PRACTICE MAKES PERMANENT: LOCK-IN WHEN ASSIGNMENT CREATES ITS OWN EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Assigning each player to the position they already play best gives them more practice in that position, making the same lineup stronger tomorrow. In a five-player team game, each match consists of five players occupying the same five positions. Any lineup must stay close to the strongest combination that can currently be assembled. We prove an exact threshold on the tolerated decline in performance: below this threshold, always fielding the strongest lineup eventually makes every alternative player–position pair permanently unreachable; for five players starting at skill 0.40 on a scale where mastery is 1, the threshold is a 24% decline. When the tolerated decline is small enough, the lineup fielded in the first match becomes permanent. If skills are observed only through noisy tryouts, no rule does better than assigning a contested position to the player with the higher average tryout score. If uncertainty is too large, decision-makers who require proof that a new lineup is safe before trying it will not risk letting a player practice a new position. We tested teams of five to fifty players and found that the observed thresholds lie within 1.4% of our predictions; even if we balance how often each player gets a desired position, five of the ten desired player–position pairs, two for each player, remain permanently unreachable. Both performative prediction and safe exploration study this assignment–practice feedback loop; here, the thresholds are exact.

1 INTRODUCTION

Many teams tend to assign the same individuals to fixed positions of the same type, and to enhance one's skill in a specific position, one primarily relies on actual work in that position (Shafer et al., 2001; Nembhard & Bentefouet, 2012). For competitive team games like *League of Legends*, or similar mobile games like *Honor of Kings* and *Mobile Legends*, a five-player team must assemble the same five positions for each match. In a restaurant, fixed individuals hold specific stations; in a software team, releases, on-call shifts, and code review repeat each iteration. In this way, planners (a coach, a matching system, or anyone responsible for arranging the lineup) who only think about assembling the strongest lineup today have subtly altered the foundation of judging everyone's abilities tomorrow. The selected team members will have the opportunity to practice, while those who are not selected will have neither the chance to practice nor the means to showcase how much they can improve. The selection policies currently in use are essentially creating the evidence for future judgments. In machine-learning terms, the planner is a deployed policy whose actions change the data it will later observe (the loop of performative prediction), under a safety constraint of the kind studied in safe exploration.

For example, consider two of the players, Ana and Bo. Both of them aspire to play the support role, and initially their skills are almost identical: on the unit skill scale where mastery is 1, Ana's skill is 0.41 while Bo's is 0.40. If the team wants to win tonight, the strongest lineup must have Ana as the support. With more practice, Ana's skill will improve; at the same time, this top lineup also raises the performance floor for future lineup selection, since teams would not be willing to field a lineup that is significantly inferior to their own best level. In the run corresponding to Figure 1 (where unused skills gradually decay; tolerance $\varepsilon = .05$, learning rate $\eta = .10$, forgetting rate $\delta = .03$), after 20 rounds Ana's support skill rose to 0.93, while Bo's dropped to 0.22 due to the lack

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.

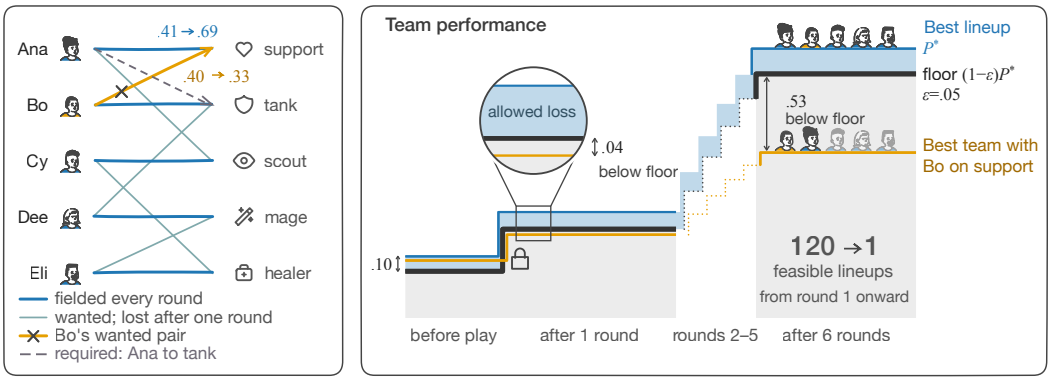


Figure 1: **One round of fielding the strongest lineup locks the assignment in.** Left: blue pairs fielded every round; desired pairs faded, Bo's orange; dashed, the swap shown: Ana to tank. Right: the team stands on P^* (best lineup value); the floor $(1-\varepsilon)P^*$ rides under it (pale band); the gap from the best lineup down to the floor widens from .10 to .17 in team value; feasible lineups: 120 to 1 from round 1 onward (padlock), the best team with Bo on support: .04 under (round 1, lens), .53 (round 6, arrow). The instance and its round-by-round margin: Figure 3, Appendix D.

of opportunity to play support (Figure 3, Appendix D). After just one match, any lineup that moved Ana to another position and let Bo practice support was no longer acceptable. There was no longer any lineup acceptable to the team that could allow Bo's skill to catch up to the performance floor. We refer to this phenomenon as *lock-in*.

Safe exploration fails simultaneously from both sides here, as the minimum performance threshold that a lineup must achieve, its *performance floor*, is determined by the team's own past choices and rises as the players holding the positions improve. Making commitments early on will have a lasting impact. With an opening that can no longer be changed, which happens when the team tolerates only a small decline, no policy can beat the minimax (best possible worst-case) error of the optimal Gaussian test in terms of the stronger player permanently losing the contested position; averaging each player's tryout scores can achieve this bound (Theorem 2). Refusing to take action means learning nothing. When uncertainty is sufficiently high, a planner that changes lineups only when it can certify the change is safe will reject unplayed pairs that could otherwise reduce uncertainty (Proposition 4).

These two types of failures are essentially priced by the same extra margin for error in team performance. If one consistently focuses on honing their skills and always allows players to stick to their strongest positions (*specialization*), the gap between the players holding the positions and the other players who want to vie for those positions will widen. To adjust positions, at least two team members must move simultaneously, as no position can remain unoccupied. The term *training* we are referring to means deliberately fielding a lineup that is not the strongest, in order to give some players practice in a new position. Once the cost of adjusting these two individuals exceeds the acceptable performance decline range for the team, training becomes infeasible, and this state will persist indefinitely. Therefore, when we evaluate fairness, we should not solely consider the current position assigned, but also take into account how many more positions there are that one has the opportunity to play in (their *option set*). Here, opportunity means reaching the skill level at which the team counts one as qualified, not being fielded (Definition 2).

This article makes four core contributions:

- Under the mechanism of a floor that rises with the team itself, the path of safely trying new lineups leads nowhere: the optimal Gaussian testing method can provide the minimax permanent loss probability (Theorem 2); however, if the uncertainty range is large enough, planners who adopt the certify first, try later approach will not even make a single training move that trains a player at a new position (Proposition 4).
- We derived a precise threshold for a symmetric team, where everyone starts at the same skill: $\varepsilon n = 2(1-\bar{s})$, where ε is the acceptable performance decline ratio for the team,

n is the total number of team members, and $1 - \bar{s}$ represents the general gap between everyone and mastery; based on this, we can also deduce a closed-form reversible window, the number of rounds in which a change is still affordable, as well as a special case where it becomes locked in one round.

- The two-lineup rotation scheme we designed has a clearly defined range of feasible tolerance; when considering pair-specific learning rates and judging whether a set of player–position pairings can all reach the qualification bar by a deadline, it is solvable in polynomial time when the number of players n is fixed. However, if the number of players n is not fixed, it becomes an NP-complete problem.
- In the simulation experiment, exact arithmetic verified that even if the playing frequency of each team member's desired position is deliberately balanced, there are still 5 out of 10 desired pairings that never have a chance to be realized in all 50 experiments. Both policy learners we test, trained only to maximize team value, failed to find evidence of being able to catch up for most of the desired pairings in all tested scenarios. In tests with a maximum of 50 players and 5 different team sizes, the error between the measured critical tolerance and the result calculated by the closed-form formula was within 1.4% (see Section 6 for details).

Relation to prior work. The closest previous research to date (Caragiannis & Narang, 2023) involves increasing the value of a certain item every time it is received; with durable skill (once learned, never forgotten), our value is also such a function of the practice count. However, it does not have a floor determined by the team's own optimal level, has no qualification bar (a skill level at which a member counts as qualified for a position), and does not consider reachability, whether a position a player wants can still be reached at all. For additive rising rewards, a constant assignment is optimal for the team (Heidari et al., 2016; Song et al., 2026); Theorem 1(a) is this optimality seen from the other players' side. Reachability of options under a system's own feedback is studied for recommenders (Dean et al., 2020; Curmei et al., 2021); the option set is the team version with a performance floor, where the planner rather than the player acts, and in none of the neighbouring literatures does the performance floor increase alongside the skill levels cultivated by the policy (Appendix A). Our feedback loop draws inspiration from the concept of performative prediction (Perdomo et al., 2020; Brown et al., 2022; Mandal et al., 2023; Jagadeesan et al., 2022), while the constraints are derived from research on safe exploration (Sui et al., 2015; Garcelon et al., 2020); our core contribution lies in integrating these two elements, with explicit thresholds and deadlines.

2 Model: Skills That Respond to Assignment

We have been using a team of five people as an example, but this model is applicable to situations with $n \geq 2$ members and n positions, and there are no empty positions in each round. We use $N = [n] = \{1, \ldots, n\}$ to represent the n team members, and R to represent the n positions. In round t, there will be a perfect matching M_t, assigning each member a unique position; $M_t(i)$ represents the position assigned to member i. We use the terms *lineup*, *assignment*, and *matching* interchangeably; in game-related examples, we refer to positions as roles. Before assignment, the skill level of team member i at position r is denoted as $s_{i,r}(t) \in [0, 1]$. The performance of a lineup is the sum of the skill levels of all members assigned to their respective positions, which can be expressed as $P(M_t, t) = \sum_{i \in N} s_{i,M_t(i)}(t)$. The best lineup performance that can be achieved under the current state is $P^*(t) = \max_M P(M, t)$.

Each time a member plays a certain position, the gap between that member and fully mastering the position (skill level 1) decreases by a proportion $\eta \in (0, 1)$, the learning rate:

$$s_{i,r}(t+1) \;=\; s_{i,r}(t) + \eta \big(1 - s_{i,r}(t)\big). \tag{1}$$

If a member does not play a certain position, the skill there may be lost by a proportion $\delta \in [0, 1)$:

$$s_{i,r}(t+1) \;=\; s_{i,r}(t)\,(1 - \delta) \qquad \text{for } r \neq M_t(i). \tag{2}$$

Both are common assumptions in this type of research (Shafer et al., 2001; Nembhard & Bentefouet, 2012); the formulas fitted there differ from ours, chosen for ease of calculation (Appendix H). The core lock-in conclusion is derived assuming $\delta = 0$, that is, skills are *durable* and once learned will not be forgotten. If $\delta > 0$, skills are *perishable*. Practice outside the team's own matches is

not modelled. Up to Section 4, the dynamics and parameters are all known to the planner; skills become latent, no longer directly observed, from Section 5, and Section 6 tests learners (Figure 2 in Appendix D gives a schematic view of this setting).

Let $\varepsilon \in (0, 1)$, the *tolerance*, signify the maximum decline in performance that the team can accept compared to the best lineup.

Definition 1 (Performance floor)**.** *An assignment M in round t is* floor-feasible *(acceptable) if and only if $P(M, t) \geq (1 - \varepsilon)P^*(t)$. A* schedule*, that is, a sequence of assignments, is floor-feasible if and only if each of its assignments is floor-feasible when played.*

This floor is endogenous. The team's past assignments hone everyone's skills, and these skills in turn determine the value of $P^*(t)$; therefore, if the floor is met today, it may be raised tomorrow. A *policy* decides how to assign positions in each round, based on what its user, the *planner*, has previously observed. The *greedy* policy always focuses on the present and maximizes the current performance. Finally, we fix the qualification bar $\theta \in (0, 1)$, and set up a desired position set $D_i \subseteq R$ for each team member. When $s_{i,r}(t) \geq \theta$, member i is qualified for position r. In the context of a team, θ represents the skill level at which the team feels comfortable entrusting a specific position to a particular member, while D_i lists the positions that member i desires to pursue (stated by the member, not chosen by the planner). Now we can clarify how many options an assignment leaves open for the future.

Definition 2 (Option set)**.** *For a horizon $H \geq 0$, the number of future rounds considered, the option set of member i in round t is $O_i(t, H) = \{r \in D_i : \exists$ a floor-feasible schedule and $0 \leq \tau \leq H$ with $s_{i,r}(t+\tau) \geq \theta\}$. If a position is not in $O_i(t, H)$ for any finite H, the position is* permanently lost *in round t.*

This is reachable qualification (whether member i can ever pass the bar θ); whether the position is then *playable*, assigned to the member by some floor-feasible lineup, is a second question (Section 4).

3 When Optimal Assignment Locks Players In

In this entire section, we assume that skills are durable ($\delta = 0$). We begin with a fully symmetric scenario. Initially, all individuals have the same skill level, denoted as $\bar{s}$; greedy arbitrarily resolves this tie, and its arrangement in round 0 (referred to as the *opening*) establishes each player's *opening position*.

Adopting a different lineup, a *deviation* from greedy, moves at least two players, so its minimum cost is twice the widening skill gap. If a state satisfies the following conditions: each member has skill d in their opening position, and skill $\bar{s}$ in all other positions, where $\bar{s} \leq d$, then we refer to it as a *symmetric* state; when everyone starts with the same level, it corresponds to the case where $d = \bar{s}$.

Lemma 1 (Deviation cost)**.** *The value of the optimal lineup is $P^* = nd$. The lowest-cost alternative assignment exchanges two players (one* transposition*) and loses $2(d - \bar{s})$, so it is floor-feasible if and only if $2(d - \bar{s}) \leq \varepsilon nd$.*

Under greedy, $d(t) = 1 - (1 - \bar{s})(1 - \eta)^t$ and all other skills stay at $\bar{s}$.

Theorem 1 (Exact symmetric lock-in threshold)**.** *Suppose that initially all skills equal $\bar{s} < \theta$, the policy is greedy, and $\delta = 0$.*

- *(a) If $\varepsilon n < 2(1 - \bar{s})$, every deviation eventually becomes permanently infeasible; every desired position other than each team member's opening position is, from some round on, permanently lost.*
- *(b) If $\varepsilon n \geq 2(1 - \bar{s})$, no desired position is permanently lost. From any time, a floor-feasible continuation exists that makes any selected team member–position pair qualified.*

The greedy policy maintains state symmetry, thus requiring only d to move; and only when $2(d - \bar{s}) \leq \varepsilon nd$ is a transposition floor-feasible (for the proof, see Appendix B.2). The losses we can tolerate, summed over the n individuals in the team, must be sufficient to cover the gap of twice the difference from initial skill to mastery; if it falls below this threshold, any lineup that does not follow

greedy will ultimately fail, and the initial arrangement becomes completely fixed and unchangeable. In part (a), by rearranging the deviation inequality of Lemma 1, it becomes evident that only when $d(t) \leq d^*$ (i.e., $d^* := 2\bar{s}/(2-\varepsilon n)$) is it permissible to switch to a different lineup. Calculating this way, the critical tolerance ε^* is $2(1-\bar{s})/n$. For instance, if there are 5 people in the team and $\bar{s} = .40$, then the critical value is 0.24.

Proposition 1 (Reversible window)**.** *Under Theorem 1(a), a floor-feasible deviation exists exactly in rounds $t < W$; equivalently, all transpositions are feasible exactly in those rounds, where $W = \lfloor t^* \rfloor + 1$ and $t^* = \ln((1-d^*)/(1-\bar{s}))/\ln(1-\eta)$. A transposition initiated within this window can continue until its target member is qualified for the target position.*

Here, t^* is calculated by setting $d(t)$ in the learning curve (Equation (1)) equal to d^*. The window grows with the team's tolerance and shrinks with the learning rate: the faster the learning, the more dangerous it is to delay adjustments. For instance, when $n = 5$, $\bar{s} = .40$, and the learning rate $\eta = .10$, if the team can tolerate a 10% performance loss, there is a window period of 3 rounds, approximately three weeks for a team that plays once a week (Appendix H).

After one round of practice, the skill value d_1 equals $\bar{s} + \eta(1-\bar{s})$.

Proposition 2 (Freeze instantly)**.** *Starting from the symmetric initial state, assume $\bar{s} < \theta$ and let $\delta \in [0, 1)$ be arbitrary. If $\varepsilon n d_1 < 2\eta(1-\bar{s})$, then every floor-feasible schedule repeats the assignment selected in round 0 forever. All other desired positions are permanently lost in round 1.*

After only one round of practice, switching to the cheapest deviation already costs twice the new skill difference, which exceeds the loss range acceptable to the team. In the case of complete symmetry, only the initial decision has long-term effects. In the first match the coach assigned either Ana or Bo to support, and thereafter that person remained there (Appendix B.6, also for starts within κ of the symmetric initial state).

The lock-in mechanism itself does not rely on symmetry. If the perturbation magnitude of the initial skills does not exceed a very small value κ, it is proven in Appendix B.3 that the critical tolerance will remain near ε^*. Several natural variations of the floor do not disrupt this mechanism, whereas fixed external goals met at the start do not exclude any possibilities (see Appendix F). If there are more players than positions (a bench), the cheapest deviation moves one player, and lock-in occurs below $\varepsilon n' = 1 - \bar{s}$ for n' positions; pair-specific learning rates keep part (a), and part (b) is open there: a direct swap that is infeasible does not by itself settle reachability, since a schedule of several moves may still reach the pair (both in Appendix B.2).

For a coach planning an entire season, when each player–position pair has its own learning rate, determining whether a given set of pairs can all meet the standard by a deadline T (given in unary, so the input grows with T itself; rates may be 0) can be calculated in polynomial time when n is fixed, but it becomes an NP-complete problem when n is variable (Appendices B.7 and C).

4 Can Rotation Keep the Options?

The reversible window indicates an obvious solution: to immediately commence cross-training, akin to a team switching positions during preseason practice matches (*scrims*). Assume that each team member wishes to play exactly two positions in the perfect matchings A and B, and that for all i, $A(i) \neq B(i)$. Can the alternation $A, B, A, B, \ldots$ consistently stay floor-feasible in a sufficient number of rounds, ensuring that all desired pairs meet the standard?

Proposition 3 (Rotation threshold)**.** *Starting with all skills at $\bar{s}$, with $D_i = \{A(i), B(i)\}$ and $A(i) \neq B(i)$ for all i, and $\delta = 0$, the A, B alternation is floor-feasible in all rounds if and only if $\varepsilon \geq \varepsilon_{\mathrm{rot}} := \eta(1-\bar{s})/\big(\bar{s} + \eta(1-\bar{s})\big)$, and the A, B alternation can then make all desired pairs qualified.*

The team must be willing to give up the value gain of the lineup brought by one round of practice; if they are unwilling to give up so much, the first switch to the B lineup is not floor-feasible (for example, when $\bar{s} = .40$, the rotation threshold $\varepsilon_{\mathrm{rot}}$ corresponding to $\eta = .05$ is .070, and when $\eta = .15$, it is .184). The A, B alternation is also the cheapest schedule that qualifies every desired pair: with k_θ the number of practices that carry one pair from $\bar{s}$ to θ, any schedule that qualifies all desired pairs by round $T \geq 2k_\theta$ gives up at least $n(1-\bar{s})(1-(1-\eta)^{k_\theta})(1-(1-\eta)^{T-k_\theta})/\eta$ of greedy's cumulative value, and the A, B alternation followed by A attains this (Appendix B.9). We call the

ratio of this loss to greedy's cumulative value the *price of optionality*; at $n = 5$, $\bar{s} = .40$, $\theta = .70$ and $T = 100$ it is 6.9% for $\eta = .05$ and 2.3% for $\eta = .15$, and it falls like $1/T$. This is the price of keeping every desired pair qualified; alternating A, B for longer and then settling on A also keeps every pair playable, at no more than 11.0% (at $\varepsilon = .10$) for $\eta = .05$ and at the same 2.3% for $\eta = .15$ over the same 100 rounds (Appendix B.9). Therefore, this time window represents the deadline for starting training rather than the deadline for completing player assessment; this distinction becomes crucial when skills must be measured (see Section 5). Even if there is a slight advantage that remains undiscovered, this pattern will not change (see Appendix E.3); Appendix B.10 discusses how many members can hold a shared position in the long run under forgetting.

5 Option Preservation Under Uncertain Skill

So far, the planner has observed the actual skill level. A real coach will evaluate the performance during tryouts, the results of scrims, and the output of the rating system, all of which are subject to noise (Herbrich et al., 2006). Skill is now latent and cannot be directly measured. The assessment result (a tryout) of player i at position r can be expressed as $y_{i,r} = s_{i,r} + \nu_{i,r}$, where $\nu_{i,r} \sim \mathcal{N}(0, \sigma^2)$ is independent, and σ measures the potential spread of a single assessment result. Before the start of round 0 (the *preseason*), the planner can obtain the assessment results for all n^2 player–position pairs, skills unchanged by these assessments. Afterwards, it can only view the pairs that are actually played.

Appendix E.1 gives two conditions on the true skills (all within κ of $\bar{s}$) under which all assignments are floor-feasible in round 0, but any floor-feasible schedule must repeat its selected opening forever, and all other desired roles are permanently lost in round 1. Since both conditions are targeted at the true skill levels, no matter what one chooses based on an initial assessment with errors, it will be permanently fixed. The two conditions can hold only when $\varepsilon n d_1 < 2\eta(1 - \bar{s})$, the band in which Proposition 2 freezes the opening after one round; at $n = 5$ and $\bar{s} = .40$ that is $\varepsilon < .028, .052, .073$ for $\eta = .05, .10, .15$. In this scenario, assume two players, Ana and Bo, who both aspire to play the same role, b, that is, competing for the position (the *contested* role). This problem is then simplified into a statistical decision-making problem: who should be assigned role b initially. Therefore, we evaluate the goodness of a rule by examining its performance under the most unfavorable true scenario, that is, the scenario that would render the rule least effective. Throughout the paper, Φ represents the cumulative distribution function of the standard normal distribution, and Φ^{-1} denotes its inverse function.

Theorem 2 (Noisy opening becomes permanent)**.** *Let $b \in D_1 \cap D_2$ be the role that players 1 and 2 both desire. All initial skills equal $\bar{s}$, except that one player $\xi \in \{1, 2\}$, unknown to the planner, has $s_{\xi,b}(0) = \bar{s} + \Delta$, where $\Delta > 0$, $\bar{s} + \Delta < \theta$ and $\sigma > 0$. Assume all initial skills lie within Δ of $\bar{s}$ with $\Delta \le \varepsilon\bar{s}/(2 - \varepsilon)$ and $(4 + \varepsilon n)\Delta < 2\eta(1 - \bar{s}) - \varepsilon n(\bar{s} + \eta(1 - \bar{s}))$ (Appendix E.1), and observe each of the two player–b pairs a positive integer m times before round 0. Let π be any policy (possibly randomized) that observes only the assessments, is fixed without knowing ξ, and is floor-feasible under both hidden instances; write $\Pr_\xi^\pi$ for probabilities under π when the better player is ξ. Then $\max_{\xi \in \{1,2\}} \Pr_\xi^\pi[\text{the better player permanently loses } b] \ge \Phi\big(-\Delta\sqrt{m}/(\sqrt{2}\,\sigma)\big)$. The natural policy assigns b to the player with the higher average assessment for b and then repeats that opening; it attains equality under both hidden instances.*

Why this is a reduction. The opening-freeze conditions of Appendix E.1 do the main work: under them, every floor-feasible policy repeats the opening it plays, so the entire problem simplifies to deciding, before round 0, which player receives b. The remaining problem is the standard one-shot test comparing the means of two Gaussians. The bound above is the minimax error probability of this test (Appendix E.2).

