human-side ranking is stable. Hedge is not a substitute for UNOP.

Policy	UCB surplus	Thompson surplus
Greedy	0.012	0.022
UNOP α=0.15	0.036	0.139
Uniform	-0.135	-0.132
Hedge	0.003	-0.003

H.4 Putting the environments together

The environments do not share a utility scale, so a single scatter is unreadable. The main-text three-case figure, the pricing/recsys tables, and Table 5 already locate each setting. If UNOP were simply the random policy, it would sit with uniform. If it were simply greedy with a different name, it would sit with greedy. On pricing and recsys it sits up and left of greedy. On both MovieLens files it sits on top of greedy. On Adult and German it sits down and left: less readable, not more

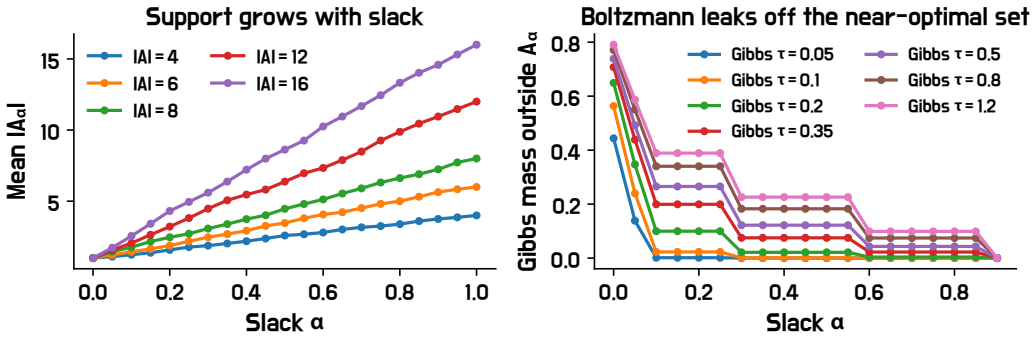


Figure 21: Left: mean size of the α-optimal set on random utility vectors. Right: Gibbs mass lying outside $\mathcal{A}_\alpha$. Temperature is not a substitute for a hard support constraint.

effective. Predictability is correlated with policy entropy across environments, but two policies with the same entropy can be very different for the human if one spends that entropy inside $\mathcal{A}_\alpha$ and the other spends it on dominated actions. Accuracy, optimizer payoff, and human payoff are therefore the reported outcomes. Entropy alone does not determine them.

H.5 A SECOND WORKED EXAMPLE: A MARGIN-SEEKING RECOMMENDATION LIST

Imagine four items with qualities $(0.90, 0.72, 0.55, 0.40)$ and a reservation $\tau = 0.60$. The platform can show three of them. A unique-best margin-seeking recommendation list that has identified τ will offer the item with quality 0.72 as the unique best (it is just above τ) and fill with 0.55 and 0.40 as filler items, because 0.90 would give the user a large surplus and a lower margin. A greedy user clicks 0.72 every time, surplus 0.12, and the platform has sat on τ. An UNOP user with $\alpha = 0.15$ sees utilities $(0.12, -0.05, -0.20, 0)$ for the three items and skip. The α-optimal IR set is $\{0.72\text{-item}, \text{skip}\}$, because skip is 0 and the offer is 0.12, which differ by less than 0.15, while the filler items are below reservation. The user therefore mixes equally between taking the offer and skipping. The platform now sees a 50% acceptance rate on an item that would have been a sure thing if τ were 0.60 with a greedy user, and UCB cannot treat 0.72 as an identified threshold. If the catalog had also included an organic 0.85 item on the recommendation list, that item would have entered $\mathcal{A}_\alpha$ as well, and the mix would have included a high-quality click. Withholding is how the platform tries to prevent that. Mixing on skip is how the user refuses to certify the withheld menu.

Uniform, on the same recommendation list, clicks the 0.40 item with positive probability, which is worse than skip, and the platform learns something even less helpful to the user: that junk is sometimes acceptable. Unpredictability without a floor is not a defense.

I LIMITATIONS AND FURTHER CHECKS

I.1 LIMITATIONS

The response-curve commitment effect applies when the platform optimizes against that curve, when feasible menus leave more than one acceptable action, and when the utility estimate places the set correctly. Noise large enough to move the estimated set removes the surplus, which is why the pricing gain and the recommendation gain should be read together with the noise and singleton checks in Section 4.2.