When $m = 1$ and $\sigma = \Delta$, even with the natural policy, the probability of error is around 24%; to achieve more accurate judgment, the amount of tryout evidence required will increase with $(\sigma/\Delta)^2$. Figure 7(a) in Appendix G tests a simpler rule, specifically the case where $n = 2$: with two roles, two lineups can be formed, corresponding to four player–role pairings. The logic behind greedy on the assessment sums is straightforward: the lineup with the higher sum of assessments will be deployed. The assessment of the other role carries only noise, so when $m = 1$, the error probability of this rule is $\Phi(-\Delta/(2\sigma))$; when $\sigma = \Delta$, the error rate is approximately 31%, which is higher than the 24% of the natural policy.

Certification can also stop learning. A cautious planner will maintain two estimates for each player–position pair: a lower limit $\ell_{i,r}(t)$ and an upper limit $u_{i,r}(t)$. When the sum of the lower limits for a certain lineup M is at least $(1-\varepsilon)$ times the largest sum of upper limits over all possible lineups, that is, $\sum_i \ell_{i,M(i)}(t) \geq (1-\varepsilon)\max_{M'} \sum_i u_{i,M'(i)}(t)$, it is said that this lineup M is *certified*. As long as the true skills $s_{i,r}$ of all player–position pairs are trapped between the corresponding lower and upper limits, this test is definitely safe, because the lower-limit value of the lineup then clears the true floor (see Appendix E.4 for proof). However, there is an inescapable issue with being overly cautious: the range of estimates for unplayed player–position pairs is always wide and inaccurate. Only by assigning them to play in that position can the uncertainty be narrowed down. To state the next conclusion, we represent each interval as $[\ell, u] = [\mu - w, \mu + w]$, where μ is the midpoint (center) of the interval and w is the half-width (radius) of the interval.

Proposition 4 (Certification starvation)**.** *Start from exact symmetry with $\bar{s} < \theta$, fix a round-0 incumbent M_0 (the lineup the planner repeats unless a change is certified), and let $\delta = 0$ and $\varepsilon n < 2$. Suppose intervals have center μ, radius w, satisfy $|\mu - s| \leq w/2$ for the true skill s of that pair, and all unplayed pairs keep radius w_0. If $w_0 > \varepsilon n \bar{s}$, a planner attempting only certified deviations and otherwise repeating M_0 never certifies a deviation.*

What does the condition $w_0 > \varepsilon n \bar{s}$ mean? It means that the uncertainty of a single pair that has never been played is already greater than the team's entire slack (the loss in team value that the floor tolerates) from the symmetric start. What is stuck is the cautious planner (*starvation*: it never gets the information it would need to move). Within the exact-symmetry regime, if additionally $\varepsilon \geq \varepsilon^*$, Theorem 1(b) says that a planner who knows the true skills could still reach them. This is the failure mode of certification in an endogenous matching system, the trap that safe exploration guards against with an expansion step (Sui et al., 2015) or with a performance budget against a baseline policy (Garcelon et al., 2020); Appendix E.5 gives the induction. The planner here has no step that expands its safe set; the same appendix also covers incumbent radii that shrink as the pairs are played.

If the interval estimates are accurate enough, the planner can still certify the A, B alternation of Proposition 3: in the loose/slow cell of Section 6 ($\varepsilon = .10$, $\eta = .05$, $\sigma = .05$, intervals of $z = 2$ standard deviations; a cell is one floor-and-learning-rate combination), each pair needs $m = 533$ assessments before the season to meet the condition (Appendix E.6).

6 Numerical Experiments

We simulated $n = 5$ players and ran a total of $T = 100$ rounds. Each player's initial skill is randomly drawn from $[.37, .43]$ (independently); the deviation from $\bar{s} = .40$ is at most $\kappa = .03$, twice the radius $\kappa_0 = .015$ that the robustness proposition of Appendix B.3 admits at $\eta = .10$ (four times at $\eta = .05$); the qualification bar is $\theta = .70$. Each player i wishes to train two roles, $D_i = \{i, (i \bmod 5) + 1\}$, their main role and a second role, which is the next number in sequence; this results in a total of ten player–desired-role pairs. $\eta = .05$ is referred to as the *slow* learning rate and $\eta = .15$ as the *fast* learning rate (to reduce the gap to mastery by half, approximately 13.5 and 4.3 practices are required; Appendix H). Public per-match role data for professional *League of Legends* exist (Oracle's Elixir, since 2014); a learning rate fitted to them is left to future work. $\varepsilon = .05$ is referred to as the *tight* floor; $\varepsilon = .10$ is referred to as the *loose* floor. After 100 rounds have been completed for each policy, we add a fixed 20-round *catch-up check* ($H = 20$) to specifically see if the roles that failed to train can still be reached. For each player–desired-role pair: as long as a lineup can be assembled that clears the floor and allows the player to practice this role, use this lineup; if it cannot be worked out, employ greedy. If it succeeds, it indicates the existence of a floor-feasible schedule to train this role. NR_{20}: out of the ten desired pairs, how many pairs failed to be reached in this 20-round check. *Performance*: over the entire process, the total value of the lineups is compared with the total value of greedy throughout on the same initial draw.

We compared a total of twelve policies. *Greedy*: indifferent to everything else, always fielding the strongest lineup. *Count-maximin*: this involves specifically focusing on the player who has been assigned a desired role the fewest times among all players, and, within the lineups that clear the floor, elevating this lowest number; *pair-count maximin* does the same for the count of each desired player–role pair. The two maximin rules also count the round about to be played. If the lowest count is the same for several lineups, the one whose total count is larger is taken; if this is also the same,

the one whose team value is larger; if it is still the same, the first one in the fixed order of the 120 permutations. The *urgency rule*: within the lineups that clear the floor, prioritize the arrangement of those desired pairs that are not yet up to standard. *In-season rotation*: as long as the floor is met, the alternation of Proposition 3 shall be followed; if it is not met, the urgency rule shall be reverted to (Algorithm 1 in Appendix D gives the specific algorithm). *Preseason rotation*: initially, treat it like scrims. The first ten rounds of warm-up matches, which are exempt from the floor, are played on a rotation basis. After that, shift to the urgency rule. *Cyclic round-robin*: in each round, everyone moves to the next role, disregarding the floor. Two *policy learners*: one is ϵ-greedy (exploration rate $\epsilon = .10$, distinct from the tolerance ε), and the other is score-update, both of which observe the true skills of players and select from lineups that clear the floor. The score-update learner raises only the score of the played lineup, a chosen-lineup rule spelled out in Appendix G. The training objective is singular: to maximize the team value. Two *option-aware learners* are the same two learners with the reward raised by λ times the number of desired pairs still *playable* after the round, a pair being playable when some lineup that clears the floor fields the player in that role ($\lambda = .2$ in the table; Appendix G sweeps it). The *option planner* looks one round ahead: among its candidate lineups, greedy and the best floor-clearing lineup that fields each unqualified desired pair, it plays the one after which the catch-up check leaves the most desired roles open for the player with the fewest, then for everyone in total, ties broken by team value.

We tested different ε values using greedy: NR_{20} remained stable at 7.82 when $\varepsilon \leq .22$, but plummeted to 0 when $\varepsilon \geq .26$. We conducted another test using the admissible $\kappa = .015$ (Appendix G), and the actual measured critical tolerance fell precisely within the predicted range of $[.222, .258]$, revolving around $\varepsilon^* = .24$. We also tested different team sizes ($n \in \{5, 8, 10, 20, 50\}$), using $\kappa = .015$: the interpolated midpoint for each team size fell within the interval $\varepsilon^*(n) \pm 6\kappa/n$, with a difference of less than 1.4% from the formula $\varepsilon^*(n) = 2(1 - \bar{s})/n$ (Figure 6 in Appendix G).

6.1 Outcome counts do not preserve options

The hypothesis: when the floor allows for leeway in training, balancing the frequency at which each player acquires *some* desired role can ensure the retention of both desired roles. In the loose/slow cell ($\varepsilon = .10$, $\eta = .05$), under count-maximin, each player was able to obtain a desired role in every round out of 100 rounds, while maintaining a performance comparable to that of greedy at .998. However, in all 50 simulation runs, the final NR_{20} count remained at five. After 100 rounds, out of the $5! = 120$ matchings in each simulation, only one clears the floor; moreover, it can be proven through exact inequalities that hold for all future rounds that the remaining 119 matchings will never clear it. Five desired pairs are qualified, while the other five pairs will never be reached.

The logic of this mechanism is quite simple: it only counts the number of times a desired role is obtained in the current round, and repeating the same role can still meet the criteria. The urgency rule enables a floor-feasible training schedule for all ten pairs, while maintaining a performance of .925. In the tight/fast cell, through exhaustive enumeration: among the five policies that respect the floor (greedy, the two maximin rules, the urgency rule, in-season rotation), in all 250 trajectories generated under 50 random seeds, only one floor-feasible matching remains, at the fastest after the first match and at the slowest after the second.

Only the loose/slow cell is above $\varepsilon_{\mathrm{rot}}$ (.070 when $\eta = .05$); and within this cell, out of 50 random seeds, 44 times the in-season rotation can reach all ten desired pairs, paying 7.4% of greedy's value on average in those runs (the cell mean .935 includes the six fallback runs) and never less than the symmetric 6.9% of Appendix B.9 (the remaining 6 instances, and the three cells below $\varepsilon_{\mathrm{rot}}$, are detailed in Appendix G). For the floor-exempt preseason rotation, in these four durable-skill cells, floor-feasible training schedules could be found for all ten desired pairs, all along one trajectory; compared to greedy, the cumulative cost was only 7.5% higher (when $\eta = .05$) or 2.4% higher (when $\eta = .15$). Once it specializes, though, only five of the ten qualified pairs stay playable (Appendix G). The urgency rule in the loose/slow cell and in-season rotation in 44 of 50 loose/slow runs qualify all ten pairs along one schedule and end the same way, with five playable. Pair-count maximin is the one floor-respecting policy that keeps all ten playable, in the 36 of 50 loose/slow runs where it qualifies them; round-robin also ends with all ten playable, but it ignores the floor. Both policy learners, which optimize team value only, fail here. The ϵ-greedy learner only tries 1.0 to 1.6 different lineups among the 120. It stays between 8.20 and 8.30 unreached pairs in every cell. Only in the most lenient loose/slow cell does the score-update learner continue to explore,

Table 1: **Outcome counts need not preserve options.** Durable skills; 50-run means with standard errors; performance relative to same-draw greedy over 100 rounds; NR_{20} counts the desired pairs not reached by the fixed $H = 20$ check (one-sided); $\kappa = .03$ (Appendix G).

	Tight floor ($\varepsilon = 0.05$)				Loose floor ($\varepsilon = 0.10$)			
	$\eta = 0.05$		$\eta = 0.15$		$\eta = 0.05$		$\eta = 0.15$	
Policy	Perf.	NR_{20}	Perf.	NR_{20}	Perf.	NR_{20}	Perf.	NR_{20}
Greedy	1.000	7.82±.20	1.000	7.82±.20	1.000	7.82±.20	1.000	7.82±.20
Count-maximin	0.995	4.98±.02	0.999	5.04±.03	0.998	5.00	0.999	5.00
Pair-count maximin	0.998	5.04±.03	0.999	5.04±.03	0.914	1.40±.32	0.999	5.00
Urgency rule	0.989	4.40±.23	0.999	5.04±.03	0.925	0.00	0.997	4.86±.05
In-season rotation	0.994	4.80±.14	0.999	5.04±.03	0.935	0.60±.23	0.999	5.00
Preseason rotation	0.926	0.00	0.976	0.00	0.925	0.00	0.976	0.00
Cyclic round-robin	0.696	0.00	0.840	0.00	0.696	0.00	0.840	0.00
ϵ-greedy learner	0.996	8.22±.16	0.999	8.22±.15	0.994	8.30±.17	0.998	8.26±.17
Score-update learner	0.980	7.88±.16	0.999	7.80±.19	0.901	7.36±.22	0.988	7.54±.19
Option-aware ϵ-greedy	0.996	8.22±.15	0.999	8.22±.15	0.994	8.30±.17	0.998	8.28±.18
Option-aware score-update	0.971	7.68±.17	0.999	7.80±.19	0.878	6.64±.24	0.984	7.18±.21
Option planner	0.929	3.94±.17	0.999	5.76±.12	0.933	4.64±.21	0.982	4.38±.17

achieving a performance of .901, and 7.36 out of ten desired pairs still failed to find a floor-feasible training schedule. Adding the option term changes little. The score-update learner improves by 0.7 pairs at $\lambda = .2$, and at $\lambda = 1$ it still leaves 4.6 pairs unreached in the loose/slow cell at a 16% cost (Appendix G). The option planner leaves 3.9 to 5.8 pairs unreached. Thompson sampling and UCB are pair-level learners that see only noisy tryouts of each player–role pair: Thompson sampling plays greedy on one random draw of every pair's skill, UCB plays greedy on every pair's optimistic estimate; they leave 7.68 to 8.08 of ten pairs unreached at $\sigma = .05$, against greedy's 7.82. The full comparison is Table 2 in Appendix G.

Under noise. Figure 7 in Appendix G tests the two predictions of Section 5 with one assessment per pair: the actual probability of denial (the better player losing the contested role) is quite close to the formula $\Phi(-\Delta/(2\sigma))$ (.304 measured against .309 when $\sigma = \Delta = .01$), and with $\varepsilon = .25$, just above $\varepsilon^* = .24$, the proportion of 50 perturbed-start runs in which the cautious planner certifies a training deviation (a deviation played to train someone) is 1.00 when $\sigma = .02$, but it drops to .38 when $\sigma = .05$, and it drops to zero directly when $\sigma \geq .10$.

7 Conclusion

In our symmetrical setting, where skills, once learned, remain effective for a long time (durable), the floor that rises with the team itself brings about a clear lock-in threshold and a deadline for starting training. After the opening is fixed, measurement noise is not corrected by practice: any errors made at the outset will affect subsequent practice and the evidence obtainable. A policy that can assign a desired role to each player in every round may also permanently eliminate half of the desired player–role options. Therefore, cross-training should commence before the skill gap widens; a single noisy tryout may determine who plays support for the entire season, rather than just one match's lineup.

Reproducibility Statement

The complete proofs are provided in the appendix. Code and materials needed to reproduce the numerical experiments are included in the anonymized supplementary material.

LLM Usage Statement

We used Claude Fable 5 to correct grammatical errors and refine the language in the main paper. GPT 5.6 pro was used to help construct part of the proofs in the appendix section. GPT 6 Pro and Fable 5.1 was used to check the correctness of the proofs presented in the appendix.

Ethics and Scope Statement

This study uses synthetic team data and does not involve real data from human participants or players. The game settings are merely illustrative: this model does not measure the abilities of any real individuals, and its simplified conclusions should not be used for personnel decisions without evidence and safeguards specific to the field. Our results are only applicable to the learning curve in Section 2, synthetic teams, and threshold scenarios close to symmetry; starts within κ of symmetry move the threshold by at most $6\kappa/n$ (Appendix B.3), and two other curve families keep the transition (Appendix H). Even so, this model serves as a reminder: assess the possibilities left open for the future, not only the assignments made.

References

Terrence Adams and Erel Segal-Halevi. Perpetually fair assignments via balanced sequences of permutations. arXiv:2602.21687, 2026.

Kenneth J. Arrow. The economic implications of learning by doing. *Review of Economic Studies*, 29(3):155–173, 1962.

Kenneth J. Arrow and Anthony C. Fisher. Environmental preservation, uncertainty, and irreversibility. *The Quarterly Journal of Economics*, 88(2):312–319, 1974.

W. Brian Arthur. Competing technologies, increasing returns, and lock-in by historical events. *The Economic Journal*, 99(394):116–131, 1989.

Jean-Pierre Aubin. *Viability Theory*. Birkhäuser, 1991.

Nader Azizi, Saeed Zolfaghari, and Ming Liang. Modeling job rotation in manufacturing systems: The study of employee's boredom and skill variations. *International Journal of Production Economics*, 123(1):69–85, 2010.

Soumya Basu, Rajat Sen, Sujay Sanghavi, and Sanjay Shakkottai. Blocking bandits. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pp. 4785–4794, 2019.

Dimitris Bertsimas, Vivek F. Farias, and Nikolaos Trichakis. The price of fairness. *Operations Research*, 59(1):17–31, 2011.

Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. Equity of attention: Amortizing individual fairness in rankings. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 405–414, 2018.

Gavin Brown, Shlomi Hod, and Iden Kalemaj. Performative prediction in a stateful world. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

Ioannis Caragiannis and Shivika Narang. Repeatedly matching items to agents fairly and efficiently. In Argyrios Deligkas and Aris Filos-Ratsikas (eds.), *Algorithmic Game Theory*, volume 14238 of *Lecture Notes in Computer Science*, pp. 347–364, Cham, 2023. Springer Nature Switzerland. doi: 10.1007/978-3-031-43254-5_20. URL https://link.springer.com/chapter/10.1007/978-3-031-43254-5_20. Proceedings of the 16th International Symposium on Algorithmic Game Theory (SAGT); journal version in Theoretical Computer Science, vol. 981, art. 114246.

Sarah H. Cen and Devavrat Shah. Regret, stability & fairness in matching markets with bandit learners. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 151 of *Proceedings of Machine Learning Research*, pp. 8938–8968. PMLR, 2022. URL https://proceedings.mlr.press/v151/cen22a.html.

Allison J. B. Chaney, Brandon M. Stewart, and Barbara E. Engelhardt. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys)*, pp. 224–232, 2018.

Emily Chen, Alexander J. Bisberg, and Emilio Ferrara. "can you play anything else?" understanding play style flexibility in league of legends. In *Proceedings of the IEEE Conference on Games (CoG)*, pp. 1–4, 2024.

Shouyuan Chen, Tian Lin, Irwin King, Michael R. Lyu, and Wei Chen. Combinatorial pure exploration of multi-armed bandits. In *Advances in Neural Information Processing Systems 27 (NIPS)*, pp. 379–387, 2014.

Stephen Coate and Glenn C. Loury. Will affirmative-action policies eliminate negative stereotypes? *American Economic Review*, 83(5):1220–1240, 1993.

Elliot Creager, David Madras, Toniann Pitassi, and Richard Zemel. Causal modeling for fairness in dynamical systems. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2185–2195, 2020.

Mihaela Curmei, Sarah Dean, and Benjamin Recht. Quantifying availability and discovery in recommender systems via stochastic reachability. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2265–2275, 2021.

Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D. Sculley, and Yoni Halpern. Fairness is not static: Deeper understanding of long term fairness via simulation studies. In *Proceedings of the 3rd Conference on Fairness, Accountability, and Transparency (FAT*)*, 2020.

Paul A. David. Clio and the economics of QWERTY. *American Economic Review*, 75(2):332–337, 1985.

Philippe De Bruecker, Jorne Van den Bergh, Jeroen Beliën, and Erik Demeulemeester. Workforce planning incorporating skills: State of the art. *European Journal of Operational Research*, 243 (1):1–16, 2015.

Sarah Dean, Sarah Rich, and Benjamin Recht. Recommendations and user agency: The reachability of collaboratively-filtered information. In *Proceedings of the 3rd Conference on Fairness, Accountability, and Transparency (FAT*)*, pp. 436–445, 2020.

Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. Runaway feedback loops in predictive policing. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAT*)*, pp. 160–171, 2018.

Tor Eriksson and Jaime Ortega. The adoption of job rotation: Testing the theories. *ILR Review*, 59 (4):653–666, 2006.

Benjamin Eysenbach, Shixiang Gu, Julian Ibarz, and Sergey Levine. Leave no trace: Learning to reset for safe and autonomous reinforcement learning. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018.

Ge Fan, Chaoyun Zhang, Kai Wang, Yingjie Li, Junyang Chen, and Zenglin Xu. CUPID: Improving battle fairness and position satisfaction in online MOBA games with a re-matchmaking system. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2):Article 439, 2024.

Guo Freeman and Donghee Yvette Wohn. Understanding esports team formation and coordination. *Computer Supported Cooperative Work (CSCW)*, 28:95–126, 2019.

Evrard Garcelon, Mohammad Ghavamzadeh, Alessandro Lazaric, and Matteo Pirotta. Conservative exploration in reinforcement learning. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1431–1441, 2020.

Gianmarco Genalti, Marco Mussi, Nicola Gatti, Marcello Restelli, Matteo Castiglioni, and Alberto Maria Metelli. Graph-triggered rising bandits. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15351–15380, 2024.

Robert Gibbons and Michael Waldman. Task-specific human capital. *American Economic Review (Papers and Proceedings)*, 94(2):203–207, 2004.

Nathan Grinsztajn, Johan Ferret, Olivier Pietquin, Philippe Preux, and Matthieu Geist. There is no turning back: A self-supervised approach for reversibility-aware reinforcement learning. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, 2021.

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

Moritz Hardt, Meena Jagadeesan, and Celestine Mendler-Dünner. Performative power. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1929–1938. PMLR, 2018. URL `https://proceedings.mlr.press/v80/hashimoto18a.html`.

Hoda Heidari and Jon Kleinberg. Allocating opportunities in a dynamic model of intergenerational mobility. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 15–25, 2021.

Hoda Heidari, Michael Kearns, and Aaron Roth. Tight policy regret bounds for improving and decaying bandits. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1562–1570, 2016.

Hoda Heidari, Vedant Nanda, and Krishna P. Gummadi. On the long-term impact of algorithmic decision policies: Effort unfairness and feature segregation through social learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2692–2701, 2019.

Jobst Heitzig and Ram Potham. A fair objective for human-empowerment-preserving ai: Desiderata, design, and likely behavioral consequences. In *Algorithmic Decision Theory (ADT)*, 2026. To appear; concurrent work; arXiv:2608.08240.

Ralf Herbrich, Tom Minka, and Thore Graepel. TrueSkill: A Bayesian skill rating system. In *Advances in Neural Information Processing Systems 19 (NIPS)*, pp. 569–576, 2006.

Wallace J. Hopp and Mark P. Van Oyen. Agile workforce evaluation: A framework for cross-training and coordination. *IIE Transactions*, 36:919–940, 2004.

Yaowei Hu and Lu Zhang. Achieving long-term fairness in sequential decision making. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, 2022.

Ayumi Igarashi, Martin Lackner, Oliviero Nardi, and Arianna Novaro. Repeated fair allocation of indivisible items. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, 2024.

Nicole Immorlica and Robert Kleinberg. Recharging bandits. In *59th IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 309–319, 2018.

Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Meena Jagadeesan, Tijana Zrnic, and Celestine Mendler-Dünner. Regret minimization with performative feedback. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9760–9785, 2022.

Ray Jiang, Silvia Chiappa, Tor Lattimore, András György, and Pushmeet Kohli. Degenerate feedback loops in recommender systems. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pp. 383–390, 2019.