We list explicitly what this extra empirical work does not establish. We have not shown that UNOP cannot be exploited by a patient optimizer who knows the policy class and chooses a best response to the mixture in closed form. We have shown that the adaptive UCB and margin-seeking recommendation-list extractors used in the experiments cannot sit on a band the way they sit on a threshold. We have not shown that users will adopt α as a slider. We have shown that if they did, the effect of changing α rises and then falls, rather than more randomness always being better. We have not shown that MovieLens would look like the synthetic recsys panel under a continuous rating. We have shown that under the ratings the file actually contains, there is no slack to mix. We

have not shown that Adult applicants should randomize their hours. We have shown that against an accuracy-maximizing logistic institution, randomization buys a policy that is harder to read, and not more utility. We have not run a user study or queried a deployed API. Every optimizer and every human policy in this paper is a simulator.

Across these checks, mixing helps inside slack and with a reservation. It does not help when the counterpart is a collaborator or a pure accuracy maximizer. A five-star grid does not create a non-singleton α-band.

I.2 The ten bimatrix games

This subsection records each bimatrix game in words, together with the mean payoff over predictors and seeds. The payoffs are those in `src/envs.py`. The numbers are means from `bimatrix.csv`.

Matching pennies. The human wants to match, the optimizer wants to mismatch, and there is no slack in the interesting sense because against a best responder the two actions are equally good at mixed equilibrium and equally terrible if played pure. A greedy human who breaks a tie toward heads twice in a row has published a peak. Frequency counts move, the optimizer plays tails, and the human's continuation value is -1 until they mix again. In the runs, greedy finishes at about -0.15 rather than -1 because of the 5% random slip and because identification is not instantaneous. The sign of the payoff is the result that matters. UNOP with α large enough to include both actions is the mixed security strategy and finishes at 0.000. Aware-UNOP finishes at $+0.054$, which is the gain from avoiding a slow predictor. Uniform is statistically the same as UNOP here, because the α-set is the whole action set. In this game, remaining effective requires remaining unpredictable.

Rock, paper, scissors. The same conclusion holds with three actions. Greedy that locks onto rock is giving the optimizer paper. UNOP that mixes on all three when they are near-ties is the usual security mix. The numbers track matching pennies closely (greedy about -0.14, UNOP 0.00, Aware-UNOP $+0.05$).

Prisoner's dilemma. Both players have a dominant strategy to defect if the other is held fixed, and an adaptive counterpart that can predict cooperation will defect against it. Every near-optimal policy in our suite therefore collapses to defection and gets about 1.11, the mutual-defection payoff, well below the cooperative payoff of 3. UNOP is not a cooperation device. If the human wants to cooperate against an extractor, they need a different theory (commitment, repeated-game arguments, outside enforcement), not a mixer on $\mathcal{A}_\alpha$.

Chicken. Swerving versus staying has an interior mixed equilibrium and two asymmetric pure equilibria. A greedy human who looks like they will stay can induce the optimizer to swerve, which is good for the human, until the optimizer's predictor becomes confident enough to stay instead. UNOP-$\alpha = 0.15$ edges greedy (1.82 versus 1.77) because mixing in the near-tie region makes it harder for the optimizer to pick the punishing pure strategy and sit on it. The gap is small. In this general-sum game a little slack is mildly useful, neither mandatory nor forbidden.

Inspection. A worker who always works is wasting effort when the inspector is not inspecting. A worker who always shirks is caught once the inspector learns that. The α-set often contains both work and shirk against the empirical inspect rate, which is the support on which UNOP mixes. UNOP-$\alpha = 0.05$ gets 0.96 against greedy's 0.92. The gap is small and in the direction predicted by the theory.

Shapley's game. A cyclic general-sum game that is a standard test for learning dynamics because no pure equilibrium is stable. Mixing is almost unavoidable, and the policies that already mix do better than greedy, which keeps trying to sit on a disappearing best response.