Ramesh Johari, Vijay Kamble, and Yash Kanoria. Matching while learning. *Operations Research*, 69(2):655–681, 2021.

Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems 29 (NIPS)*, pp. 325–333, 2016.

Sampath Kannan, Aaron Roth, and Juba Ziani. Downstream effects of affirmative action. In *Proceedings of the 2nd Conference on Fairness, Accountability, and Transparency (FAT*)*, pp. 240–248, 2019.

Ian A. Kash, Ariel D. Procaccia, and Nisarg Shah. No agent left behind: Dynamic fair division of multiple resources. *Journal of Artificial Intelligence Research*, 51:579–603, 2014.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016.

Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi-Yadkori, and Benjamin Van Roy. Conservative contextual linear bandits. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pp. 3910–3919, 2017.

Jooyeon Kim, Brian C. Keegan, Sungjoon Park, and Alice Oh. The proficiency-congruency dilemma: Virtual team design and performance in multiplayer online games. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016.

Alexander S. Klyubin, Daniel Polani, and Chrystopher L. Nehaniv. Keep your options open: An information-based driving principle for sensorimotor systems. *PLoS ONE*, 3(12):e4018, 2008.

Victoria Krakovna, Laurent Orseau, Ramana Kumar, Miljan Martic, and Shane Legg. Penalizing side effects using stepwise relative reachability. arXiv:1806.01186, 2018.

Morimitsu Kurino and Yoshinori Kurokawa. Job rotation or specialization? a dynamic matching model analysis. *Review of Economic Design*, 28:243–273, 2024.

Nir Levine, Koby Crammer, and Shie Mannor. Rotting bandits. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pp. 3074–3083, 2017.

Qian Li, Qinshan Sun, Sha Tao, and Xinglin Gao. Multi-skill project scheduling with skill evolution and cooperation effectiveness. *Engineering, Construction and Architectural Management*, 27(8): 2023–2045, 2020.

Eugene Lim, Vincent Y. F. Tan, and Harold Soh. Stochastic bandits for egalitarian assignment. *Transactions on Machine Learning Research*, 2024. OpenReview: kmKVJl2JWo.

Eugene Lim, Tzeh Yuan Neoh, and Nicholas Teh. Fairness in repeated matching: A maximin perspective. In *Proceedings of the 40th AAAI Conference on Artificial Intelligence*, volume 40, pp. 17111–17119, 2026.

Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 3150–3158, 2018.

Lydia T. Liu, Horia Mania, and Michael I. Jordan. Competing bandits in matching markets. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1618–1628, 2020a.

Lydia T. Liu, Ashia Wilson, Nika Haghtalab, Adam Tauman Kalai, Christian Borgs, and Jennifer Chayes. The disparate equilibria of algorithmic decision making when individuals invest rationally. In *Proceedings of the 3rd Conference on Fairness, Accountability, and Transparency (FAT*)*, 2020b.

Debmalya Mandal, Stelios Triantafyllou, and Goran Radanovic. Performative reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pp. 23642–23680, 2023.

Shie Mannor and John N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5:623–648, 2004.

Alberto Maria Metelli, Francesco Trovò, Matteo Pirola, and Marcello Restelli. Stochastic rising bandits. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pp. 15421–15457, 2022.

Cristiano Migali, Marco Mussi, Gianmarco Genalti, and Alberto Maria Metelli. Tightening regret lower and upper bounds in restless rising bandits. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 18437–18478. Curran Associates, Inc., 2025. doi: 10.52202/085713-0622. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/1aa670dafbd4b32ebf483f8372e612c5-Paper-Conference.pdf.

Teodor Mihai Moldovan and Pieter Abbeel. Safe exploration in Markov decision processes. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pp. 1711–1718, 2012.

Hussein Mouzannar, Mesrob I. Ohannessian, and Nathan Srebro. From fair decision making to social equality. In *Proceedings of the 2nd Conference on Fairness, Accountability, and Transparency (FAT*)*, pp. 359–368, 2019.

Marco Mussi, Alessandro Montenegro, Francesco Trovò, Marcello Restelli, and Alberto Maria Metelli. Best arm identification for stochastic rising bandits. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36953–36989, 2024.

David A. Nembhard and Frank Bentefouet. Parallel system scheduling with general worker learning and forgetting. *International Journal of Production Economics*, 139:533–542, 2012.

Jaime Ortega. Job rotation as a learning mechanism. *Management Science*, 47(10):1361–1370, 2001.

Vishakha Patil, Ganesh Ghalme, Vineet Nair, and Y. Narahari. Achieving fairness in the stochastic multi-armed bandit problem. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, volume 34, pp. 5379–5386, 2020.

Vishakha Patil, Vineet Nair, Ganesh Ghalme, and Arindam Khan. Mitigating disparity while maximizing reward: Tight anytime guarantee for improving bandits. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 4100–4108, 2023.

Prasanta K. Pattanaik and Yongsheng Xu. On ranking opportunity sets in terms of freedom of choice. *Recherches Économiques de Louvain*, 56(3–4):383–390, 1990.

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

Gaurab Pokharel, Diptangshu Sen, Sanmay Das, and Juba Ziani. Fixed points and stochastic meritocracies: A long-term perspective. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2026.

Canice Prendergast. The role of promotion in inducing specific human capital acquisition. *Quarterly Journal of Economics*, 108(2):523–534, 1993.

Ruwen Qin, David A. Nembhard, and Walter L. Barnes II. Workforce flexibility in operations management. *Surveys in Operations Research and Management Science*, 20(1):19–33, 2015.

Nasim Rahaman, Steffen Wolf, Anirudh Goyal, Roman Remme, and Yoshua Bengio. Learning the arrow of time for problems in reinforcement learning. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2020.

Scott M. Shafer, David A. Nembhard, and Mustafa V. Uzumeri. The effects of worker learning, forgetting, and heterogeneity on assembly line productivity. *Management Science*, 47(12):1639–1653, 2001.

Virag Shah, Jose Blanchet, and Ramesh Johari. Bandit learning with positive externalities. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, 2018.

Simrita Singh, Naireet Ghosh, and Tinglong Dai. Managing the human fallback: Skill investment under improving AI and worker mobility. arXiv:2606.29111, 2026.

Seockbean Song, Youngsik Yoon, Siwei Wang, Wei Chen, and Jungseul Ok. Combinatorial rising bandits. In *Proceedings of the 14th International Conference on Learning Representations (ICLR)*, 2026.

Yanan Sui, Alkis Gotovos, Joel Burdick, and Andreas Krause. Safe exploration for optimization with Gaussian processes. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pp. 997–1005, 2015.

Cem Tekin and Mingyan Liu. Online learning of rested and restless bandits. *IEEE Transactions on Information Theory*, 58(8):5588–5611, 2012.

Matteo Turchetta, Felix Berkenkamp, and Andreas Krause. Safe exploration in finite Markov decision processes with Gaussian processes. In *Advances in Neural Information Processing Systems 29 (NIPS)*, pp. 4305–4313, 2016.

Alexander M. Turner, Dylan Hadfield-Menell, and Prasad Tadepalli. Conservative agency via attainable utility preservation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2020.

Alexander M. Turner, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. Optimal policies tend to seek power. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, 2021.

Rolf van Lieshout and Bart van Rossum. The fair periodic assignment problem. In *Proceedings of the 25th Symposium on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems (ATMOS)*, volume 137 of *Open Access Series in Informatics (OASIcs)*, 2025.

Tonghan Wang, Heng Dong, Victor Lesser, and Chongjie Zhang. ROMA: Multi-agent reinforcement learning with emergent roles. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 9876–9886, 2020.

Tonghan Wang, Tarun Gupta, Anuj Mahajan, Bei Peng, Shimon Whiteson, and Chongjie Zhang. RODE: Learning roles to decompose multi-agent tasks. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.

Min Wen, Osbert Bastani, and Ufuk Topcu. Algorithms for fairness in sequential decision making. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1144–1152, 2021.

Peter Whittle. Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability*, 25(A):287–298, 1988.

Yifan Wu, Roshan Shariff, Tor Lattimore, and Csaba Szepesvári. Conservative bandits. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pp. 1254–1262, 2016.

Tongxin Yin, Reilly Raab, Mingyan Liu, and Yang Liu. Long-term fairness with unknown dynamics. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Bin Yu. Assouad, Fano, and Le Cam. In David Pollard, Erik Torgersen, and Grace L. Yang (eds.), *Festschrift for Lucien Le Cam*, pp. 423–435. Springer, New York, 1997. doi: 10.1007/978-1-4612-1880-7_29.

Xueru Zhang, Ruibo Tu, Yang Liu, Mingyan Liu, Hedvig Kjellström, Kun Zhang, and Cheng Zhang. How do fair decisions fare in long-term qualification? In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

The appendix is organized as follows. Appendix A gives the detailed comparison with the seven adjacent literatures. Appendix B contains the complete deterministic proofs, including a robustness bracket for perturbed starts, the planning-complexity theorem, and the forgetting-capacity result. Appendix C gives the full hardness construction for Theorem 3(b). Appendix D draws the setting and states the rotation policy in full. Appendix E proves the uncertain-skill results, including the conditional certified-rotation guarantee and a tightly scoped two-player alternation calculation. Appendix F shows how lock-in depends on the floor: a lagged benchmark gives the same threshold, an additive allowance gives an exactly solvable analogue, and a fixed floor feasible at the symmetric start leaves every assignment feasible. Appendix G collects experimental details. Appendix H tests the lock-in picture under power-law and logistic learning curves.

A Related Work in Detail

Further neighbours: performative power (Hardt et al., 2022); delayed-impact and downstream effects of fair decisions (Mouzannar et al., 2019; Kannan et al., 2019; Creager et al., 2020); degenerate feedback loops and algorithmic confounding in recommenders (Jiang et al., 2019; Chaney et al., 2018); irreversibility in reinforcement learning (Eysenbach et al., 2018; Rahaman et al., 2020); rested and restless bandits (Tekin & Liu, 2012); the "price of X" convention (Bertsimas et al., 2011); job rotation in firms (Eriksson & Ortega, 2006); strategic classification (Hardt et al., 2016); fair stochastic bandits (Patil et al., 2020); and a mechanism twin in economics (Singh et al., 2026).

Repeated and temporal fair division. The research on dynamic fair division starts from the scenario where "agents arrive one after another over time" (Kash et al., 2014); Igarashi et al. (2024) studies the scenario of multiple rounds of allocation of the same set of goods and chores, comparing the differences between single-round fairness and overall fairness; Lim et al. (2026) focuses on the task of repeatedly matching items to agents, seeking maximin fairness not only ultimately but also after each round; Adams & Segal-Halevi (2026) constructs a balanced sequence of permutations, ensuring fairness after each day; van Lieshout & van Rossum (2025) provides a periodic worker–task scheduling scenario and bounds the price of fairness. If we distribute the fairness of a single decision across a long series of decisions, as in our "count-maximin" policy, this is essentially the concept of equity of attention (Biega et al., 2018). These methods treat "valuation" as a fixed and unchanging number, which is predetermined at the outset of each instance. The model most similar to ours is the one proposed by Caragiannis & Narang (2023): in which an agent's value for an item may depend on how many times they have previously acquired that item. This valuation follows history and is indexed by counts; however, in their model, the state change of "skill + forgetting" is not clearly stated, nor are the concepts of "floor compared to the current endogenous optimal level," "qualification thresholds," and "set of reachable options" mentioned. Our model explores the scenario where these elements are combined: how tasks are assigned now will alter the skills that the allocator needs to optimize later; and fairness concerns whether those "qualified future possibilities" can still be retained and reachable.

Workforce scheduling with learning. The concept that "repetitive task assignment can cultivate skills" has been thoroughly studied in operations research and is a well-established field: Nembhard & Bentefouet (2012) developed a task assignment model with multiple periods and parallel workers, which also took into account each individual's learning and forgetting curves; Shafer et al. (2001) fitted learning–forgetting curves that vary among individuals based on actual data; Hopp & Van Oyen (2004) developed a framework for cross-training; when Azizi et al. (2010) optimized job rotation, they took into account learning, forgetting, and boredom; in the model proposed by Li et al. (2020), skill growth also took into account cooperation within the team; there are also several surveys that have systematically organized the research in this field (De Bruecker et al., 2015; Qin et al., 2015). The first difference from these studies is the "feasibility constraint": it uses "the optimal level achieved by oneself at present" as the reference (our floor, which will automatically move up as a

reference standard, referred to as the "lock-in mechanism"; the conclusions regarding thresholds and window periods in Section 3 are derived from this); the second aspect is the definition of fairness: the fairness we are referring to is "ensuring that each worker can maintain their feasible *future* role set", rather than simply imposing requirements and restrictions on the scheduling plan from the production side, as was done in the past. The issue of "job rotation" versus "specialization" has long been studied in economic theory: Kurino & Kurokawa (2024) developed a dynamic matching model with training, and calculated that there are two profit-optimal schemes. Ortega (2001) stated that job rotation was not intended to enable employees to acquire skills, but rather was a tool used by companies to gauge the true abilities of their employees. As for the concept of "assigning tasks to individuals and allowing them to develop specialized skills for certain types of work," this is a well-established classical theory that has been written about by many prominent scholars (Gibbons & Waldman, 2004; Arrow, 1962; Prendergast, 1993). The above are all calculated as the "optimal equilibrium state" from the perspective of the company; The term *lock-in* itself originates from the economic theory of "increasing returns + path dependence" (Arthur, 1989; David, 1985); it refers to fields such as technology and the market, where seemingly insignificant matters in the early years will eventually become a foregone conclusion. The "threshold results" calculated in our article are to narrow this principle down to "within each individual and each team", and precisely calculate the critical value where "the situation completely changes once it reaches a certain point". "Don't make irreversible decisions; retaining the right to choose is valuable": this is also a classic concept, known as quasi-option value (Arrow & Fisher, 1974); the option set in our research is the "combinatorial version + constraint-compliant version" of this concept, except that we evaluate value based on "performance quality" rather than money.

Rising, rotting, restless, and conservative bandits. If the more one chooses a certain "arm" (i.e., an option/opportunity), the higher the return, then in the improving-bandit model with "total duration known", directly focusing on this arm is the most profitable (Heidari et al., 2016); Patil et al. (2023) studied the disparity issues that arise when opportunities can strengthen arms, as well as the anytime guarantees under the condition of "not knowing the total duration and being able to stop at any time". Additionally, the stochastic line can calculate the guarantee of "regret" (Metelli et al., 2022), be able to study the problem and lower bound of "finding the best arm under a fixed budget" (Mussi et al., 2024), it is also possible to study "evolution triggered by graph structure": that is, one selects an action and can make the other arms linked to it stronger (Genalti et al., 2024). Finally, Song et al. (2026) extended the concept of "increasing returns" to "combinatorial actions" (including matchings); they found that only under the special condition where "returns are simply added together" is it optimal to consistently use the same combined action, not for every combinatorial reward. When it comes to the concept of "decay", there has actually been considerable research in the field of bandit algorithms (also known as multi-armed bandits): Heidari et al. (2016) studied the situation where "arms can become stronger or weaker"; rotting bandits (Levine et al., 2017) specifically discuss the setting where the more arms are pulled, the less reward is given. If the reward depends on "how long the arm hasn't been pulled," then it corresponds to the recharging and blocking model (Immorlica & Kleinberg, 2018; Basu et al., 2019). This concept bears the closest resemblance to the "forgetting dynamics" and revisit-gap (the time between two turns at the same role) analysis discussed in our paper (specifically, in Proposition 8), especially within the context of bandit algorithms. There is another scenario: if the forgetting rate δ is greater than 0, the unplayed pairs will also undergo changes on their own: this is the core characteristic of restless bandits (arms change even while not played) (Whittle, 1988). Migali et al. (2025) studied the situation where rewards increase in this restless setting; however, in our model, when δ equals 0, the entire process is a deterministic rested (skills move only when played) process over matchings. There is also a category with "baseline constraints": for example, conservative bandits (Wu et al., 2016; Kazerouni et al., 2017), which both of those works have done. It requires the learner to ensure that the *cumulative* total reward in each round cannot be less than $(1-\alpha)$ times the *exogenous* baseline: the former's baseline is a preset default arm, while the latter's baseline is a baseline policy. There are also bandits with "fairness constraints", which were proposed by Joseph et al. (2016). Their "fairness constraint of 'choosing based on merit' " has been proven to indeed slow down the learning speed. Our paper's Section 5 reverses the situation: while others' fairness constraints harm the learner, our learner itself aims for optimality, but ends up creating unfairness. There are also numerous studies focusing on "matching while learning numerical values from the matching process": for example, matching while learning (Johari et al., 2021), competing bandits in matching markets (Liu et al., 2020a), fairness and stability in matching markets with bandit learners (Cen & Shah, 2022), and stochastic egalitarian

assignment (Lim et al., 2024). In these studies, the value of a match remains constant; however, it's different here: once a pair is assigned, its subsequent skill will change. The number of assessments m in Theorem 2 is actually somewhat related to two types of research: one is pure exploration over improving arms (Mussi et al., 2024), and the other is combinatorial pure exploration (Chen et al., 2014). However, it has a unique aspect: once this endogenous floor is crossed, mistakes made due to misidentification can never be corrected. Sections 3 to 4 still discuss the scenario of an "omniscient planner": these conclusions of "not feasible" have no relation to "information uncertainty", so it is not due to learning too slowly that things go awry. There is another scenario of "getting stronger by oneself", known as positive externalities, which was mentioned in the study on bandit problems by Shah et al. (2018).

Safe and constrained exploration. There is a method of "safe exploration" (Moldovan & Abbeel, 2012; Sui et al., 2015; Turchetta et al., 2016), which suggests that planners only dare to act in places where "confidence intervals prove absolute safety". It has a recognized Achilles' heel: the safety zone can no longer be expanded, because the actions that can help reduce uncertainty are precisely those that have not yet been proven safe. Our Proposition 4 is to fully articulate this Achilles' heel in our scenario: there is a clearly defined "radius", beyond which there will never be such a thing as a "training deviation" in the safe set; and this is at tolerances where the fully informed planner is provably safe. The first difference is that the benchmark for the constraint is endogenous: this floor follows the team's own best performance; the second point is that it is not the reward that is "starved to death", but rather the future opportunities of others. The planner of Proposition 4 also has no step that expands its safe set (the lineups it can certify), unlike the expander-based methods above (Sui et al., 2015; Turchetta et al., 2016). Conservative exploration (Garcelon et al., 2020) keeps a learner's cumulative performance within a budget of a baseline policy; our floor is a constraint of that form, applied round by round against the team's own best lineup.

Fairness dynamics and performativity. A "feedback loop" is formed between algorithmic decision-making and the real world it affects: in plain terms, the algorithm makes decisions, the reality changes accordingly, and these changes in turn influence the algorithm. This phenomenon has been studied in many fields: such as predictive policing (Ensign et al., 2018), loan approval (Liu et al., 2018), models for repeatedly reducing losses (Hashimoto et al., 2018), long-term simulation (D'Amour et al., 2020), situations involving long-term qualification changes (Zhang et al., 2020), and selection systems where "those who are chosen will have more advantages" (Pokharel et al., 2026); there have also been studies on people deliberately working to "enhance their qualifications" to comply with rules, in which case "merit" is not fixed, as in Coate & Loury (1993); Liu et al. (2020b). The first article, Heidari & Kleinberg (2021), focuses on "opportunity allocation": when someone is given an opportunity, and after receiving it, their socioeconomic status improves, the distribution of the available population changes when opportunities are redistributed later; the second article, Heidari et al. (2019), discusses how decision-making rules guide individuals to make efforts, which in turn alters the distribution of "qualified individuals", meaning that the rules themselves shape the objects that will be evaluated later; the third article, Wen et al. (2021), designs policies with fairness constraints for Markov decision processes (i.e., decision-making models where each step affects the next), where every decision made changes the state generating future features. These three articles share a common premise: the decision-making system itself "produces" the "merit" it uses to make judgments. However, the specific subjects of our research differ: their conclusions are targeted at the *distribution across the entire population*, such as differences in mobility, effort levels, or discounted reward disparities between groups; whereas our research focuses on *each individual within a specific team*, and must also meet the hard condition of a per-round performance floor, as well as a clear irreversibility threshold of "once on, not coming off". Their research neither includes the concept of the option set nor of permanence quantified over all feasible schedules. There is also a research direction called performative prediction (Perdomo et al., 2020), which does not involve "state" (i.e., it does not consider previous influences). Later, Brown et al. (2022) added state to it. Performative reinforcement learning (Mandal et al., 2023) is a broader concept that can incorporate the dynamic situation where "the more practice, the more proficient" and "skills change with policies". There are two differences between the research questions we study and the research conducted by predecessors in that direction: firstly, its solution concepts (performative stability and optimality) evaluate the decision-maker's reward; there is neither the concept of "each individual corresponds to a set of feasible future positions" like $O_i(t, H)$ nor the restriction of "being floor-feasible in each round". Secondly, the phenomenon we focus on is "constrained irreversibility": that is, under the optimal policy, once leaving a certain type of recoverable state, it can never be returned to, unlike

the situation where "repeated retraining eventually leads to an equilibrium state". Additionally, the feedback channels actually studied in previous literature (such as data collection, population composition, and strategic investment) differ from ours (practice assigned by the algorithm).

Option preservation. Numerous studies have already proposed formal "keep your options open" objectives, such as empowerment (Klyubin et al., 2008), reachability preservation and attainable-utility preservation in the field of AI safety (Krakovna et al., 2018; Turner et al., 2020), the concept of "power" in Markov decision processes (Turner et al., 2021), axiomatic rankings of opportunity sets (Pattanaik & Xu, 1990), and long-term fairness in sequential decision processes (Jabbari et al., 2017; Hu & Zhang, 2022; Yin et al., 2023). A study conducted concurrently with ours, Heitzig & Potham (2026), derived axioms for a goal, which is to preserve each individual's future options in a general human–AI assistance scenario. Additionally, reversibility-aware reinforcement learning (Grinsztajn et al., 2021) relies on experiential learning to avoid irreversible state transitions. In the terminology of control theory, the option set is a reachability object with constraints: $O_i(t, H)$ represents a collection of targets, and as long as the qualification region of a target can be reached by a "dynamic system constrained by a floor", it will be included in this set; it is somewhat akin to a viability kernel (the states that can stay feasible forever) (Aubin, 1991), but specifically tailored for individual agents and within a finite horizon.

Team games, roles, and skill estimation. Our choice of this research direction is supported by actual data: Kim et al. (2016) used data from 3.36 million players to empirically confirm a contradiction: whether to assign players to the roles they are best at or to roles that complement their teammates, the two often conflict; Chen et al. (2024) specifically examined the relationship between the flexibility of players' play styles and their final performance; Fan et al. (2024) conducted "rematching by team and position" on a real industrial scale, but the players' skills were treated as fixed inputs; Freeman & Wohn (2019) also documented how esports teams are assembled and how they divide and cooperate in roles. However, none of these studies have touched upon the matter of "long-term role assignment within a persistent team": that is, the role assigned today will directly determine the "measured merit" of tomorrow. In the field of multi-agent reinforcement learning, the role-based approach (Wang et al., 2020; 2021) enables AI to *learn* a set of role divisions, as this can help a team of artificial agents cooperate better; however, the roles we are referring to are institutional and assigned, with real people doing the work. What we want to ask is not "whether roles are useful to the team", but "what impact does such role assignment have on the people who do these tasks". Estimating the true skill level of players from noisy match outcomes is a classical method (Herbrich et al., 2006); our observation model in Section 5 is actually a one-skill-per-pair analogue of it. However, what's new is that which skills generate data at all depends on which roles the assignment gives.

B Complete Proofs

B.1 Proof of Lemma 1

In the symmetric state, each player's skill when playing their own role is d, and their skill when playing any other role is exactly $\bar{s}$, where $\bar{s} \leq d$. Take any assignment M' as an example. Here, k represents the number of players who are not in their own roles. A matching cannot move just one player, so k is either 0 (everyone remains in place) or at least 2. For each moved player, the value at their new role is $\bar{s}$, which is inferior to the value d they would achieve in their own role. Therefore, the total value of M' is $P(M') = nd - k(d - \bar{s})$ (with a total of n players, who would otherwise achieve nd in their own roles, and each additional player moved results in a loss of $(d-\bar{s})$ in value). To find the "most cost-effective deviation," it is definitely most profitable when $k = 2$, that is, when two players exchange roles. No matter how many players are involved in circular swaps or even more chaotic rearrangements, it cannot be more cost-effective, as the loss scales linearly in k, meaning the more players switched, the greater the loss. The floor we require is that the total value cannot be less than $(1 - \varepsilon)nd$, which translates to $nd - 2(d - \bar{s}) \geq (1 - \varepsilon)nd$. Alternatively, it can be written as $2(d - \bar{s}) \leq \varepsilon nd$. □

B.2 PROOF OF THEOREM 1

Sketch. The greedy policy maintains state symmetry, thus requiring only d to move; and only when $2(d - \bar{s}) \leq \varepsilon nd$ is talent swapping within the "floor-feasible" range. First, consider case (a): for every increase of d by 1, the value on the left will increase by 2, while the value on the right will only increase by εn, which is less than 2. The two values will intersect only once and will never meet again. As a result, all off-role skills will forever remain at their initial level $\bar{s}$ and will not reach the θ standard. In case (b), the same swap will be repeated. Here, the deficit is $2\max(d - a_m, 0)$, where a_m is the common skill value of the newly swapped-in players after repeating this operation m times. As this skill value increases, the gap becomes smaller and smaller, while the available budget εP^* becomes larger and larger. Ultimately, the target player–position pair will meet the criteria.

What if d happens to be equal to $\bar{s}$? In that case, no matter how we arrange the players, the total value will be equally good, so every assignment is optimal. Let's just use the matching played by the greedy policy in round 0 as M (if there are multiple, choose any one), and call $M(i)$ "the own role of player i". From round 1 onwards, M becomes the only optimal assignment, and thereafter, the greedy policy always assigns players according to M. At this point, everyone's own-role skills synchronously increase as

$$d(t) \;=\; 1 - (1 - \bar{s})(1 - \eta)^t,$$

climbing higher and higher towards 1; whereas all off-role skills remain unchanged at $\bar{s}$, because they have never played these roles before and there is no forgetting.

Part (a). Since $\varepsilon n < 2(1 - \bar{s})$, we can calculate and obtain $2\bar{s} < 2 - \varepsilon n$. Therefore, the threshold $d^* = 2\bar{s}/(2-\varepsilon n)$ is less than 1. The own-role skill $d(t)$ is increasing towards 1, and it will definitely exceed d^*. Therefore, there exists a first round t_0 such that $d(t_0) > d^*$. At this point, according to Lemma 1, all assignments that are not the greedy one cannot meet the floor, and we can only honestly follow the greedy approach. If we always follow greedy play, this symmetric form will be maintained, and d will continue to increase. The "deficit due to a swap between two players" is $2(d - \bar{s})$, which increases by 2 for every unit increase of d; while the budget we can accept is εnd, which only increases by εn for every unit increase of d. Since $\varepsilon n < 2$, once the deficit exceeds the budget (i.e., the assignment becomes infeasible), it will always exceed it and never return to the previous state. From round t_0 onwards, no feasible schedule will assign anyone an off-role, thus ensuring that everyone's off-role skills remain permanently at $\bar{s} < \theta$: these options become completely lost and can no longer be utilized.

Part (b). Pick any player i and any role r that is not their own role. Assume that this role is owned by player j. The swap τ we are referring to involves having i take on r and having j take on i's own role $M(i)$. In the symmetric state, the two swapped-in skills $s_{i,r}$ and $s_{j,M(i)}$ are both stuck at the same off-role level $\bar{s}$, so a single trajectory can clarify the situation of both. Let us first confirm: regardless of the round in which τ is first attempted, it will be feasible. If $\varepsilon n \geq 2$, this can be seen at a glance; after all, $2(d - \bar{s}) \leq 2d \leq \varepsilon nd$. If $2(1 - \bar{s}) \leq \varepsilon n < 2$, then for all finite t, $d^* = 2\bar{s}/(2 - \varepsilon n) \geq 1 > d(t)$ (when exactly at the boundary $\varepsilon n = 2(1 - \bar{s})$, $d^* = 1$, while $d(t)$ is strictly less than 1). Therefore, Lemma 1 applies.

Now, start from a round where own-role skills equal d and repeatedly play τ. After repeating m times, both swapped-in skills have increased to $a_m = 1 - (1 - \bar{s})(1 - \eta)^m$; the two displaced own-role skills $s_{i,M(i)}$ and $s_{j,r}$ remain frozen at d; every *other* player k continues to play their own role under τ. Therefore, the value of this swap is:

$$P(\tau, m) \;=\; \sum_{k \neq i,j} s_{k,M(k)}(m) \;+\; 2\,a_m.$$

As a reference benchmark, note that the total value of any matching will not exceed the sum of the row maxima one per player: for players $k \neq i, j$, the row maximum is their own-role skill (all other skills of k are $\bar{s}$); for the pair $\{i, j\}$ over the two roles $\{M(i), r\}$, it is $2\max(d, a_m)$. This upper bound is attained (except for this pair, everyone else plays their own role; within this pair, the choice is made based on the higher value between "each doing their own" and "swapping"), so the benchmark and the deficit are respectively:

$$P^*(m) \;=\; \sum_{k \neq i,j} s_{k,M(k)}(m) \;+\; 2\max(d,\, a_m),$$

and hence:

$$P^*(m) - P(\tau, m) \;=\; 2\max(d - a_m,\; 0).$$

From this perspective, the larger m (i.e., the higher a_m), the smaller or unchanged the deficit (it is nonincreasing in m); whereas the budget $\varepsilon P^*(m)$ only increases and never decreases (skills only rise). In the two subcases mentioned earlier, at $m = 0$ the swap is feasible. Afterwards, with each repetition, the swap remains feasible; and after repeating $\lceil \log \frac{1-\theta}{1-\bar{s}} / \log(1-\eta) \rceil$ times, $s_{i,r} = a_m$ exceeds any bar θ less than 1. □

Two remarks. *A bench.* Assuming there are $n' < n$ positions and $n - n'$ idle players, each idle player has skill $\bar{s}$ in all positions, and the idle players are not subject to the floor. So one deviation can mobilize just one player: an idle player replaces the holder of a certain position, with a cost of $d - \bar{s}$ and a corresponding budget of $\varepsilon n' d$; the deviation cost of transferring $k \geq 1$ holders is $k(d - \bar{s})$, so a single-player transfer is the lowest cost. At this point, the argument for Theorem 1(a) can be concluded that when $\varepsilon n' < 1 - \bar{s}$, lock-in occurs: the cost of single-player mobilization increases to $1 - \bar{s}$, while the budget increases to $\varepsilon n'$. There is no conclusion in this article on whether all desired pairs can still be reached when the value is higher than this; in an example of three players and two positions ($\bar{s} = .40$, $\eta = .30$), when $\varepsilon n' = 1 - \bar{s}$, the idle player's target pair can still be reached throughout the entire horizon, but a starter's target pair is not. Therefore, the converse needs a margin. *Pair-specific learning rates.* If the update (1) uses a rate $\eta_{i,r}$ within the $(0, 1)$ interval for each pair, then the conclusion in part (a) remains unchanged: under the greedy policy, each player's own-role skill will still improve to 1, while all off-role skills remain at $\bar{s}$, so the cost of any deviation tends to be at least $2(1 - \bar{s})$, while the budget tends to $\varepsilon n < 2(1 - \bar{s})$. Part (b) is not settled here: when $n = 5$, $\bar{s} = .40$, learning rates are $.9, .9, .01, .01, .01$, and εn is exactly equal to $2(1 - \bar{s})$, the exchange of the two fast players is floor-feasible only in round 0, because their own-role skills outrun the budget, while the skills of the slow players have not yet increased the budget; swapping with a slow player first, followed by a three-cycle, still allows each fast player to complete training in the other's opening role.

B.3 Robustness of the Threshold Under Perturbed Starts

Proposition 5 (Robustness of the threshold)**.** *Let $\delta = 0$ and $\bar{s} < \theta$, and let κ satisfy $0 \leq \kappa \leq \kappa_0 := \min\{\frac{\eta(1-\bar{s})}{4},\; \frac{\theta-\bar{s}}{2},\; \frac{\bar{s}}{3}\}$. Run the greedy policy from any start in which every skill lies in $[\bar{s}-\kappa,\; \bar{s}+\kappa]$.*

(a) If $\varepsilon n < 2(1 - \bar{s}) - 6\kappa$, then after finitely many rounds every deviation from greedy play is infeasible and remains infeasible forever, and every desired role other than a player's round-0 greedy role is permanently lost, for every player.

(b) If $\varepsilon n \geq 2(1 - \bar{s}) + 6\kappa$, no desired role is ever permanently lost.

At $\kappa = 0$ the two conditions meet at the threshold of Theorem 1; the transition region has width at most $12\kappa/n$ in ε.

Proof. We refer to the matching played by the greedy policy in round 0 as M, and the role assigned to player i within M as "player i's own role". If δ equals 0, the skill values corresponding to those unplayed roles will remain unchanged. Therefore, throughout the process of greedy play, the skill values of all off-roles will always remain within the range $[\bar{s} - \kappa, \bar{s} + \kappa]$.

Greedy is locked to M from round 1 on. After round 0, the own-role skill value of player i is at least $\bar{s} - \kappa + \eta(1 - \bar{s} - \kappa)$, which, after conversion, is at least equal to $\bar{s} - \kappa + \eta(1 - \bar{s}) - \eta\kappa$. Compared to the skill value of an off-role (which is at most $\bar{s} + \kappa$), it is at least $\eta(1 - \bar{s}) - (2 + \eta)\kappa$ higher, or simplified to at least $\eta(1 - \bar{s}) - 3\kappa$. Since we initially set $\kappa \leq \eta(1 - \bar{s})/4$, the difference is at least $\eta(1 - \bar{s})/4$, which is definitely a positive number. Any matching that is not M will result in at least two players leaving their own roles, so the total score of the matching will definitely be lower than M. As a result, in each round, M will be greedily chosen, and the own-role skills will become increasingly stronger, leading to a widening gap between them and the off-roles. We denote the initial own-role skill value of player i as ζ_i. After t rounds, their own-role skill $s_{i,M(i)}(t)$ equals $1 - (1 - \zeta_i)(1 - \eta)^t$. Compared to the symmetric trajectory $d(t) = 1 - (1 - \bar{s})(1 - \eta)^t$, the deviation is at most $\kappa(1 - \eta)^t$, and it will definitely not exceed κ.

Deviation costs. In round t, if a "deviation" occurs, that is, k players (k is at least 2) are moved at once, each moved player will replace their own-role skill value (at least $d(t)-\kappa$) with their off-role skill value (at most $\bar{s}+\kappa$). In this way, compared to $P^*(t)=\sum_i s_{i,M(i)}(t)$ (which is the sum of all players' own-role skill values), the deficit of this deviation (that is, the part less than optimal) is at least $k\,(d(t)-\bar{s}-2\kappa)$; as long as $d(t)\geq\bar{s}+2\kappa$ (which holds from round 1 on, since $\kappa\leq\eta(1-\bar{s})/4$ gives $d(1)\geq\bar{s}+4\kappa$), then the deficit is at least $2\,(d(t)-\bar{s}-2\kappa)$. If only the positions of two players i and j are swapped, the maximum deficit for this transposition is $2(d(t)-\bar{s})+4\kappa$. Let us now examine the range of the budget: $\varepsilon n\,(d(t)-\kappa)\leq\varepsilon P^*(t)\leq\varepsilon n\,(d(t)+\kappa)$.

Part (a). Let us define a function $f(d):=2(d-\bar{s}-2\kappa)-\varepsilon n\,(d+\kappa)$. After simplifying, we obtain $(2-\varepsilon n)\,d-2\bar{s}-(4+\varepsilon n)\kappa$. Since $\varepsilon n<2(1-\bar{s})-6\kappa\leq 2$, $f(d)$ strictly increases as d increases. Calculate when $d=1$: $f(1)=\big[2(1-\bar{s})-\varepsilon n\big]-(4+\varepsilon n)\kappa$. Combining with $\varepsilon n\leq 2$, it can be inferred that $(4+\varepsilon n)\kappa\leq 6\kappa$. Substituting the values, we find that $f(1)>6\kappa-6\kappa=0$. Because $d(t)$ is increasing towards 1, there must be a first round t_0 such that $f(d(t_0))>0$; from this round onwards, any deviation will have a deficit larger than the budget, in plain terms, it is "not enough to make up the gap", so one can only follow the greedy policy honestly. $d(t)$ will continue to increase, and $f(d(t))$ will always remain positive. This state of infeasibility is absorbing. Furthermore, all off-role skill values are frozen at a maximum of $\bar{s}+\kappa$, which is less than θ (because $\bar{s}+\kappa\leq\bar{s}+(\theta-\bar{s})/2<\theta$). Therefore, except for their own roles, players can never reach other desired roles: those roles are permanently lost.

Part (b). Take any round t, any player i, and any role $r=M(j)\neq M(i)$ (where r is the own role of player j). Let's try swapping i and j (denote this transposition as τ). When attempting this transposition for the first time, the maximum deficit is $2(d(t)-\bar{s})+4\kappa$, while the minimum budget is $\varepsilon n(d(t)-\kappa)$. If $\varepsilon n<2$, we define a function $g(d):=\varepsilon n(d-\kappa)-2(d-\bar{s})-4\kappa$, which decreases as d increases. When calculating $d=1$: $g(1)=\big[\varepsilon n-2(1-\bar{s})\big]-(4+\varepsilon n)\kappa\geq 6\kappa-6\kappa=0$, so this transposition τ is feasible at every t (the budget is sufficient to cover the deficit). Firstly, if $\varepsilon n\geq 2$, we have at least $2(d(t)-\kappa)$ in our "budget", while the "deficit" is at most $2(d(t)-\bar{s}+2\kappa)$; this number can be bounded further, since $\kappa\leq\bar{s}/3$. Substituting this value, we get $2(d(t)-\bar{s}+2\kappa)\leq 2(d(t)-3\kappa+2\kappa)\leq 2(d(t)-\kappa)$, which is at most the budget. Therefore, this matter can be successfully achieved (meaning it is "feasible"). Next, we will repeatedly use the swap τ, with each use constituting a round. After using it m times, we refer to it as the state after m repetitions. If $\delta=0$, then for every "edge" (understood as a player–role pairing) used in τ, the corresponding skill will only increase and never decrease; for those edges not in τ, the skill will remain unchanged. In this way, no matter which matching M' is chosen for comparison, the difference $P(M',m)-P(\tau,m)$ will only become smaller or stay unchanged. Why is this so? Because the edges shared by M' and τ cancel each other out when subtracted; the remaining edge skills of M' remain unchanged, while those of τ increase, which naturally leads to a smaller or unchanged difference. Compare all M' and take the maximum difference, which is $P^*(m)-P(\tau,m)$. This difference will only become smaller or stay unchanged. Moreover, the value of each matching itself only increases and never decreases. Therefore, $P^*(m)$ (which represents the highest value among all matchings) will only increase, and our budget $\varepsilon P^*(m)$ will also only increase and never decrease. This swap was feasible at the very beginning ($m=0$), and it remained feasible thereafter until $s_{i,r}$ (representing the skill of player i in role r) crossed any bar θ below 1: in any case, the desired role would never be permanently lost. □

B.4 The One-Round Instance, in Full

Example 1 (One-round permanent lock-in). *Let $n=2$ with roles A and B, $\eta=0.5$, $\delta=0$, $\theta=0.9$, $\varepsilon=0.35$, $s_{1,A}=s_{2,B}=0.15$, $s_{1,B}=s_{2,A}=0.10$, with player 1 desiring role B; the two assignments are the* diagonal *($1\to A$, $2\to B$) and the* swap. *The swap is feasible at $t=0$; after one greedy round it is infeasible at every future round, and player 1's desired role B is permanently lost in round 1.*

Let's calculate and it will become clear: in round 0, the total value of the diagonal was $0.15+0.15=0.30$, which was the highest at that time, that is, $P^*(0)=0.30$; the total value of the swap is $0.10+0.10=0.20$, while the feasible requirement is "not less than $(1-0.35)$ times the maximum value", which is $0.30\times 0.65=0.195$. Since 0.20 is greater than 0.195, the swap is feasible. However, in a greedy approach, one would choose the option with the highest immediate value, that is, selecting the diagonal; according to the learning formula (i.e., equation (1)), both skills used in the diagonal will increase: $0.15+0.5\times 0.85=0.575$, both reaching 0.575; therefore, the highest value

$P^*(1)$ in round 1 is $0.575+0.575 = 1.15$; what about the value of the swap? Since $\delta = 0$ and neither of the two off-roles involved in the swap has been played before, the value remains unchanged at 0.20. At this point, we can calculate the feasible requirement: $1.15 \times 0.65 = 0.7475$. 0.20 is much smaller than 0.7475, so the swap is completely infeasible. In round 1, only the diagonal is feasible, and this principle holds true for every subsequent round (just follow this logic inductively): playing the diagonal will only increase the value of P^*; conversely, the swap's value will always remain stuck at 0.20, without any movement. So, no matter when it is, $s_{1,B}$ will always be 0.10, which is much smaller than $\theta = 0.9$: player 1 will never be able to obtain role B, and will permanently lose it. □

B.5 Proof of Proposition 1

Let us discuss the rule for whether a transposition can be used in round t: it must satisfy $d(t) \le d^*$ (Lemma 1). Here, $d(t) = 1 - (1 - \bar{s})(1 - \eta)^t$, which becomes larger as t increases. Why mention transpositions? Because they are the cheapest among all deviations. When can some deviation be used? It is feasible exactly when the condition $(1 - \eta)^t \ge (1 - d^*)/(1 - \bar{s})$ is met; both sides of the inequality lie in $(0, 1]$. Therefore, we take the logarithm of both sides and divide by $\ln(1 - \eta)$ (this number is negative), and the inequality sign flips. Finally, we get $t \le t^*$, with t^* as displayed. Therefore, the feasible rounds are $t = 0, 1, \ldots,$ up to $\lfloor t^* \rfloor$; in other words, all rounds where $t < W$, where $W = \lfloor t^* \rfloor + 1$. If t^* is not an integer, W equals $\lceil t^* \rceil$; if t^* is exactly an integer, then $d(t)$ in the boundary round $t = t^*$ is exactly equal to d^*, which is still feasible, and $W = t^* + 1$ also counts this round. As long as a chosen transposition is begun in any feasible round, it can be sustained until qualification: this relies on the cancellation argument in the proof of Theorem 1(b). After the first swap round, there is no longer a need for symmetry: the deficit $2\max(d - a_m, 0)$ will only decrease or remain unchanged, and the budget in hand will only increase or remain unchanged, which is definitely sufficient. Let us discuss the monotonicity of W: the larger ε is, the larger d^* becomes, and the smaller (more negative) the numerator $\ln((1 - d^*)/(1 - \bar{s}))$ is, leading to a corresponding increase in t^*; the larger η is, the more negative the denominator $\ln(1 - \eta)$ becomes, and t^* will decrease; and $W = \lfloor t^* \rfloor + 1$ is nondecreasing in t^*. Finally, for $\bar{s} > 0$, as $\varepsilon n \uparrow 2(1 - \bar{s})$ from below, $d^* \uparrow 1$ from below, the numerator tends towards $-\infty$, $t^* \to \infty$, and $W \to \infty$. □

B.6 Proof of Proposition 2

We first present a perturbed-start form; Proposition 2 is the special case where the perturbation is 0 (i.e., $\kappa = 0$). Let us assume that the initial skill values of all players fall within the interval $[\bar{s} - \kappa,\ \bar{s} + \kappa]$ and the upper limit $\bar{s} + \kappa$ has not yet reached the bar θ; meanwhile, the following condition must also be met:

$$(4 + \varepsilon n)\,\kappa \ <\ 2\,\eta\,(1 - \bar{s}) \ -\ \varepsilon\, n\,\big(\bar{s} + \eta(1 - \bar{s})\big). \tag{3}$$

Condition (3) forces $\varepsilon n < 2$ and $\kappa < \eta(1 - \bar{s})/2$; using the parameters from the tight-floor cell of Section 6 (i.e., $\varepsilon = 0.05$, $\eta = 0.15$, and $n = 5$) for calculation, this condition is satisfied as long as κ is less than 0.0135. We denote the nonempty set consisting of "all feasible assignments in round 0" as $\mathcal{F}_0$.

State. Assume we follow a feasible schedule: the first t rounds (i.e., rounds 0 to $t - 1$, where t is at least 1) consistently use the assignment M_0, and then record each player's initial skill value at their role assigned by M_0 as $\zeta_i = s_{i,M_0(i)}(0)$. According to the learning formula (1) and forgetting formula (2), calculate up to round t: the skill level of each player in their consistently played M_0 role is given by $s_{i,M_0(i)}(t) = 1 - (1 - \zeta_i)(1 - \eta)^t$, and the difference between this value and $d(t)$ will not exceed κ; for other roles that have not been played, $r \ne M_0(i)$, the skill value is $s_{i,r}(t) = s_{i,r}(0)(1 - \delta)^t$, which is at most $\bar{s} + \kappa$. The impact of forgetting is actually only reflected in the last inequality above, and forgetting will only cause skills to decline, which in turn will only make this inequality more valid; therefore, all the following derivations, regardless of the value of the forgetting rate δ in $[0, 1)$, will still hold true.

M_0 is the unique optimum at every $t \ge 1$. At this point, each player's "skill in their own consistently played role" is at least $d(1) - \kappa = \bar{s} + \eta(1 - \bar{s}) - \kappa$, while their "skills in other roles they have not played" are at most $\bar{s} + \kappa$. Therefore, the gap between a player's "own-role skill minus other skills" is at least $\eta(1 - \bar{s}) - 2\kappa$, and this difference is positive. A matching cannot move only

one player; therefore, any assignment M that differs from M_0 must involve at least $k \geq 2$ players being moved from their M_0-roles. Consequently, the total value $P(M,t)$ of M in round t is at least $k\,(\eta(1-\bar{s})-2\kappa)$ less than the total value $P(M_0,t)$ of M_0, that is, $P(M,t) < P(M_0,t)$; and the optimal total value $P^*(t) = \sum_i s_{i,M_0(i)}(t)$ in round t is the total value of M_0, which is at most $n\,(d(t)+\kappa)$.