Battle of the sexes. Two pure equilibria, disagreement about which one, and a mixed equilibrium that both dislike. A greedy human who looks determined can sometimes pull the optimizer onto their preferred equilibrium, which is a case where being predictable is helpful even though the game is not common-interest. UNOP does not destroy that, because small α still leaves a peak on the currently better equilibrium, but large α mixes toward the mixed equilibrium and pays the coordination tax of Proposition 3.

Coordination and pure coordination. Greedy, small-α UNOP, and Aware-UNOP all sit at the Pareto-optimal pure profile (payoff 1.95 on the $2, 0$ and $0, 1$ game, 0.965 on three-action pure coordination, 4.88 on stag hunt, and suite mean 2.598). Uniform falls to 1.00 on the two-action game. Predictor accuracy is 1.0 for the peaked policies. If a deployment cannot tell whether the counterpart is a collaborator or an extractor, UNOP run by default against a helper misses the meeting point.

Stag hunt. The Pareto-optimal profile is (stag, stag) at 5, the safe profile is (hare, hare) at 3, and a greedy human who is predicted to hunt stag will be met there by an adaptive partner. Mixing toward hare fails to reach the stag equilibrium. Our runs put greedy and UNOP together at 4.88 and uniform at 3.96, which is the coordination tax. UNOP is not an appropriate default when the counterpart is a cooperative assistant completing a joint task, which is the content of Proposition 3.

I.3 Hyperparameters

Table 16 lists α and the other experimental settings.

Table 16: Experimental settings. Only α is a user-facing parameter. The rest are simulation choices checked by ablations.

Setting	Value(s) used	Role
Bimatrix T	1000 (ablation: 200 to 2000)	Long enough for frequency counts to concentrate. Ablations check that the ranking does not flip.
Bimatrix seeds	12	Enough to see seed spread.
Bimatrix random slip	5% uniform	Prevents exact freeze on a misidentified pure profile.
Pricing T	500 (ablation: 150 to 800)	UCB has time to leave exploration. Longer T helps UNOP relative to greedy.
Price grid	21 points on $[0, 1]$ (ablation: 11 to 41)	Mesh $\Delta \approx 0.05$. Finer grids extract greedy more, consistent with Proposition 4.
v	$U[0.25, 0.95]$, 24 draws	Avoids trivial v near 0 or 1.
Recsys T	300	Catalog identification is cheaper than pricing.
Catalog / list size	48 / 8 (ablation: 24 to 72 / 4 to 12)	Unique-best offer plus filler items. Ranking stable across the grid.
Reservation τ	0.42-quantile of quality	Interior outside option. Neither never-skip nor always-skip.
Promo fraction	0.35	Enough ads to make withholding tempting.
Strategic T	700 synth. / 3000 Adult / 2000 German	Online logistic has time to retrain.
Manipulation levels	5 deltas	Small discrete action set so UNOP's α-set is interpretable.
Cost c	0.55 synth. / 0.45 real (ablation: 0.15 to 1.05)	Interior: some gaming, not always-max or never.
Lock-in T / window	350 / 30	Window defines 70% lock-in. Greedy always hits at 30.
Population users	40, $T = 500$, 12 seeds	Shared optimizer. Fraction of UNOP from 0 to 1 in steps of 0.1.
α main	$\{0.05, 0.15, 0.3\}$ extraction. $\{0.1, 0.2\}$ MovieLens	Named slack in payoff units. MovieLens uses smaller values because ratings are already in $[0, 1]$ with 0.25 star gaps.
IR floor	0 in pricing, τ in recsys, off in bimatrix	Ablation in Appendix F.4 shows why the first two are not optional.
Utility noise σ	0 main. $\{0.02, 0.05, 0.10, 0.20\}$ ablation	α absorbs small σ. $\sigma > \alpha$ does not.
Workers	96 CPU processes	CPU only.

The only user-facing hyperparameter is α, in payoff units (dollars, stars, minutes of extra commute, points of approval probability). The remaining choices are experimental, and the ablations indicate that they do not carry the result.