Every deviation is infeasible at every $t \geq 1$. For any M that is different from M_0, the deficit against the optimum is at least $2\,(d(t)-\bar{s}-2\kappa)$; and the budget is at most $\varepsilon n\,(d(t)+\kappa)$. We denote the difference between the two as $f(d) = 2(d-\bar{s}-2\kappa) - \varepsilon n\,(d+\kappa)$. Since $\varepsilon n < 2$, the larger d is, the larger $f(d)$ becomes (it is strictly increasing). Condition (3) simply states that $f(d(1)) > 0$; and since $d(t) \geq d(1)$ when $t \geq 1$, in every round from the first onwards, any assignment other than M_0 violates the floor.

Induction. In round 0, according to the rules defined in Definition 1, an assignment M_0 belonging to $\mathcal{F}_0$ will be adopted; similarly, as long as $t \geq 1$, the only feasible assignment is always M_0 (it is feasible because it is optimal), so the entire scheduling arrangement remains unchanged from start to finish. Conversely, any fixed schedule based on a certain $M \in \mathcal{F}_0$ is feasible: firstly, it meets the requirements in round 0, and secondly, it remains optimal in subsequent rounds. The skills corresponding to roles that have not been assigned will never improve and will always remain below θ; moreover, from any moment $t \geq 1$ onwards, no matter how the arrangement is subsequently made, this situation cannot be changed, because the inductive logic just mentioned can be directly applied verbatim to the subsequent process: every desired role other than $M_0(i)$ lies outside $O_i(t,H)$ for every H, permanently lost at round 1. In round 0, as long as the time span is long enough for a certain own role to qualify, then $O_i(0,H)$ equals the intersection $D_i \cap \{M(i) : M \in \mathcal{F}_0\}$: in plain terms, all desired roles that can be obtained through a feasible initial assignment still have opportunities, and in each schedule, each player can only retain at most one of them.

Sharpness at $\delta = 0$, $\kappa = 0$. If $\varepsilon n(\bar{s} + \eta(1-\bar{s})) \geq 2\eta(1-\bar{s})$, then starting from a symmetric initial state, regardless of the initial assignment, in the state of round 1, the level of the own role is $d(1)$ and all other skills are $\bar{s}$; the deficit $2\eta(1-\bar{s})$ brought by a transposition fits within the budget $\varepsilon n\, d(1)$, so the transposition is feasible, and the cancellation argument in the proof of Theorem 1(b) can support this transposition until the role qualifies. Nothing is completely lost in round 1, so the freeze does not hold. □

Example 1 in Appendix B.4 works one two-player instance of the freeze through by hand.

When the perturbation value κ is set to 0.03 in the experiment, the worst-case form of condition (3) does not hold, and the freeze can be broken by an adversarial initial state within this range. The freeze certificates in Section 6.1 cover 250 policy–seed trajectories selected from the tight/fast cell in testing, confirming that each one eventually reaches a singleton feasible set.

B.7 QUALIFICATION-SCHEDULING COMPLEXITY

We allow each pairing of "player + role" to have a unique learning rate $\eta_{i,r}$, which ranges over $[0,1)$ (widening $\eta \in (0,1)$ of Section 2). We set the forgetting rate δ to 0 (meaning that learned skills do not decay at all). What is the input for this QUALIFICATION-SCHEDULING problem? It will clearly list all the designated player–role pairs (the pairs that must qualify), as well as provide initial values for all skills, learning rates, and the two parameters θ and ε. These numbers are all rationals written in binary; only the total number of rounds T is given in unary. All numerical comparisons are calculated using exact rational arithmetic without rounding. The core of this problem is to ask whether it is possible to arrange a floor-feasible schedule with a total number of rounds not exceeding T, such that all designated pairs are qualified in the terminal state. Because we have set $\delta = 0$, once qualification is obtained, it will not be lost. Therefore, this matter can be equated to: by the deadline T, all designated pairs must be trained until qualification is obtained.

Theorem 3 (Complexity of qualification scheduling). *(a) For every constant n,* QUALIFICATION-SCHEDULING *is polynomial-time solvable.*

(b) When n is part of the input, it is NP-complete already at $T = 2$ and strongly NP-complete for general horizons.

Proof of the fixed-n statement. Because $\delta = 0$, skills will only improve through practice, and the order in which they are practiced does not affect the outcome: let $k_{i,r}(t)$ denote the total number of times player i practices role r before round t. The skill value can be directly calculated using this number: $1 - s_{i,r}(t) = (1 - s_{i,r}(0))(1 - \eta_{i,r})^{k_{i,r}(t)}$. Let us define the "state" after round t as this count matrix $\big(k_{i,r}(t)\big)_{i,r}$, that is, a table listing how often each player has practiced each role; each number is at most T (the total number of rounds), so if n is fixed, then all possible states add up to at most $(T+1)^{n^2}$ types. Since n is a constant, when T is input in unary, the scale of this number is polynomial. The current round can be deduced from a given state by simply adding up all the entries and dividing by n ($t = \frac{1}{n}\sum_{i,r} k_{i,r}$, as each round involves exactly n players practicing a different role). Furthermore, from this state, one can calculate all skill values, consider all $n! = O(1)$ possible assignments (i.e., who does what in this round), identify the optimal P^*, as well as list all floor-feasible assignments. The time taken for each of these steps is polynomial in T and the size of the input; after all, when n is fixed, the number of assignment possibilities remains constant. Even if the number of digits in the exact skill rationals increases with T, the computation time is still polynomial. From a state in round t, the states reachable in round $t+1$ are found as follows: find a floor-feasible assignment, add 1 to the n entries corresponding to that assignment, and the new state obtained is the one that can be reached. How can we find all the reachable states? We start with a matrix filled with zeros and search layer by layer (i.e., breadth-first search), up to a maximum of T layers. By doing so, all reachable states can be found, and the time taken is still polynomial. What is meant by "a yes-instance"? As long as we find a state in any layer not exceeding layer T where the skill $s_{i,r}$ corresponding to each designated pair (i, r) is $\geq \theta$, then it is a yes-instance. When we search, we can directly check for this. □

B.8 Proof of Proposition 3

Let us first give names to the two edge-disjoint (sharing no player–role edges) alternating matchings, one called A and the other called B. What does this mean? It means that for each player, the role assigned to them in A is definitely different from the one assigned to them in B, $A(i) \neq B(i)$. The two roles that each player desires are the one given to them in A and the one given to them in B, which are different; apart from the skills corresponding to these two roles, they never play any other role and always maintain the initial level of skill $\bar{s}$. The rounds are numbered starting from 0 as in the model: even rounds are played using matching A, while odd rounds are played using matching B. In round 0, A is played (initially, everyone is symmetric, and all assignments are equally good, so A is definitely feasible). After the round, all A-skills become $\bar{s} + \eta(1 - \bar{s})$, while all B-skills remain at their original $\bar{s}$. The optimal assignment is A, where the total value P^* equals $n(\bar{s} + \eta(1 - \bar{s}))$. When can B also be feasible? Exactly when:

$$n\bar{s} \;\geq\; (1-\varepsilon)\, n\big(\bar{s} + \eta(1-\bar{s})\big) \;\iff\; \varepsilon \;\geq\; \frac{\eta(1-\bar{s})}{\bar{s} + \eta(1-\bar{s})} \;=\; \varepsilon_{\mathrm{rot}}.$$

For the subsequent rounds, we will proceed two rounds at a time, alternating between A and B. We define a sequence: $x_0 = \bar{s}$, and then $x_j = x_{j-1} + \eta(1 - x_{j-1})$. This sequence is increasing and approaching 1. Let us state a claim: before the j-th use of B (i.e., before round $2j - 1$), the level of all A-skills is x_j, and the level of all B-skills is x_{j-1}. At $j = 1$ this was just calculated above; it matches perfectly. What about after that? First, play round B, and the skill level of the B-roles will increase to x_j (the same as the A-roles, a tie); then play round A again, and the skill level of the A-roles will increase to x_{j+1}. At this point, j has become $j + 1$, and the claim still holds, because when $\delta = 0$, the skill level of the matching that is resting will not change. In the tie rounds (i.e., the A-rounds after the first), the deficit is 0, so only the B-rounds bind: the optimal assignment at this time is A, with the corresponding $P^* = n\, x_j$; if B is chosen, the deficit becomes $n(x_j - x_{j-1}) = n\,\eta(1 - x_{j-1})$, which strictly decreases as j increases, while the budget $\varepsilon\, n\, x_j$ strictly increases. Therefore, as long as the first B-round satisfies the feasibility requirement (i.e., the case where $j = 1$, corresponding to the displayed condition $\varepsilon \geq \varepsilon_{\mathrm{rot}}$), every subsequent alternation round will also satisfy the feasibility requirement. If the alternation is maintained, each desired skill will approach 1 along the values x_j. Regardless of how close the bar θ is to 1 (as long as it is less than 1), it can be crossed within a finite number of rounds. When $\delta = 0$, once a skill reaches qualification, it will never fall below it. Therefore, when $\varepsilon \geq \varepsilon_{\mathrm{rot}}$, as long as the alternation is maintained until all desired pairs qualify, every desired option is preserved; thereafter, even if any floor-feasible policy is adopted, the already qualified options will not be lost in the sense of

Definition 2. Playability is another matter: if the alternation stops early and the team specializes in A, the A-skills keep rising while the B-skills stand still, and once $2(x_A - x_B) > \varepsilon n x_A$ no floor-feasible lineup plays a B-role again, although every B-pair stays qualified; Proposition 7 in Appendix B.9 says how long the alternation must run to avoid this. □

B.9 The Price of Optionality: Statement and Proof

The rotation of Proposition 3 is not only feasible above $\varepsilon_{\mathrm{rot}}$; among all schedules that qualify every desired pair, it is the cheapest.

Proposition 6 (Price of optionality)**.** *Assume the setting of Proposition 3: symmetric start at $\bar{s} < \theta$, $D_i = \{A(i), B(i)\}$ with $A(i) \neq B(i)$, and $\delta = 0$; let $T \geq 2k_\theta$. Write $x_j = 1 - (1-\bar{s})(1-\eta)^j$ for the skill of a pair after j practices and $V(S) = \sum_{t<T} P(M_t, t)$ for the cumulative value of a T-round schedule $S = (M_0, \ldots, M_{T-1})$; greedy has $V_{\mathrm{g}} = n \sum_{j<T} x_j$. Let $L_{\min}(T)$ be the smallest value of $V_{\mathrm{g}} - V(S)$ over all T-round schedules S, floor-feasible or not, after which every desired pair is qualified. Then*

$$L_{\min}(T) \;=\; \frac{n\,(1-\bar{s})\,\big(1-(1-\eta)^{k_\theta}\big)\big(1-(1-\eta)^{T-k_\theta}\big)}{\eta},$$

and $V_{\mathrm{g}} - V(S) = L_{\min}(T)$ exactly when every player i plays only $A(i)$ and $B(i)$, one of them exactly k_θ times. The alternation $A, B, \ldots, A, B$ for $2k_\theta$ rounds followed by A is such a schedule; it is floor-feasible if and only if $\varepsilon \geq \varepsilon_{\mathrm{rot}}$. The price of optionality *is* $\mathrm{PoO}(T) := L_{\min}(T)/V_{\mathrm{g}}$.

Counts determine value. With $\delta = 0$ a skill changes only when its pair is played, so a pair that has been played j times has skill x_j, whatever the order of play. Let $k_{i,r}$ be the number of rounds of S in which player i plays role r; then $\sum_r k_{i,r} = T$. At the j-th play of pair (i, r) the pre-update skill is x_{j-1}, so player i contributes $\sum_r g(k_{i,r})$ to $V(S)$, where $g(k) := \sum_{j<k} x_j$. Pair (i, r) is qualified at the end exactly when $x_{k_{i,r}} \geq \theta$, that is, when $k_{i,r} \geq k_\theta$.

One player at a time. Fix i and consider the relaxed problem of maximizing $\sum_r g(k_r)$ over nonnegative integers with $\sum_r k_r = T$, $k_{A(i)} \geq k_\theta$, and $k_{B(i)} \geq k_\theta$; the matching structure across players is ignored, which can only enlarge the maximum. Since x_j strictly increases, moving one practice from a role with count $c \geq 1$ to a role with count $c' \geq c$ changes the sum by $x_{c'} - x_{c-1} > 0$. Let k^* be a maximizer and suppose some role $r \notin \{A(i), B(i)\}$ has $k^*_r > 0$. Let r' be the desired role with the larger count. If $k^*_{r'} \geq k^*_r$, moving one practice from r to r' strictly improves the sum, a contradiction. If $k^*_{r'} < k^*_r$, exchanging the two counts leaves the sum unchanged and keeps both constraints (now $k_{r'} = k^*_r > k^*_{r'} \geq k_\theta$), after which the same move strictly improves it, again a contradiction. Hence a maximizer uses only $A(i)$ and $B(i)$. Writing $k_{B(i)} = k$ and $k_{A(i)} = T - k$ with $k_\theta \leq k \leq T - k_\theta$, the value is $F(k) = g(T-k) + g(k)$, whose increment $F(k+1) - F(k) = x_k - x_{T-k-1}$ is negative for $k < (T-1)/2$ and positive for $k > (T-1)/2$. So F is largest only at the two endpoints $k = k_\theta$ and $k = T - k_\theta$, which tie. Therefore player i contributes at most $g(T-k_\theta) + g(k_\theta)$, with equality exactly at these two count vectors.

Summing over players. Adding over i gives $V(S) \leq n\,[g(T-k_\theta) + g(k_\theta)]$, while $V_{\mathrm{g}} = n\,g(T)$, so $V_{\mathrm{g}} - V(S) \geq n\,[g(T) - g(T-k_\theta) - g(k_\theta)]$. From $g(k) = k - (1-\bar{s})\big(1-(1-\eta)^k\big)/\eta$,

$$\begin{aligned} g(T) - g(T-k_\theta) - g(k_\theta) &= \frac{1-\bar{s}}{\eta}\Big[1 - (1-\eta)^{k_\theta} - (1-\eta)^{T-k_\theta} + (1-\eta)^{T}\Big] \\ &= \frac{(1-\bar{s})\big(1-(1-\eta)^{k_\theta}\big)\big(1-(1-\eta)^{T-k_\theta}\big)}{\eta}, \end{aligned}$$

which is the displayed $L_{\min}(T)/n$; equality holds exactly when every player is at an endpoint.

Attainment. In the alternation $A, B, \ldots, A, B$ for $2k_\theta$ rounds every player plays each desired role k_θ times; the remaining $T - 2k_\theta$ rounds of A give counts $(T - k_\theta,\, k_\theta)$, so the bound is met. By Proposition 3 (Appendix B.8) the alternation rounds are floor-feasible if and only if $\varepsilon \geq \varepsilon_{\mathrm{rot}}$, the first B-round being the binding one; after round $2k_\theta$ every A-skill is at least the corresponding B-skill, so A is optimal and floor-feasible thereafter. □

The price of playability. If the alternation never stops, every desired pair stays both qualified and playable, and by the same count argument the loss against greedy over T rounds is $n\,[g(T) -$

$g(\lceil T/2 \rceil) - g(\lfloor T/2 \rfloor)]$; at $n = 5$, $\bar{s} = .40$, $\theta = .70$, $T = 100$ this is 51.1 (11.6%) for $\eta = .05$ and 20.0 (4.2%) for $\eta = .15$. The alternation stays floor-feasible in every round above ε_{rot} (Appendix B.8), so this is attainable. It is not the cheapest way, though: the team can stop alternating after enough rounds, and the pairs it trained stay playable on their own.

Proposition 7 (Finite alternation keeps every pair playable)**.** *Assume the setting of Proposition 6 with $\varepsilon \geq \varepsilon_{\text{rot}}$, and let $K \geq k_\theta$ be an integer for which at least one of the two conditions holds:*

(i) $x_K \geq 1 - \varepsilon$, or

(ii) $\varepsilon n \geq 2 - x_K - \bar{s}$.

Alternate $A, B, \ldots, A, B$ for $2K$ rounds and play A in every later round. This schedule is floor-feasible, and from round $2K$ on every desired pair is qualified and playable in every round. Its loss against greedy over $T \geq 2K$ rounds is $n\,[g(T) - g(T-K) - g(K)]$, which is smaller than the loss of the never-ending alternation whenever $T > 2K + 1$.

Proof. Feasibility. By Proposition 3 (Appendix B.8), each of the $2K$ alternation rounds is floor-feasible because $\varepsilon \geq \varepsilon_{\text{rot}}$. After round $2K$ every A-pair has been played at least as often as the corresponding B-pair, so A is an optimal lineup and clears the floor in every later round.

Playability. After round $2K$ every B-pair has been played exactly K times and is not played again, so its skill stays at x_K (skills are durable, $\delta = 0$), and $K \geq k_\theta$ makes it qualified. Every A-pair is qualified and fielded by A itself. It remains to field each B-pair $(i, B(i))$ by some lineup that clears the floor. Write d for the common A-skill in the round considered; then $P^* = nd$ (no lineup can beat the row maxima) and $d \geq x_K$. Under condition (i), take the lineup B itself: its value is $nx_K \geq n(1-\varepsilon) \geq (1-\varepsilon)P^*$, since no lineup is worth more than n. Under condition (ii), take the lineup that moves player i to $B(i)$, moves the player who holds $B(i)$ under A to $A(i)$, and leaves everyone else in A. The first moved player contributes x_K; the second contributes at least $\bar{s}$; the other $n-2$ players contribute d each. The floor asks for $(n-2)d + x_K + \bar{s} \geq (1-\varepsilon)nd$, that is, $(2-\varepsilon n)d \leq x_K + \bar{s}$. When $\varepsilon n \leq 2$ the left side grows with d and is at most $2 - \varepsilon n \leq x_K + \bar{s}$ at $d = 1$; when $\varepsilon n > 2$ it is negative. So the swap lineup clears the floor in every later round, whatever the pattern in which B moves the players.

Loss. Every player plays $A(i)$ in $T - K$ rounds and $B(i)$ in K rounds, so the count argument of Proposition 6 gives $V(S) = n\,[g(T-K) + g(K)]$ and a loss of $n\,[g(T) - g(T-K) - g(K)]$. The increments of g are the skills x_k, which increase in k, so g is convex: of two count pairs with the same sum, the more unequal one has the larger value of g plus g. When $T > 2K + 1$ the pair $(T-K, K)$ is more unequal than $(\lceil T/2 \rceil, \lfloor T/2 \rfloor)$, so the finite schedule loses strictly less than the never-ending alternation; when $T = 2K + 1$ the two coincide. □

Numbers. At $n = 5$, $\bar{s} = .40$, $\theta = .70$ and $T = 100$ (exact rational arithmetic), condition (ii) reads $x_K \geq 1.6 - 5\varepsilon$. For $\eta = .15$ we have $k_\theta = 5$ and $x_5 = .7338$, so condition (ii) needs $\varepsilon \geq .1733$, which every $\varepsilon \geq \varepsilon_{\text{rot}} = .1837$ satisfies: the schedule of Proposition 6 itself keeps every pair playable, and the price of playability equals the price of optionality, 11.1 (2.3%). For $\eta = .05$ we have $k_\theta = 14$ and $x_{14} = .7074$, so $K = k_\theta$ suffices from $\varepsilon = .179$ on and the two prices again coincide at 6.9%; below that, the smallest admissible K is the first with $x_K \geq \min(1-\varepsilon,\, 1.6-5\varepsilon)$. At $\varepsilon = .10$ this is $K = 35$ ($x_{35} = .9003$), with loss 48.3 (11.0%) against 51.1 (11.6%) for the never-ending alternation; at $\varepsilon = \varepsilon_{\text{rot}} = .0698$ it is $K = 42$ and 50.3 (11.4%). Whether a still cheaper schedule keeps every pair playable at $\eta = .05$ and $\varepsilon = .10$ is open; its loss lies between 6.9% and 11.0%.

B.10 Sharing Capacity Under Forgetting: Statement and Proof

Proposition 8 (Sharing capacity under forgetting)**.** *Suppose q players all hold skill at least θ at one common role r, in every round from some time on, under $\delta > 0$. Then $q \leq k_{\max} = \lfloor 1 + \eta(1-\theta)/(\delta\theta) \rfloor$: at most $k_{\max}$ players can share one role, regardless of the floor and of how the turns are arranged.*

This sharing capacity (how many players can hold one role at once) represents the upper limit of the number of players who "maintain qualification simultaneously and indefinitely". It does not mean

that all options for these q players are gone: in Definition 2, it is possible to test each player under a different hypothetical continuation.

Assume that a player returns to the same role every k rounds, repeating this cycle indefinitely. Their skill forms a cycle (increasing slightly with each practice, followed by a gradual decline in the subsequent $k-1$ rounds); to hold the role at qualification, their skill must be $s \geq \theta$ in each round. The most dangerous moment in this cycle is the trough (the lowest point of the skill) just before the next practice; therefore, the condition for gap-k cycling (returning to the role every k rounds) to sustain qualification is precisely $\rho(k)(1-\delta)^{k-1} \geq \theta$, where $\rho(k)$ represents the peak skill level during this cycle. Next, let's calculate this value.

Derivation of $\rho(k)$ for the cyclic case: suppose one practices a skill once and reaches a peak level p. If one does not practice for $k-1$ rounds, the skill will continue to decline. How much will it decline? It will decline to $p(1-\delta)^{k-1}$. At this point, if one practices again, the new level after practice will be $(1-\eta)(1-\delta)^{k-1}p + \eta$. Every time, starting from the previous peak value p, after training, it becomes $(1-\eta)(1-\delta)^{k-1}p + \eta$. This map $p \mapsto (1-\eta)(1-\delta)^{k-1}p + \eta$ is a contraction, and the peak value gradually approaches a fixed point monotonically, getting closer and closer without any deviation. This fixed point is $\rho(k) = \eta/(1-(1-\eta)(1-\delta)^{k-1})$. Furthermore, the larger k is (that is, the longer the interval between two practices), the smaller this $\rho(k)$ becomes: it is strictly decreasing in k.

Next, let's discuss why there is an upper limit to the number of players. This requires a "drift" argument. What does it mean? It means that regardless of whether everyone's turn-taking is regular or not, it can still be calculated. Assume that there are currently q players, and from a certain point onwards, their skill $s_{i,r}$ at the role r is not lower than θ in each round. Let's choose any one of them and observe their performance in a stretch of G rounds: in these G rounds, they practice the role r exactly c times. How much can one's skill increase with each practice? It is $\eta(1-s)$, because skill can only reach a maximum of 1. Currently, one's skill is already at least θ, so the increase must not exceed $\eta(1-\theta)$. What about the remaining $G-c$ rounds that have not been practiced? How much will skill decrease in each round? It is δs. Since skill is already at least θ, then the decrease in each round must be at least $\delta\theta$. However, the level has always been stuck in the bounded interval $[\theta, 1]$ and hasn't fallen out, what does this indicate? It indicates that the gains must cover the losses up to a constant:

$$c\,\eta(1-\theta) \;\geq\; (G-c)\,\delta\theta \;-\; 1.$$

Let's calculate a "long-run practice frequency" $f_i = \liminf_{G\to\infty} c_i(G)/G$ for each of the q players. This is the number of times $c_i(G)$ they practice their role r in the first G rounds, divided by G, taking the limit inferior as G becomes infinitely large. Why use the limit inferior? Because everyone's turn-taking may not follow a regular pattern, with varying amounts of practice, so using this is more reliable. The displayed inequality holds at every G. If we divide both sides by G and pass to the limit along a sequence realizing f_i, we obtain: $f_i\,\eta(1-\theta) \geq (1-f_i)\,\delta\theta$. After rearranging, we have that f_i must be at least $\delta\theta/(\eta(1-\theta)+\delta\theta)$. However, it should be remembered that only one player can play the role r in each round at most, so regardless of which G it is, the total number of practices by all players will not exceed G: $\sum_i c_i(G) \leq G$. Since the limit inferior of a sum is at least the sum of the limits inferior, the sum of all players' f_i will also not exceed the limit inferior of the total number of practices divided by G, that is, $\sum_i f_i \leq \liminf_{G\to\infty} \sum_i c_i(G)/G \leq 1$. This leads to the following conclusion: if none of the f_i of the q players is less than that number, and their sum is still ≤ 1, then q can at most be equal to $1+\eta(1-\theta)/(\delta\theta)$; this is the upper limit of the number of players. Let us first clarify its theoretical upper bound: $q \leq 1+\eta(1-\theta)/(\delta\theta) =: K^*$. Since q is an integer, the maximum value of q, denoted as $k_{\max}$, is the result of rounding down K^*: $q \leq k_{\max} = \lfloor K^* \rfloor$. Next, it is necessary to clarify what this conclusion governs and what it does not, which can be discussed from two perspectives. Firstly, the same frequency bound also constrains the cyclic revisit gap: if the cyclic gap is k, the frequency is $1/k$, which must be greater than or equal to $1/K^*$, and k is an integer, so k can only be $k_{\max}$ at most; however, whether a certain gap k can be maintained for a long time depends on the trough before practice, that is, $\rho(k)(1-\delta)^{k-1} \geq \theta$. Using $\theta = 0.7$ with the forgetting rate $\delta = 0.03$ of Figure 1 for calculation, the maximum gap that can be maintained, as certified by the trough criterion, is precisely $k_{\max}$: $\eta = 0.15$ with a gap of 3 (trough approximately 0.7048), $\eta = 0.10$ with a gap of 2 (trough approximately 0.7638), and $\eta = 0.05$ with a gap of 1; if the gaps are set to 4, 3, or 2 respectively, they fail already at the peak. Secondly, the operation of rounding down is only effective for the number of players and strict cyclic gaps: if the schedule is irregular, the *average* gap can be perfectly stuck between $k_{\max}$ and K^* while still

consistently sustaining qualification (for example, when $\eta = 0.05$, the pattern of "practice–practice–rest" can consistently maintain a level $s \geq 0.75 > \theta$, with an average gap of 1.5, which is larger than $k_{\max} = 1$). Therefore, in this proposition, the number of players is limited by $\lfloor K^* \rfloor$, but for individuals, what is limited is the frequency, not the gap for each visit. □

C THE FULL HARDNESS CONSTRUCTION

Next, we prove Theorem 3(b): when the skill growth rate of the QUALIFICATION-SCHEDULING problem varies with different pairings ($\eta_{i,r} \in [0, 1)$) and the forgetting rate $\delta = 0$, as long as the input includes the number of players n, it is NP-hard; even if the time span is only $T = 2$, it can be proven using the PARTITION problem; it is also strongly NP-hard, as can be proven using the 3-PARTITION problem. The construction of this proof is divided into two key parts: (i) designing a "small device" (a gadget) that consumes exactly one item size per "training swap"; (ii) setting an upper limit, the growth cap, on how much any skill can ever grow. How small is this upper limit? It is even smaller than the minimum cost incurred by making a single swap. Therefore, no matter how skills increase, the optimal lineup will never change: the benchmark P^*, whose value is fixed, will not budge even a cent. Moreover, regardless of the schedule or round of operation, the total cost of switching roles multiple times in a single round is equivalent to the sum of the costs of each individual swap, without any discrepancy.

Membership in NP. A schedule of at most T matchings is a certificate of polynomial size, because T is given in unary. To check it, replay the rounds one by one: with $\delta = 0$, a pair played k times has skill $1 - (1 - s_{i,r}(0))(1 - \eta_{i,r})^k$, whose numerator and denominator have at most k times the bit length of the inputs, and the floor test in each round needs one maximum-weight perfect matching, which the Hungarian method finds in polynomial time. Together with the reductions below, this makes the problem NP-complete at $T = 2$ and strongly NP-complete in general.

The instance. Now let's first give a 3-PARTITION instance: there are $3m$ positive integers a_1 to a_{3m}, whose sum is $\sum_j a_j = mB$, and each number is greater than $B/4$ and less than $B/2$, with all numbers being within the polynomial range. Next, let's form a team: there are a total of $n = 6m$ team members, divided into two categories: there are $3m$ individuals referred to as *trainees* (the players who must qualify), denoted as X_1 to X_{3m}; additionally, there are $3m$ individuals referred to as *anchors* (the players who hold the target roles), denoted as Y_1 to Y_{3m}. There are also $6m$ roles, which are also divided into two categories: there are $3m$ *target* roles, denoted as R_1 to R_{3m}; there are $3m$ *parking* roles (i.e., temporary holding roles), denoted as Q_1 to Q_{3m}. Let us first establish two numerical values: the scale $c = 1/(2B)$, the growth cap $\gamma = c/(4m)$. Then, establish the rules one by one:

- $\theta = 1 - \gamma/2$, and the designated pairs are (X_j, R_j) for all j;
- initial skills: $s_{X_j,R_j}(0) = s_{X_j,Q_j}(0) = 1 - \gamma$; $\quad s_{Y_j,R_j}(0) = 1$; $\quad s_{Y_j,Q_j}(0) = 1 - c\,a_j$; every other skill 0;
- rates: $\eta_{X_j,R_j} = 1/2$ for all j; every other rate 0;
- tolerance: ε chosen so that the round budget is exactly $\varepsilon\, P^*(0) = cB$ (see below); horizon $T = m$.

Structure of assignments. Let's name the most basic lineup as the *baseline* matching M_{base}, $\{X_j \to Q_j,\ Y_j \to R_j\}_j$: that is, all trainees should stay in the parking role Q_j, and all anchors should stay in the target role R_j. The total value of this lineup, denoted as V_0 (the lineup value P of Section 2), is equal to $\sum_j \big[(1-\gamma)+1\big] = 3m(2-\gamma)$ (i.e., the sum of each individual's skill: $1-\gamma$ for each trainee, 1 for each anchor, and the sum of $3m$ groups is $3m \times (1-\gamma+1)$). Define another operation called the *swap* τ_j of block j: swap the roles of the two individuals in block j; trainee X_j goes to target role R_j, anchor Y_j goes to parking role Q_j, and the others remain the same as in the baseline lineup M_{base}. Finally, a conclusion: if any player in a lineup is assigned to a role where their skill is 0 (outside their two nonzero-skill roles), the score for that seat will be directly 0. Let's do some more calculations: even if the remaining seats experience an increase in skill, the magnitude of the increase cannot exceed the growth cap. Furthermore, even with all possible calculations, a maximum of $3m$ seats can experience an increase, and the magnitude of each increase is less than γ. Upon calculation, the total value of this cross-assigned lineup is more than $(1-\gamma)-3m\gamma \geq 1-c$

below the baseline V_0. This is more than the budget per round $cB = 1/2$, because for 3-PARTITION instances ($B \geq 3$), $c = 1/(2B) \leq 1/6$; in the PARTITION variant, the loss exceeds $1 - c/2 \geq 3/4$, which is greater than $1/2$. Therefore, such matchings are infeasible from start to finish. All feasible assignments are of the form: the baseline with some set S of simultaneous swaps applied (which sets S are feasible is determined by Lemma C.1's deficit against the budget).

Lemma C.1 (exact deficits, all rounds). We denote the current value of the baseline matching as V_0. For each block j, we denote the amount by which the target skill of X_j has grown above $1 - \gamma$ so far as ℓ_j; growth is capped, so ℓ_j falls within the range $[0, \gamma)$. *In every round of every schedule, applying the swap set S yields value $V_t(S) = V_0 - \sum_{j \in S}(c\,a_j - \ell_j)$. Proof.* Only the skills s_{X_j,R_j} ever change (all other rates are 0, and $\delta = 0$). Meanwhile, the player pairs involved in the swaps are disjoint, so the total value can be added up seat by seat. If block j is not swapped, the contribution is $(1 - \gamma) + 1$; if swapped, the contribution is $s_{X_j,R_j} + (1 - c\,a_j) = (1 - \gamma + \ell_j) + 1 - c\,a_j$. The difference between the two is $c\,a_j - \ell_j$, and the total growth is capped by $\ell_j < 1 - (1 - \gamma) = \gamma$, as the skill can never exceed 1. □

Lemma C.2 (P^* is exactly invariant). *$P^*(t) = V_0$ in every round of every schedule. Proof.* According to Lemma C.1, the value of the baseline matching remains constant at V_0 (with its seats' skills never changing), and the value change brought about by each swap is $\ell_j - c\,a_j \leq \gamma - c \cdot 1 < 0$, because $\gamma = c/(4m) < c \leq c\,a_j$; the situation with cross-assignment is only worse. Therefore, no matching's value can exceed V_0. □

Now set $\varepsilon = cB/V_0$, so that the budget for each round is always exactly cB (according to Lemma C.2).

Qualification arithmetic. For practicing (X_j, R_j): when the initial level is at $1-\gamma$, practicing once can increase it by $\eta_{X_j,R_j}\,(1 - s) = \frac{1}{2}\,\gamma$, landing exactly on the line $1 - \gamma/2 = \theta$; in plain terms, qualifying with just one practice. Conversely, X_j starts $\gamma/2$ below θ, so at least one practice round is necessary.

Completeness (3-partition gives a schedule). Given a 3-partition into triples $S_0, \ldots, S_{m-1}$, with each group consisting of three numbers and the sum of each group being equal to B: in each round t, we apply the swap set S_t. According to Lemma C.1 (where ℓ_j is 0 at first practice), the deficit in this round is $\sum_{j \in S_t} c\,a_j = cB$, which exactly consumes the budget, so it is feasible. After completing m rounds, each X_j has practiced once and all are qualified.

Soundness (schedule gives a 3-partition). Thinking about it in reverse: assuming there is a feasible schedule with a total of at most m rounds, and ultimately all (X_j, R_j) are qualified; then each X_j must have been swapped in at least once in one of the rounds. (If the schedule uses fewer than m rounds, the grouping below yields fewer than m groups, each summing to at most B, which cannot cover the total mB; so exactly m rounds are used.) In a certain round, the swap set is S. The deficit for this round is $\sum_{j \in S}(c\,a_j - \ell_j)$, which is greater than $\sum_{j \in S} c\,a_j - 3m\gamma$ (because there are at most $3m$ swaps in a round, and each ℓ_j is less than γ). Since the schedule is feasible, it follows that $\sum_{j \in S} c\,a_j < cB + 3m\gamma = cB + \frac{3c}{4}$. Dividing both sides by c yields $\sum_{j \in S} a_j < B + \frac{3}{4}$. Since the a_j are all integers, it follows that $\sum_{j \in S} a_j$ can at most be B. In each round, the sum of those items that are practiced for the *first* time will definitely not exceed the full swap-set sum of that round. If we divide these $3m$ "first-practice" events into m groups based on the round, the sum of each group will be at most B, but the total sum of all groups will be mB; which means that the sum of each group must be exactly B. Coupled with the condition that $B/4 < a_j < B/2$, each group will consist of exactly three items: this is nothing but a 3-partition.

Weak hardness (numbers written in binary) at $T = 2$. Applying the same construction to a PARTITION instance (m' items summing to $2B$, two rounds, with a budget of cB per round and a cap $\gamma = c/(4m')$), it can be observed that even with only two rounds, the problem is already NP-hard. The term "strongly NP-hard" referred to here means: even if the problem is restricted to instances where "both the numerator and denominator of all rational parameters are bounded by a polynomial in the input length," it remains NP-hard. The 3-PARTITION reduction satisfies this condition, because if the input of that problem is given in unary, the sizes of m, B, and all a_j are polynomially bounded. These constructions can all be computed in polynomial time. □

D THE SETTING DRAWN, AND THE ROTATION POLICY IN FULL

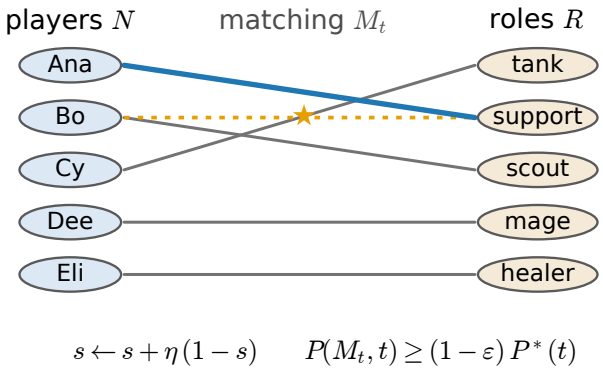


Figure 2: **The setting.** Five players, five roles, one lineup per round: a perfect matching M_t (solid gray edges; Ana's assignment drawn bold blue). Playing a role trains it by the update $s \leftarrow s+\eta(1-s)$ of equation (1); the chosen lineup's value must clear the floor $P(M_t, t) \geq (1-\varepsilon)P^*(t)$; and the season repeats for T rounds. The dotted orange edge, marked by a star, is a desired role its player is not assigned (Bo's support). Such a role stays in the option set exactly while some floor-feasible schedule can still deliver enough practice; that reachability is the paper's central object.

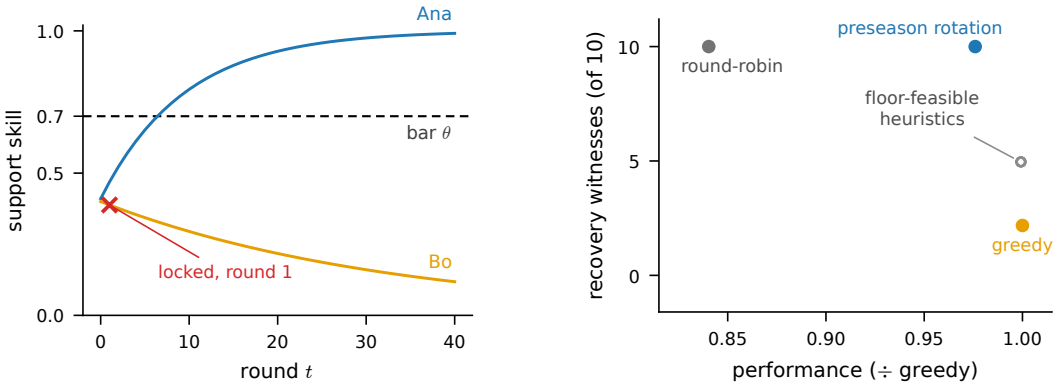


Figure 3: **The introduction's run, continued.** *Left:* the same run as Figure 1, over 40 rounds in skill space (every skill starts at .40 except Ana on support at .41; $\eta = .10$, $\delta = .03$, $\varepsilon = .05$, $\theta = .70$; after round 1 the best of the 119 other lineups is worth 2.165 against the floor 2.19355, and the floor minus that best alternative, $1.75 - 1.0505(.9)^t - .8(.97)^t$, rises every round, so the opening is forced from round 1 on): Ana (blue) passes the bar θ (black dashed) while Bo's unused skill (orange) decays; the red cross marks the round-1 lock. *Right:* the performance relative to greedy at the tested tight/fast condition of Table 1 (50-run means): greedy has 2.2 of 10 desired-pair floor-feasible training schedules at full performance; the four other floor-feasible heuristics coincide at full performance around 5.0, so their open gray markers plot as one; preseason rotation has floor-feasible training schedules for all 10 at 2.4% cost; round-robin, which ignores the floor, pays 16%.

Algorithm 1 states rotate-then-specialize as evaluated in Section 6.

E PROOFS FOR SECTION 5: ASSIGNMENT UNDER UNCERTAIN SKILL

The skills in the entire appendix are latent and change according to the formulas (1) and (2). Before the start of round 0, each of the n^2 pairs can obtain m independent assessments $y_{i,r}^{(j)} = s_{i,r}(0) + \nu_{i,r}^{(j)}$; in season, for each pair selected for play, a reading of the current skill will be measured before skill updating. All noise terms are independent normally distributed with mean 0 and variance σ^2, and the floor constrains true performance.

Algorithm 1 Rotate-then-specialize, with the urgency rule as floor-feasible fallback

Require: initial skills $s(0)$, desired sets $D_1, \dots, D_n$, rates η, δ, tolerance ε, bar θ, rotation cap $t_{\max}$

$A, B \leftarrow$ two fixed edge-disjoint matchings ($A(i) \neq B(i)$ for all i) whose union covers every desired pair
$U_t(M) \leftarrow \sum_{i:\, M(i) \in D_i} \max\{0, \theta - s_{i,M(i)}(t)\}$ {urgency, defined locally}
for $t = 0, 1, \dots, T-1$ **do**
  $\mathcal{F}_t \leftarrow \{M : P(M,t) \geq (1-\varepsilon)\, P^*(t)\}$ {the floor-feasible assignments}
  if $s_{i,r}(t) < \theta$ for some desired pair (i,r) **and** $t < t_{\max}$ **then**
    $M_t \leftarrow A$ if t is even, else B {rotation phase}
    if $M_t \notin \mathcal{F}_t$ **then**
      $M_t \leftarrow \arg\max_{M \in \mathcal{F}_t} U_t(M)$, breaking ties by $P(M,t)$ {floor enforced: fall back}
  else
    $M_t \leftarrow \arg\max_{M \in \mathcal{F}_t} U_t(M)$, breaking ties by $P(M,t)$ {specialization phase}
  play M_t; skills update by equations (1) and (2)

E.1 Proof of Corollary 1

Corollary 1 (Estimates freeze; from the perturbed-start form of Proposition 2, Appendix B.6)**.** *Suppose all initial skills lie within κ of $\bar{s}$, with $\bar{s} + \kappa < \theta$, and*

$$\kappa \leq \frac{\varepsilon \bar{s}}{2-\varepsilon} \qquad \textit{and} \qquad (4+\varepsilon n)\,\kappa \;<\; 2\eta(1-\bar{s}) - \varepsilon n\big(\bar{s} + \eta(1-\bar{s})\big), \tag{4}$$

the second being Proposition 2's perturbed freeze condition (Appendix B.6). Then all assignments are floor-feasible in round 0, but any floor-feasible schedule must repeat its selected opening forever. All other desired roles are permanently lost in round 1.

We first demonstrate that all assignments are feasible in round 0. Regardless of the assignment M, everyone's skills hover within the range of "$\bar{s}$ minus κ" to "$\bar{s}$ plus κ". Therefore, according to M, the total performance $P(M,0)$ is at least n times ($\bar{s}$ minus κ); while the theoretically maximum total performance $P^*(0)$ is at most n times ($\bar{s}$ plus κ). By rearranging the first inequality in formula (4), we can obtain the expression $\bar{s} - \kappa \geq (1-\varepsilon)(\bar{s} + \kappa)$. From this derivation, it is evident that the overall performance based on M can at least reach $(1-\varepsilon)$ times the theoretical maximum overall performance, $P(M,0) \geq (1-\varepsilon)P^*(0)$. In this way, all the assignments that can be arranged in round 0 are counted in the feasible set $\mathcal{F}_0$; especially the assignment $\widehat{M}_0$ that the planner estimates and selects on their own, which is definitely included. The previously mentioned conclusion of "perturbed-start freeze" (see the proof of Proposition 2 in Appendix B.6) can be directly applied here without any alteration: as long as the schedule can maintain the floor, the assignment chosen in round 0 will be used consistently and will not be changed again; for roles not assigned, the corresponding skills will never improve; except for the role $\widehat{M}_0(i)$ already assigned to each individual, other desired roles originally intended for them will be permanently lost at round 1. The "choose one from two" mentioned in the corollary is clear: from round 1 onwards, the only feasible assignment that can also maintain the floor is $\widehat{M}_0$ (this is what the "uniqueness" step in the proof refers to). Therefore, the policy is either to consistently use this assignment, or else the true floor cannot be maintained. The entire reasoning process completely disregards how $\widehat{M}_0$ is calculated, so as long as the policy is observation-measurable (a function of the readings alone), this conclusion holds. □

E.2 Proof of Theorem 2

In this type of instance we provided, there are at least two players ($n \geq 2$), and there is also a role b, which is present in both desired sets ($b \in D_1 \cap D_2$). Initially, everyone's skills are at $\bar{s}$, except for one player ξ (either 1 or 2), unknown to the planner, whose initial skill on role b is $s_{\xi,b}(0) = \bar{s} + \Delta$, where Δ is a positive number with $\bar{s} + \Delta < \theta$, and the noise level σ is also a positive number; all parameters satisfy the previously mentioned formula (4), where κ equals Δ. Before the planner officially starts, each "player + role" pair can obtain m assessments with error, where the reading $y_{i,r}^{(j)} = s_{i,r}(0) + \nu_{i,r}^{(j)}$ is the true initial skill plus random noise. We call the average of these m

readings $\bar{y}_{i,r}$, which follows a normal distribution with mean being the true skill and variance being σ^2/m. Moreover, the readings of different pairs are independent.

Step 1: every floor-feasible policy reduces to a preseason-measurable (fixed by the preseason readings alone) opening rule. According to Corollary 1 (where $\kappa = \Delta$, and all the premises of the corollary are satisfied): all assignments in round 0 are feasible; and any schedule that can maintain the floor will perpetually adhere to the assignment chosen in round 0, without any further changes. Firstly, the initial round-0 assignment: at this stage, the season hasn't started yet, and there are no match data available. Therefore, this step's assignment can only be determined based on the limited data from the preseason assessment; moreover, from round 1 onwards, there is no choice of policy (only the opening set at the beginning can be used). A floor-feasible policy is a measurable rule that relies solely on preseason assessment data to determine the opening matching $\widehat{M}_0$. The player with a higher actual level will consistently fail to obtain the desired role b, precisely when the opening matching $\widehat{M}_0$ did not assign role b to this truly better player (denoted as ξ). Let φ denote "who gets the role b in the opening", where φ can be either 1, 2, or $\perp$ (i.e., someone other than 1 or 2, which only occurs when the total number of players $n > 2$). Then "error" occurs when $\varphi \neq \xi$. If φ is $\perp$, regardless of whether ξ is 1 or 2, it is considered an error.

Step 2: the two-point lower bound (only two possible truths are compared) (Yu, 1997). When ξ is 1 and ξ is 2, the distribution difference in the preseason assessment data is only reflected in the average values of the two pairs: "role b for 1" and "role b for 2"; therefore, the likelihood ratio is a strictly monotone function that only follows a single number D; D is the difference $\bar{y}_{1,b} - \bar{y}_{2,b}$ between "the average reading of role b for 1" and "the average reading of role b for 2". When $\xi = 1$, D follows a normal distribution with a mean of $+\Delta$; when $\xi = 2$, D follows a normal distribution with a mean of $-\Delta$, and the variance for both is $2\sigma^2/m$. Regardless of the judgment rule used, the "maximum value of the error probability in two scenarios" is definitely greater than or equal to the "average value of the error probability in two scenarios", $\max_\xi \Pr_\xi(\varphi \neq \xi) \geq \frac{1}{2}\sum_\xi \Pr_\xi(\varphi \neq \xi)$; and this average value is precisely the Bayes risk of the "two normal distributions with a mean difference of 2Δ and a standard deviation of $\sqrt{2}\sigma/\sqrt{m}$ for a binary choice test" (and $\perp$ is considered incorrect in both hypotheses, which only increases the risk). The optimal approach in this test is to use $D = 0$ as the cutoff: if $D > 0$, it is judged as 1, and if $D < 0$, it is judged as 2; the error probability is $\Phi\big(-\Delta\sqrt{m}/(\sqrt{2}\sigma)\big)$; no rule can be better than this.