I.4 AWARE-UNOP

Unaware UNOP is uniform on $\mathcal{A}_\alpha$. Aware-UNOP reweights that set by $(P(a) + \varepsilon)^{-1}$, so that currently over-predicted near-optimal actions are downweighted and currently under-predicted ones are upweighted. It is a small step that avoids the current forecast, not a no-regret algorithm and not a full best response to the optimizer. In matching pennies it is the difference between 0.000 and $+0.054$, because avoiding a slow predictor raises payoff. In pricing it matches unaware surplus (0.041 versus 0.040) at slightly lower accuracy (0.785 versus 0.807), which is the same-effectiveness, harder-to-read refinement. In recommendation it is the least predictable effective policy (accuracy 0.241) at a small surplus cost relative to unaware $\alpha = 0.15$ (0.124 versus 0.148). In Adult it is the least readable policy in the table and not the most effective, which is again the weak-dilemma pattern.

The reason we did not make awareness the default is practical as much as theoretical. It needs a forecast P_t in the same action simplex the human is mixing on. A user who sees a posted price can build a crude $P(\text{buy})$ from how often they have been seeing that price. A user who sees a recommendation list can less easily know the platform's predictive distribution over clicks. Unaware UNOP only needs u and α. If an assistant can see P_t, it should use it. If it cannot, uniform-on-the-band is the distribution characterized by the theorems.

A numerical snapshot, again with $v = 0.70$ and $\alpha = 0.15$. Suppose the posted price is 0.65, so $\mathcal{A}_\alpha = \{\text{skip}, \text{buy}\}$, and the monopolist's current $P(\text{buy}) = 0.8$. Unaware mixes equally. Aware mixes in proportion to $\big((P(\text{skip}) + \varepsilon)^{-1}, (P(\text{buy}) + \varepsilon)^{-1}\big) \approx (1/0.2,\, 1/0.8) = (5, 1.25)$ on $(\text{skip}, \text{buy})$, i.e., more skip than buy, which is the direction that surprises a seller who has become confident that this price still sells. The next $P(\text{buy})$ comes down, and the next mix tilts back. That oscillation is why awareness reduces accuracy without needing to touch dominated actions.

I.5 OBSERVED FAILURE CASES

The following negative results are in the released files.

First, MovieLens-100K and MovieLens-1M, complete files, UNOP indistinguishable from greedy at the surplus level that matters. The failure is the absence of slack: the five-star grid does not give $\mathcal{A}_{0.1}$ anything to contain except the top star.

Second, accuracy-seeking logistic classification on Adult and German Credit, complete tables. UNOP reduces the accuracy of a manipulation-level forecast and does not raise agent utility. The failure is of the adversary's objective: there is no surplus to protect against an institution that is not extracting it.

Third, UNOP without an IR floor, pricing environment. Surplus changes sign, from $+0.036$ to -0.018. The failure is of the definition: mixing into buy-above-value is a teaching signal. The definition in the main text therefore includes the floor.

Fourth, utility noise $\sigma = 0.20$ with $\alpha = 0.15$. UNOP pricing surplus goes negative. The failure is that noise exceeds slack. A separate pricing sweep at the same σ with α up to 0.50 stays negative, so the binding fix is a better utility estimate, not a larger band.

Fifth, prisoner's dilemma and the common-interest suite, where UNOP is either irrelevant (everyone defects) or actively unhelpful (mixing is a coordination tax). The failure is of the environment class, which the three-case pattern separates from surplus extraction.

Sixth, Hedge against a posted-price extractor. Near greedy against UCB, slightly negative against Thompson. No-regret on realized utilities does not protect a threshold. Mean-based learning still publishes one.

Each of these cases limits where mixing helps. Slack can be absent, the counterpart can be maximizing accuracy instead of surplus, and exploration below the reservation can teach the extractor.

I.6 STATISTICAL PRESENTATION

We report means over seeds, with error bars or seed scatters on the figures. We do not report p-values for the main contrasts. The pricing gap is 0.042 versus 0.018 surplus, with greedy's standard deviation 0.016 and a positive gap on most matched draws of v ($n = 24$). The recommendation gap

is 0.148 versus 0.052. The MovieLens comparison is a per-user scatter on the diagonal for all 943 users and then all $6{,}040$ users (Figure 18). The Adult and German pattern repeats on both tables and on the synthetic population. The oracle-pricing and recommendation contrasts in Section 4.2 can be recomputed from `supplement/sprint`, matched by value or catalog.