Step 3: the natural policy attains it. Let us first clarify this rule: "assign b to $\arg\max_{i\in\{1,2\}} \bar{y}_{i,b}$, complete the matching arbitrarily." This is essentially a "threshold test." Regardless of the hypothetical scenario, the probability of it making an error is $\Phi(-\Delta\sqrt{m}/(\sqrt{2}\sigma))$. According to Corollary 1, all openings in round 0 are feasible, so playing and repeatedly playing in this way is floor-feasible (i.e., it will definitely not violate the floor).

Appendix comparison: greedy on the assessment sums at $n = 2$. If there are only two roles ($n = 2$, roles a and b), there are a total of two matchings. Greedy on the assessment sums directly compares the two total scores: $\bar{y}_{1,a} + \bar{y}_{2,b}$ and $\bar{y}_{1,b} + \bar{y}_{2,a}$. The difference between these two scores follows a normal distribution, with a mean of plus or minus Δ and a variance of $4\sigma^2/m$. Therefore, the probability of assigning b to the incorrect player is exactly $\Phi\big(-\Delta\sqrt{m}/(2\sigma)\big)$; as for a tie? The probability is 0, so there's no need to even consider it. The disparity between it and the natural policy is quite illustrative: the readings in column a are distributed identically under both hypotheses, pure noise for decision-making purposes. However, greedy on the assessment sums insists on using these useless numbers for calculation, which is simply a waste of time.

Sample complexity. When $0 < \rho < 1/2$, rearranging the inequality $\Phi(-\Delta\sqrt{m}/(\sqrt{2}\sigma)) \leq \rho$ yields

$$m \geq m_{\min} := \left\lceil 2(\sigma z_{1-\rho}/\Delta)^2 \right\rceil, \qquad z_{1-\rho} = \Phi^{-1}(1-\rho).$$

The lower bound ensures that this is a necessary condition, while the natural policy further ensures that it is a sufficient condition; this is the classical Gaussian sample-complexity bound (Mannor & Tsitsiklis, 2004; Kaufmann et al., 2016). If $\sigma = 0$ (i.e., no noise), then one noiseless reading of each relevant pair orders them correctly. □

Remark (greedy on the assessment sums at larger n). If n is greater than 2, there is no ready-made formula to directly calculate the errors made by greedy on the assessment sums (after all, finding the maximum value involves selecting among $n!$ estimated sums). However, there are two structural patterns that are stable. In the case of a contested role, the remaining $n - 1$ rivals with skill level $\bar{s}$

are exchangeable (interchangeable in distribution). Therefore, the probability that any specific rival can push the more skilled individual out of the role is $(1 - p_\xi)/(n - 1)$, where p_ξ represents the probability of the more skilled individual obtaining the role. Furthermore, as Δ increases, p_ξ does not decrease (after all, if a certain entry in the estimate table increases, the matchings containing it can only remain in the argmax and not be pushed out). When σ approaches infinity (i.e., the noise is extremely high), p_ξ tends towards $1/n$. Therefore, the denial probability (the better player losing the contested role) gradually increases from 0 (when $\sigma = 0$) toward $1 - 1/n$; this value was measured in the case of $n = 5$ in Section 6.

From the perspective of the team, if each player only undergoes one tryout, and the random fluctuations during the tryout are exactly as large as the true skill gap between the two players, then even under the best rule, approximately every fourth season, the contested role will be mistakenly given to the wrong player. Subsequent practice may actually make this incorrect choice appear quite correct.

E.3 A TWO-PLAYER ALTERNATION CALCULATION

Proposition 9 (Two-player alternation with one hidden advantage)**.** *Consider two players and roles a, b, let $q = 1 - \eta$, $\delta = 0$, and $0 < \Delta \le 1 - \bar{s}$. In each of two hidden instances, exactly one player begins at $\bar{s} + \Delta$ on b and all other skills equal $\bar{s}$. Fix the opening matching independently of the observations. Strict alternation is floor-feasible under both hidden instances whenever*

$$\varepsilon \ge \varepsilon_{\mathrm{alt}}(\Delta) := \frac{2\eta(1 - \bar{s}) + q\Delta}{2\bar{s} + 2\eta(1 - \bar{s}) + q\Delta}.$$

Below this threshold, the instance in which the opening matching is the better one makes the first switch infeasible.

Suppose further that $\sigma > 0$, each relevant b-pair receives m_0 preseason assessments, and a played pair reveals its pre-update skill. After L fixed alternating rounds, let $k_1 = \lceil L/2 \rceil$, $k_2 = \lfloor L/2 \rfloor$, and

$$g(k) = \sum_{j=0}^{k-1} q^{2j} = \frac{1 - q^{2k}}{1 - q^2}.$$

The exact minimax probability of misidentifying which player began ahead is

$$p_L^* = \Phi\left(-\frac{\Delta}{2\sigma} \sqrt{2m_0 + g(k_1) + g(k_2)} \right).$$

If $b \in D_1 \cap D_2$ and $\bar{s} < \theta < 1$, both b-pairs qualify within $2k_\theta$ rounds of feasible alternation.

Proof. The initial stronger matching and weaker matching have values of $H_0 = 2\bar{s} + \Delta$ and $L_0 = 2\bar{s}$, respectively. After the stronger matching is played once, the value changed to $H_1 = 2\bar{s} + 2\eta(1 - \bar{s}) + q\Delta$. The first switch is floor-feasible, precisely when $L_0 \ge (1 - \varepsilon)H_1$. Rearranging this formula gives the threshold written above. After completing j complete cycles, the deficit on the same switch is $q^j[2\eta(1 - \bar{s}) + q\Delta]$; this deficit will become smaller and smaller, while the benchmark value will become larger and larger. If the weaker matching is used initially, the requirements will not be higher, so the first robust switch is the most challenging step.

During the jth play of a pair, the observed mean difference between the two hidden instances is $q^j\Delta$. Therefore, the squared distance between the mean vectors of the two Gaussian distributions is $\Delta^2[2m_0 + g(k_1) + g(k_2)]$, and their covariances are both $\sigma^2 I$; the a-column readings contribute nothing. When conducting a likelihood ratio test between two Gaussian distributions with identical covariances, the probability of error is $\Phi\big(-\Delta\sqrt{2m_0 + g(k_1) + g(k_2)}/(2\sigma)\big)$. This leads to the aforementioned formula. Finally, it should be noted that each b-pair is practiced once every two rounds, and a total of k_θ practices are required. Therefore, both qualify within $2k_\theta$ rounds. □

Practice serves two purposes: firstly, it directly preserves both options; secondly, the geometric sum $g(k)$ shows that the later the observation, the less information it carries about who began ahead. As long as alternation is feasible, there is no need to distinguish who is who in order to preserve the options.

E.4 CERTIFIED POLICIES: THE SAFETY LEMMA

The remaining conclusions all pertain to a "planner": this planner solely operates based on the feasibility that it can certify. For each pair, we set two numbers: $\ell_{i,r} = \mu_{i,r} - w_{i,r}$ (lower bound) and $u_{i,r} = \mu_{i,r} + w_{i,r}$ (upper bound). The experiment employs a stricter coverage convention: the difference between the true value and the center should not exceed $w_{i,r}/2$, $|\mu_{i,r} - s_{i,r}| \le w_{i,r}/2$; this ensures that the true value must fall between the lower and upper bounds, $\ell_{i,r} \le s_{i,r} \le u_{i,r}$, and the maximum difference between the bounds and the true value is $3w_{i,r}/2$. For pairs that have never been observed, the initial radius w_0 is maintained. Let us write down two sums: $L(M) = \sum_i \ell_{i,M(i)}$ (the sum of all pair lower bounds), and $U(M) = \sum_i u_{i,M(i)}$ (the sum of all pair upper bounds). An assignment M is *certified* in round t if it satisfies the following condition:

$$L(M) \ge (1-\varepsilon) \max_{M'} U(M'). \tag{5}$$

Lemma 2 (Certification is truly safe)**.** *If all bounds contain the true skills, all certified assignments are floor-feasible.*

Proof of Lemma 2. If all the upper and lower bounds encompass the true value, then $L(M) \le P(M,t)$ holds, and the maximum upper bound among all assignments satisfies $\max_{M'} U(M') \ge P^*(t)$ (the optimal value). Combining these two inequalities with the certified condition (5), we obtain $P(M,t) \ge (1-\varepsilon)P^*(t)$, that is, the performance of this assignment is at least $(1-\varepsilon)$ times optimal. □

E.5 PROOF OF PROPOSITION 4

We fix the incumbent M_0 for round 0. We demonstrate that as long as the initial radius w_0 is greater than $\varepsilon n\bar{s}$ and εn is less than 2, there will never be another matching that can be certified. Prove the tth round using induction: assume that the planner has been using M_0 up to now. The skill of each incumbent pair is represented by $d(t) = 1 - (1-\bar{s})(1-\eta)^t$, while the skill of other pairs that have never been incumbent pairs is simply $\bar{s}$. Therefore, the optimal total performance is $P^*(t) = nd(t)$.

For any distinct matching M, it will result in $k \ge 2$ players being paired with unplayed pairs, with each such pair's radius remaining at w_0. Under $|\mu - s| \le w/2$, every such term in $L(M)$ is at most its true skill minus $w_0/2$. Therefore, according to Lemma 1, it can be deduced that

$$L(M) \le P(M,t) - w_0 \le P^*(t) - 2(d(t) - \bar{s}) - w_0.$$

Additionally, there exists $\max_{M'} U(M') \ge P^*(t)$. Therefore, to complete the certification, it is necessary to satisfy the requirements

$$\varepsilon n d(t) \ge 2(d(t) - \bar{s}) + w_0.$$

When $d(t) = \bar{s}$, the reverse strict inequality is precisely $w_0 > \varepsilon n\bar{s}$. As $d(t)$ increases, the growth rate on the right side is 2, while the growth rate on the left side is εn (less than 2). Therefore, this reverse inequality always holds. No certified deviation will occur; the fallback plan will repeatedly play M_0, and the induction is thus completed. □

A sharper condition. The proof above discards the information the planner does gain on the incumbent pairs. Keeping it gives a condition that reaches into the noise range of Figure 7(b).

Proposition 10 (Certification starvation with shrinking incumbent radii)**.** *Assume the setting of Proposition 4: exact symmetry at $\bar{s} < \theta$, a round-0 incumbent M_0, $\delta = 0$, $\varepsilon n < 2$, intervals with center μ and radius w satisfying $|\mu - s| \le w/2$, and radius w_0 on every unplayed pair. Suppose in addition that at round t every incumbent pair carries the same radius $w_{\mathrm{inc}}(t) \le w_0$. Write $d(t) = 1 - (1-\bar{s})(1-\eta)^t$ and $c_n = (2-\varepsilon)\, n/2 - 1$. If*

$$w_0 + c_n\, w_{\mathrm{inc}}(t) \;>\; \varepsilon n\bar{s} - (2-\varepsilon n)\big(d(t) - \bar{s}\big) \qquad \textit{for every } t \ge 0, \tag{6}$$

then a planner attempting only certified deviations and otherwise repeating M_0 never certifies a deviation. Proposition 4 is the case in which the incumbent term is dropped ($w_{\mathrm{inc}} \equiv 0$): the right side of (6) *is largest at $t = 0$, where it equals $\varepsilon n\bar{s}$.*

Proof of Proposition 10. Induct on t, assuming M_0 has been played in rounds $0, \dots, t-1$. Then every incumbent pair has skill $d(t)$ and radius $w_{\mathrm{inc}}(t)$, every other pair has skill $\bar{s}$ and radius w_0, and $P^*(t) = n\,d(t)$. Coverage $|\mu - s| \le w/2$ gives $\ell = \mu - w \le s - w/2$ and $u = \mu + w \ge s + w/2$ for every pair.

Let $M \neq M_0$ move $k \ge 2$ players off their incumbent roles. Its k moved players sit on unplayed pairs and its $n-k$ others on incumbent pairs, so by Lemma 1

$$L(M) \;\le\; P(M,t) - \frac{k\,w_0}{2} - \frac{(n-k)\,w_{\mathrm{inc}}(t)}{2} \;=\; n\,d(t) - k\big(d(t)-\bar{s}\big) - \frac{k\,w_0}{2} - \frac{(n-k)\,w_{\mathrm{inc}}(t)}{2}.$$

Also $\max_{M'} U(M') \;\ge\; U(M_0) \;\ge\; n\,d(t) + n\,w_{\mathrm{inc}}(t)/2$. Certification, $L(M) \;\ge\; (1-\varepsilon)\max_{M'} U(M')$, therefore requires

$$\varepsilon n\,d(t) \;\ge\; k\big(d(t)-\bar{s}\big) + \frac{k\,w_0}{2} + \frac{(n-k)\,w_{\mathrm{inc}}(t)}{2} + \frac{(1-\varepsilon)\,n\,w_{\mathrm{inc}}(t)}{2}. \tag{7}$$

The right side of (7) equals $k\big[(d(t)-\bar{s}) + (w_0 - w_{\mathrm{inc}}(t))/2\big] + (2-\varepsilon)\,n\,w_{\mathrm{inc}}(t)/2$, which is nondecreasing in k because $d(t) \ge \bar{s}$ and $w_0 \ge w_{\mathrm{inc}}(t)$; so if (7) fails at $k=2$ it fails for every deviation. At $k=2$ it reads $\varepsilon n\,d(t) \ge 2(d(t)-\bar{s}) + w_0 + c_n\,w_{\mathrm{inc}}(t)$, and since $\varepsilon n\,d(t) - 2(d(t)-\bar{s}) = \varepsilon n\bar{s} - (2-\varepsilon n)(d(t)-\bar{s})$, its failure is exactly (6) at round t. Hence no deviation is certified at round t, the planner repeats M_0, and the induction continues. □

Numbers. Under the Gaussian beliefs of Appendix G (prior variance σ^2/m on every pair, an observation of precision $1/\sigma^2$ on each played pair followed by the propagation $v \leftarrow (1-\eta)^2 v$, radius $z\sqrt{v}$ with $z = 2$), $w_{\mathrm{inc}}(t) = r_t w_0$ with r_t free of σ, so condition (6) reads $\sigma > (\sqrt{m}/z)\max_t [\varepsilon n\bar{s} - (2-\varepsilon n)(d(t)-\bar{s})]/(1 + c_n r_t)$. At the parameters of Figure 7(b) ($n=5$, $\varepsilon = .25$, $\bar{s} = .4$, $\eta = .10$, $m = 1$) the maximum is .164 at $t = 4$, so $\sigma > .082$; Proposition 4 alone needs $\sigma > .25$. In the tight/fast and loose/slow cells of Table 2 the conditions are $\sigma > .010$ and $\sigma > .022$ under exact symmetry, and $\sigma > .040$ and $\sigma > .063$ at the table's perturbation $\kappa = .03$; the tested $\sigma = .05$ is covered in the tight/fast cell, and the loose/slow freeze is observed rather than covered. These radii are rounded; to four decimals the sufficient conditions read $\sigma > .0103$, .0220, .0403, .0628, .0819 and .1277. For starts within κ of $\bar{s}$, own-role skills lie within $\kappa(1-\eta)^t$ of $d(t)$ and off-role skills within κ of $\bar{s}$, and the same induction adds $2\kappa + (2n - 2 - \varepsilon n)\kappa(1-\eta)^t$ to the right side of (6); at $\kappa = .03$ the requirement at the Figure 7(b) parameters becomes $\sigma > .13$, which still covers the plotted points at $\sigma \ge .20$ and the point at $\sigma = .10$ only under exact symmetry.

E.6 Proof of Conditional Certified Rotation

Recall $d_1 = \bar{s} + \eta(1-\bar{s})$ from Proposition 2, and let $k_\theta = \lceil \ln((1-\theta)/(1-\bar{s}))/\ln(1-\eta) \rceil$ be the number of practices required to qualify one pair.

Proposition 11 (Conditional certified rotation)**.** *Assume the two-matching setting of Proposition 3, start with A as the round-0 incumbent, and let $\delta = 0$. Until qualification, suppose every bound endpoint is within q_0 of the true value. Thus $0 \le s_{i,r}(t) - \ell_{i,r}(t) \le q_0$ and $0 \le u_{i,r}(t) - s_{i,r}(t) \le q_0$. If $\varepsilon \ge \varepsilon_{\mathrm{rot}} + 2q_0/d_1$, all subsequent matchings in $B, A, B, A, \dots$ are certified, and all desired pairs qualify within $2k_\theta$ rounds.*

The first round of B here is the most difficult, as uncertainty will drag down its value, while the benchmark will also be raised by q_0 per player. Taking the scenario of "loose floor + slow learning rate" used in the experiments in Section 6 (with parameters $\varepsilon = .10$, $\eta = .05$, $\sigma = .05$) as an example, to obtain two-standard-deviation intervals, each pair needs to undergo $m = 533$ assessments before the season; if $n = 5$, a total of $25 \times 533 = 13{,}325$ pair assessments need to be conducted under the condition of fixed skill levels (the *preseason*).

Proof of Proposition 11. Use the edge-disjoint matchings A and B from Proposition 3, set $x_0 = \bar{s}$, and define x_j as $x_{j-1} + \eta(1 - x_{j-1})$. Before the start of the j-th round of B, all skill values of the A pairs are x_j, all skill values of the B pairs are x_{j-1}, and $P^* = n x_j$. Based on the assumed endpoint error, it can be derived that

$$L(B) \ge n(x_{j-1} - q_0) \quad \text{and} \quad \max_{M'} U(M') \le n(x_j + q_0).$$

Therefore, the condition for the B-round to pass certification is

$$x_{j-1} - q_0 \geq (1-\varepsilon)(x_j + q_0),$$

Alternatively, it is equivalent to

$$\varepsilon x_j \geq (x_j - x_{j-1}) + (2-\varepsilon)q_0.$$

The learning gap $x_j - x_{j-1} = \eta(1 - x_{j-1})$ will decrease as j increases, and x_j is increasing, so the first round of B is the most constrained. At this point, $x_1 = \bar{s} + \eta(1-\bar{s}) = d_1$, and $x_1 - x_0 = \varepsilon_{\text{rot}} x_1$; since $2 - \varepsilon \leq 2$, the condition given by the proposition, $\varepsilon \geq \varepsilon_{\text{rot}} + 2q_0/d_1$, is sufficient. The true lineup gap in the A-round, which is interspersed in the middle, is zero, and only needs to meet the same uncertainty margin. Therefore, the most constrained B-round condition can also simultaneously certify these A-rounds. According to Lemma 2, each certified round indeed satisfies floor feasibility, and the alternating arrangement of edge-disjoint matchings allows each desired pair to be practiced every other round. Each pair can reach θ after k_θ practices, so all desired pairs can qualify within $2k_\theta$ rounds. □

The preseason assessment in a frozen state (not a practice round) will provide m independent assessment data for each pair. Assuming Gaussian beliefs, let z be the selected posterior-standard-deviation multiplier, and denote $w_0 = z\sigma/\sqrt{m}$; the subsequent error radius will not increase further. According to the stricter coverage convention in the paper, $|\mu - s| \leq w/2$, the maximum difference between each endpoint and the true value is $q_0 = 3w_0/2$. Therefore, if $\varepsilon > \varepsilon_{\text{rot}}$, an integer assessment budget that satisfies the requirements is

$$m \;\geq\; \left\lceil \left(\frac{3z\sigma}{d_1(\varepsilon - \varepsilon_{\text{rot}})} \right)^2 \right\rceil. \tag{8}$$

Under the assessment scenario of the loose-floor cell ($\varepsilon = 0.10$, $\eta = 0.05$, $\sigma = 0.05$), with $d_1 = 0.43$ and $z = 2$, the conditional sufficient point is that each pair requires $m = 533$ assessments. Keeping the factor $2-\varepsilon$ in the proof instead of bounding it by 2 lowers the same sufficient budget to $m = 481$. Under the Gaussian beliefs, $|\mu - s| \leq w/2$ with $w = z\sigma/\sqrt{m}$ and $z = 2$ is a one-standard-deviation event for each pair, so all 25 pairs satisfy it at once with probability $(2\Phi(1) - 1)^{25} \approx 7 \times 10^{-5}$; the budget grows with z^2.

F Robustness of Lock-In to the Form of the Floor

Using yesterday's best performance as the benchmark, the threshold obtained is the same; whereas if a fixed absolute margin β is used, the threshold is $\beta = 2(1-\bar{s})$, which represents the maximum loss that can occur during a single training swap. In comparison, an exogenous floor met at the start does not exclude any possibilities, as durable skills only continue to strengthen.

Throughout the entire process, the settings of Theorem 1 are employed: the initial state is symmetric, with all players at skill level $\bar{s}$; greedy play is adopted; $\delta = 0$; the own-role trajectory is $d(t) = 1 - (1-\bar{s})(1-\eta)^t$.

Proposition 12 (Floor variants)**.** *Under greedy play from the symmetric start with $\delta = 0$:*

- *(a)* Lagged floor. *Replace the floor by $P(M_t, t) \geq (1-\varepsilon)\, P^*(t-1)$, with $P^*(-1) := P^*(0)$. Then Theorem 1 holds unchanged: deviations become permanently infeasible after finitely many rounds if $\varepsilon n < 2(1-\bar{s})$, and no desired role is ever permanently lost if $\varepsilon n \geq 2(1-\bar{s})$. The reversible window is weakly longer than under the current-best floor.*

- *(b)* Additive floor. *Replace the floor by $P(M_t, t) \geq P^*(t) - \beta$ for a fixed allowance $\beta > 0$. If $\beta < 2(1-\bar{s})$, a deviation exists exactly in the rounds t with $2(d(t) - \bar{s}) \leq \beta$, and every deviation becomes permanently infeasible from the first round past $t^*_\beta = \ln\big((1-\bar{s} - \beta/2)/(1-\bar{s})\big)/\ln(1-\eta)$; if $\beta \geq 2(1-\bar{s})$, every transposition is feasible at every round, and sustaining a chosen transposition qualifies any designated pair.*

- *(c)* Exogenous floor. *Replace the floor by $P(M_t, t) \geq c$ for a fixed constant c, with $n \geq 2$. If $c \leq n\bar{s}$, every assignment remains feasible after every history, and repeatedly assigning any designated player–role pair qualifies it; if $c > n\bar{s}$, no assignment is feasible at round 0.*

Proof. *(a)* In round 0, we adhere to the convention $P^*(-1) = P^*(0)$. At this point, the two floors are identical, and the values of all assignments are equal to $n\bar{s}$, without distinction of superiority or inferiority. For the case of $t \geq 1$, the lowest-cost deviation method is to swap two players (see Lemma 1), with a corresponding value of $n\,d(t) - 2(d(t) - \bar{s})$; while the lagged floor requirement is $(1-\varepsilon)n\,d(t-1)$. Substituting $d(t) = 1 - u(t)$ and $d(t-1) = 1 - u(t)/(1-\eta)$ (where $u(t) = (1-\bar{s})(1-\eta)^t$), the difference between the two is

$$G(t) = \big[\varepsilon n - 2(1-\bar{s})\big] + u(t)\Big(\tfrac{(1-\varepsilon)n}{1-\eta} - (n-2)\Big).$$

Since $u(t)$ is always positive, strictly decreasing, and eventually approaches 0, G varies monotonically with t, with a limit value of $\varepsilon n - 2(1-\bar{s})$. Therefore, it crosses zero at most once. If $\varepsilon n < 2(1-\bar{s})$, then from a certain round onwards, $G(t)$ will always be less than 0: any deviation will hit the floor, and only the greedy policy can be followed. Moreover, as in Theorem 1(a), the same monotonicity will make all subsequent deviations infeasible, so all non-own desired roles will be permanently lost. If $\varepsilon n \geq 2(1-\bar{s})$, then in each round, the lagged floor is weaker than the current-best floor ($P^*(t-1) \leq P^*(t)$). Therefore, the sustained swap in Theorem 1(b) is still feasible and can qualify any designated pair. Finally, the lagged budget is greater than the current-best budget by $(1-\varepsilon)n\,(d(t) - d(t-1)) \geq 0$, so the crossover point will not occur earlier than in Theorem 1: the reversible window will only be longer (not shorter).

(b) In the additive comparison, the benchmark value cancels: the necessary and sufficient condition for M to be feasible is that its deficit $P^*(t) - P(M,t)$ does not exceed β, and according to Lemma 1, the deficit of the lowest-cost deviation is $2(d(t) - \bar{s})$. If $\beta < 2(1-\bar{s})$, the deficit will strictly increase toward $2(1-\bar{s}) > \beta$, and solving for $2(d(t) - \bar{s}) > \beta$ yields $t > t^*_\beta$; after this point, only the greedy policy can be followed, and the deficit will continue to increase. Therefore, no matter how the process continues thereafter, the infeasible state will not change. If $\beta \geq 2(1-\bar{s})$, the deficit will always remain below β, so a two-player swap is always feasible; under a sustained swap, the two practiced skills will improve, while the contributions of all other players will cancel in the deficit, so the deficit will never increase, exactly as in Theorem 1(b).

(c) When $\delta = 0$, used skills will only improve, while unused skills remain unchanged. Therefore, regardless of the schedule, each skill will not fall below its initial value $\bar{s}$, and after every history, the value of each assignment will be at least $n\bar{s}$. If c does not exceed $n\bar{s}$, then any assignment will always work; one can repeatedly use a perfect matching (i.e., an assignment where every player gets a role) that contains the designated pair to qualify it. However, if c is greater than $n\bar{s}$, then the value of all assignments in round 0 will be exactly $n\bar{s}$, which is definitely less than c. □

These three parts frame the setting of the model: there is a floor that rises with the team, eventually tying lock-in to the limit cost of one training swap, $2(1-\bar{s})$; this limit is reached either as the tolerance εn or as the allowance β; while another exogenous floor, which holds in the symmetric initial state, never restricts the assignment. Let's illustrate with an example: substituting $n = 5$ and $\bar{s} = .40$ into the two closed-form solutions (i.e., formulas that directly yield results) yields the thresholds: $\varepsilon = .24$ and $\beta = 1.20$.

G Experimental Details

Bimodality behind Table 1's means. The tight-floor means do not follow the typical distribution pattern of being high in the middle and low at both ends. Instead, they are bimodal, divided into two distinct groups with significant differences: under the parameters ($\varepsilon = 0.05, \eta = 0.15$), this situation occurred 48 times out of 50 runs: throughout the entire 100 rounds, the least-served player was always able to secure a desired role, and ultimately, there were exactly 5 pairs that are NR_{20}; the remaining 2 runs were completely different: in round 0, there was no floor-feasible lineup that could cover everyone's desired role, and ultimately, 6 pairs are NR_{20}, and the least-served player did not secure a desired role from start to finish.

The boundary sweep at an admissible perturbation. The perturbation magnitude used in our main experiments ($\kappa = 0.03$) is larger than the worst-case radius $\kappa_0 = 0.015$ that can be tolerated at $\eta = 0.10$ as stated in Proposition 5. Therefore, we adjusted κ to 0.015 and retested the boundary sweep (the sweep of the tolerance across the critical value; with 50 new random seeds, the check

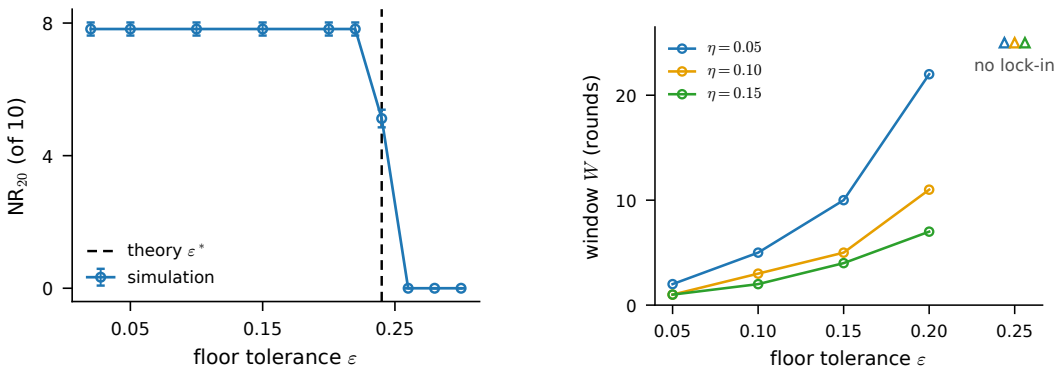


Figure 4: **The threshold value appears at the position predicted by theory.** *Left:* NR_{20} (meaning: the number of desired pairs for which the fixed $H = 20$ catch-up check finds no floor-feasible training schedule) under greedy, varying with tolerance ε ($\delta = 0$, $\eta = 0.10$, $\kappa = .03$). The points on the graph represent the means over 50 runs, the error bars indicate the standard errors, and the dashed line represents the exact-symmetry threshold $\varepsilon^* = 0.24$ of Theorem 1. *Right:* the size of the reversible window varying with ε under three learning rates; the 12 circle points are the closed form of Proposition 1. The triangles at $\varepsilon = 0.25$ are drawn at the top of the axis, indicating that no lock-in occurred within the budget of 120 rounds; these parameter settings are all above the threshold.

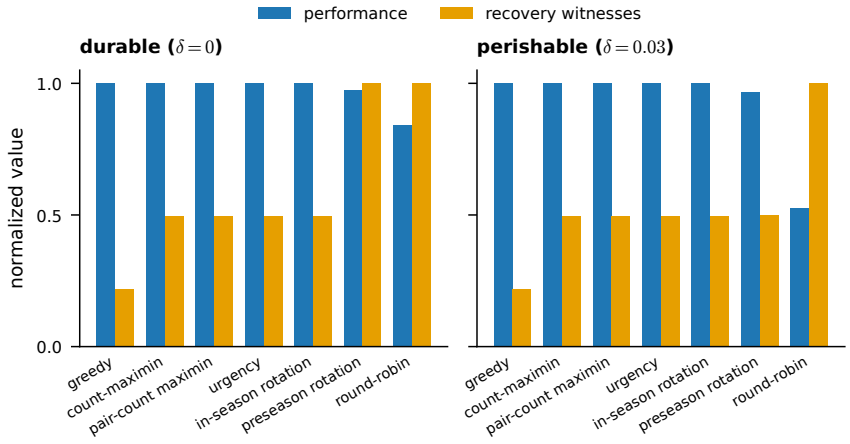


Figure 5: **The performance across policies, relative to greedy.** The performance relative to greedy is the share of greedy's cumulative value a policy keeps; one minus this share is the loss, and the price of optionality of Proposition 6 is the least such loss over schedules that qualify every desired pair. The figure shows this performance for the various policies, as well as the fraction of desired pairs with an $H = 20$ floor-feasible training schedule, when the parameters are set to $\varepsilon = 0.05$ and $\eta = 0.15$: the left side represents the durable regime, and the right side represents the perishable regime ($\delta = .03$); the values in the bar chart are the means over 50 runs. In the durable regime, preseason rotation and round-robin can achieve 97.6% and 84.0% of greedy performance, respectively, while finding floor-feasible training schedules for all pairs; however, in the perishable regime, only round-robin among the tested policies can achieve this.

criteria remain the same as before): under greedy, the mean NR_{20} is 7.94 when $\varepsilon = 0.20$ and 0.22, 4.8 when $\varepsilon = 0.24$, and 0.00 when $\varepsilon = 0.26$ and 0.28. That is to say, the actual transition we measured falls precisely within the band predicted by the proposition, $\varepsilon^* \pm 6\kappa/n = [0.222, 0.258]$; this point has also been mentioned in Section 6. That plateau belongs to the main sweep at $\kappa = .03$; the admissible sweep's plateau is 7.94. The value of the plateau before the transition is 7.82, which is entirely attributed to greedy's opening: since the assignment at the beginning is completely determined by minor initial differences, there is only about a $2/5$ chance for a player to obtain a

desired role. Calculating this, it roughly amounts to $10 - 5 \cdot (2/5) = 8$ pairs left unreached from the beginning.

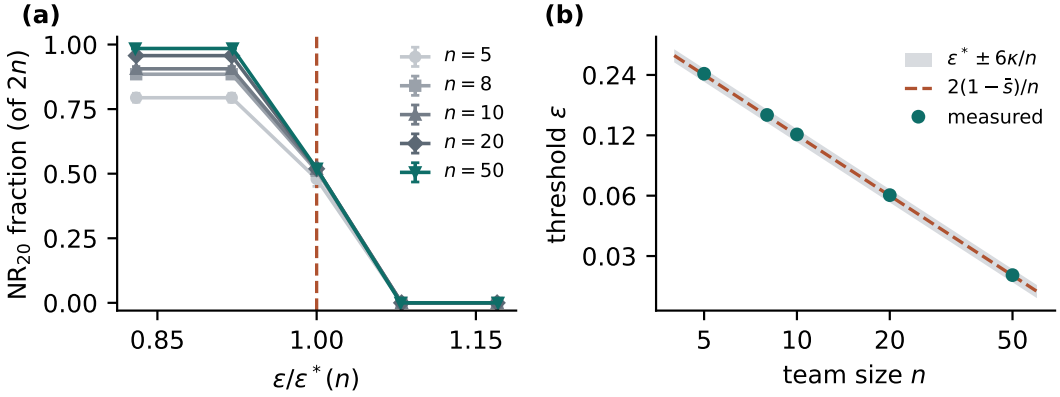


Figure 6: **The lock-in threshold scales as $\varepsilon^*(n) = 2(1 - \bar{s})/n$.** **(a)** Under greedy, NR_{20} as a fraction of the $2n$ desired pairs varies with the relative tolerance $\varepsilon/\varepsilon^*(n)$. The tested values of n are $\{5, 8, 10, 20, 50\}$ ($\eta = .10$, $\delta = 0$, $\kappa = .015$; data are the means with standard errors from $50/50/50/25/15$ random seeds). All curves are in the transition phase at the point where $\varepsilon/\varepsilon^*(n) = 1$ (dashed line), and they complete the decline and reach zero within the multiplier range of 0.92 to 1.08; the plateau height approaches 1 as n increases ($.794$ when $n = 5$ and $.985$ when $n = 50$). The reason is that greedy can match approximately two out of the $2n$ desired pairs at the beginning. **(b)** On logarithmic axes, the measured midpoints of transitions vary with n, accompanied by the prediction line $2(1 - \bar{s})/n$ (dashed line) and the band $\varepsilon^*(n) \pm 6\kappa/n$ given by Proposition 5 (shaded area); all five midpoints fall within this band. All NR_{20} values are one-sided catch-up-check counts.

The threshold at other team sizes. The n-scaling sweep runs greedy for $n \in \{5, 8, 10, 20, 50\}$, setting $D_i = \{i, (i \bmod n)+1\}$, with parameters $\eta = .10$, $\delta = 0$, and $\kappa = .015$. The grid of relative tolerance values is $\varepsilon = \varepsilon^*(n) \times \{0.83, 0.92, 1.00, 1.08, 1.17\}$; $n = 5/8/10/20/50$ corresponds to $50/50/50/25/15$ random seeds, respectively. The midpoint of the transition is extracted according to the predefined rule before the runs: take half of the maximum mean NR_{20} fraction on the grid, find the position where this value is crossed downward for the last time, and calculate the midpoint using linear interpolation. The measured midpoints are respectively $1.014/1.011/1.010/1.006/1.004$ times $\varepsilon^*(n)$, all falling within the interval $\varepsilon^*(n) \pm 6\kappa/n$. The set of data with $n = 5$ fully replicates the boundary sweep results obtained previously with $\kappa = .015$; we also verify the result numerically: when $n = 5$, all 120 permutations are brute-forced to reproduce the best-lineup values, with a maximum deviation of only 9×10^{-16}. For these two grid points below the threshold, regardless of the value of n, the means are exactly the same. Why is that? Because greedy's moves have nothing to do with ε (the tolerance); and in a state of deep lock-in, regardless of which tolerance is used, all the pairs that fail the check are from the same batch. We also measured the window when ε is equal to $0.5\,\varepsilon^*(n)$: starting from the symmetric state, only one edge is .01 higher ($s_{1,1} = .41$), and counting up to the round where the pair $(2, 1)$ along the greedy trajectory fails to pass the one-sided $H = 20$ check for the first time. The result is always 4 rounds regardless of n, which matches the closed form for the symmetric start in Proposition 1. On this ray, ε multiplied by n is a constant, so the window remains unchanged. Moreover, this result is fully consistent when n is greater than 5, which was obtained through the runs, not pre-assumed. All the counts in this sweep are one-sided, and no exact certification has been run regardless of the value of n.

Windows, scaling, and the rotation cells. All 12 finite window values (4 tolerances $\times$ 3 learning rates) match the closed form. If viewed in terms of the relative ratio $\varepsilon/\varepsilon^*(n)$, all curves are mid-transition at 1.00, and the positions where they fall to 0 are all between 0.92 and 1.08 (Figure 6). In the loose/slow cell, the remaining 6 instances in which in-season rotation does not reach all ten desired pairs were due to the perturbed start blocking the first B round, causing the policy to revert to the urgency rule. In the end, 5 pairs failed to meet the criteria; this accounts for the mean of 0.60. In the three cells below $\varepsilon_{\mathrm{rot}}$, out of 50 random seeds, there are 48 to 50 instances where the

rotation gets stuck in the first two rounds; this is consistent with the prediction of Proposition 3 for the symmetric start.

Joint and playable options. The NR_{20} count is pairwise: each desired pair is checked from the final state with its own continuation. Replaying the four durable cells of Table 1 separates three readings. One common schedule qualifies all ten pairs along the policy's own trajectory in all 50 runs for preseason rotation in every cell and for the urgency rule in the loose/slow cell, in 44 runs for in-season rotation and 36 for pair-count maximin in that cell, and in no run for greedy or count-maximin. At the final state, however, a qualified pair is playable (some floor-feasible lineup assigns the player to it) for exactly five of the ten pairs in every one of those runs of the floor-respecting policies other than pair-count maximin in the loose/slow cell: after the rotation ends the policies specialize, the floor-feasible set at round 100 is a single lineup, and that lineup gives each player one of their two desired roles. Apart from those 36 pair-count-maximin runs, only floor-ignoring round-robin ends with all ten pairs playable. Pair-count maximin in the loose/slow cell is the one exception among the seven rule-based policies: in its 36 all-qualified runs every pair stays playable, because its final floor-feasible set is not a single lineup. Count-maximin in the tight/slow cell ends with a two-lineup feasible set in 2 of its 50 runs; apart from those 2 runs, the 36 non-singleton runs of pair-count maximin in the loose/slow cell and the floor-ignoring round-robin, every run of the seven rule-based policies in the four durable cells ends with a single feasible lineup. Under the floor, a rotation of k_θ practices per role keeps qualifications, and a longer finite one keeps playability as well (Proposition 7 in Appendix B.9); the rotations run here stop at k_θ.

The check horizon is not driving the counts. We used the same set of paired random seeds to recalculate the statistics in Table 1 with H values of 10, 20, 40, 80, and 160: the results for greedy, count-maximin, and the urgency rule, respectively, showed no change in the means within the tight/fast and loose/slow durable cells, retaining two decimal places. Despite increasing the horizon by a factor of 8, no additional pair was recovered.

Horizons under forgetting. According to the definition in Definition 2, regardless of the value of δ, $O_i(t, H)$ only increases and does not decrease as H increases: after all, if a certain training schedule can reach the desired role within the horizon $\tau \leq H$, it can still reach it with a larger horizon, and this still holds true mathematically.

The contested variant in detail. In the rules of the contested variant, each player not only desires their own personal role, but also desires the carry role (the contested role of Section 5). Player 1, whose own role is the carry role, desires role 2 as its second role, so there are again ten desired pairs, and the odd rotation rounds pass the carry role to one player at a time. When $\eta = 0.15$ and $\varepsilon = .05$, every tested floor-feasible alternative leaves 5.0–5.2 pairs unreached, of which 3.2–3.7 are carry pairs of the four non-incumbents and 1.4–1.8 are other desired pairs. When $\eta = 0.05$, the urgency rule adds floor-feasible training schedules to 1–2 more pairs, leaving 3.0–3.8 pairs in total. Under the loose floor $\varepsilon = .10$ at $\eta = 0.15$ the urgency rule leaves 3.00 pairs unreached, all of them carry pairs, and the other three alternatives 5.00. What is mentioned here is the situation of recovery within a limited time; conversely, the capacity result limits all five players *holding* the role indefinitely. Among all the policies tested, only the floor-ignoring round-robin can gather enough floor-feasible training schedules ($H = 20$) for all five players. However, under this "forgetting condition" ($\delta = .03$), it incurs a cost increase of 48–65%.

Policies under noise: the full comparison. Table 2 lists the policy comparisons from Section 6.1: it uses the average values of the same batch of 50 random seeds as Table 1, and tests two cells: the tight floor ($\varepsilon = 0.05$, $\eta = 0.15$) and the loose floor ($\varepsilon = 0.10$, $\eta = 0.05$), $\delta = 0$, and observation noise $\sigma = 0.05$. "Viol." (violations) only counts the number of times the true floor is actually violated in the rounds where it is in effect. For example, in preseason rotation, the first ten rounds are exempt from the floor, so these ten rounds are not counted, while other policies must adhere to the floor rule for a full 100 rounds. The oracle is aware of the true skills (i.e., it is the greedy of Table 1), unlike posterior-mean greedy (greedy run on the posterior-mean skill estimates). Thompson draws one Gaussian sample per pair, clipped to $[0, 1]$, and plays greedy on the sample; without noise, Thompson and UCB coincide with greedy. After its ten warm-up rounds the noisy preseason rotation plays the urgency rule over the certified lineups and repeats its incumbent when none is certified, that incumbent being fixed once, at the first constrained round, as the certified lineup with the best urgency key or, when nothing is certified there, the posterior-mean greedy lineup; the certified planner plays the certified lineup that covers the most desired roles and

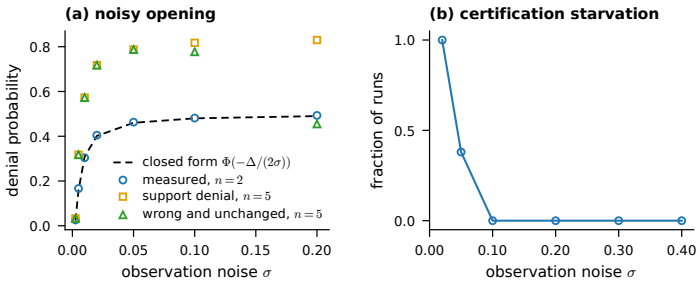


Figure 7: **Noisy opening and certification starvation. (a)** The relationship between the role denial probability and noise ($\Delta = .01$). The black dashed line represents the closed-form expression $\Phi(-\Delta/(2\sigma))$ for greedy on the assessment sums (see Section 5); in the Monte Carlo simulation points, circles represent the case of $n = 2$, and squares represent the case of $n = 5$ (the number of trials is shown in Appendix G). Triangle markers indicate that after an initial mistake, the complete assignment remains unchanged for 100 consecutive rounds, but this does not necessarily mean a permanent loss. **(b)** In 50 perturbed-start runs, the proportion in which the certified planner of Proposition 4 certifies a training deviation at $\varepsilon = .25$; zero is an observed proportion. Both radii of Propositions 4 and 10 are sufficient conditions under its coverage event $|\mu - s| \leq w/2$; the plotted zeros are unconditional observed proportions. Proposition 10 in Appendix E.5 predicts starvation for $\sigma > .082$ at these parameters under exact symmetry and for $\sigma > .13$ at the plotted perturbation ($\kappa = .03$), which covers the zero points at $\sigma \geq .20$.

otherwise repeats its incumbent, the posterior-mean greedy lineup when nothing is certified at round 0 (the source of its 0.04 violations in Table 2, two round-0 openings in 50 runs; no certified round there violates the true floor); the fixed-z heuristic plays the rotation lineup when it is certified, else the best certified lineup, else its incumbent. The radius used for the certified planner and the fixed-z rotation heuristic is twice the posterior standard deviation; the latter is a purely empirical policy tested through trial and error, while the conclusion of Proposition 11 holds only when the premise of simultaneous endpoint coverage is satisfied.

A noisy opening, and starved certification. When two players are competing, greedy on the assessment sums (Section 5) will "snatch away" the more powerful player's role and prevent them from using it (*denial*). The probability of this happening can be calculated using the formula $\Phi(-\Delta/(2\sigma))$. Additionally, Proposition 4 states that if the upper and lower limits are set too wide, the certified planner (someone who is cautious and only dares to make changes after they are certified) will simply not dare to try any training deviation. Figure 7 is specifically designed to test the accuracy of these two claims; during the test, each pair of players is only assessed once before the season, and only the data from actual matches played is considered. First, let's look at panel (a): when there are two players, the actual probability of denial is quite close to the value calculated by the formula, with the error falling within the normal fluctuation range of Monte Carlo simulation (for example, when both σ and Δ are .01, the actual measured value is .304, while the value calculated by the formula is .309, showing a very small difference). As for the denial curve for five players, it can only be used as a rough reference. Why? When there is a lot of noise, this rule, which relies on estimates, sometimes breaches the true floor; after all, it operates based on estimated numbers, not actual numbers. What does the triangular mark in the graph mean? It means that there was a wrong opening, and the assignment remained unchanged for 100 rounds. Only when the freeze conditions described in Corollary 1 are met will this denial turn into a permanent loss where the role is completely unreachable. Let us look at panel (b) again: ε is set to .25, which is just slightly higher than the critical value $\varepsilon^* = .24$. Starting from the symmetric initial state, according to Theorem 1(b), all options should have been open originally; if any options were indeed frozen, it was all due to the cautious planner itself, not the theorem. We conducted 50 runs starting from different perturbed initial states and found that the proportion in which the certified planner dares to certify a training deviation is 1.00 when $\sigma = .02$, but it drops to .38 when $\sigma = .05$, and it drops to zero directly when $\sigma \geq .10$. These perturbed initial states lie outside the exact-symmetry guarantee, so this panel shows what the certified planner's freezing looks like at the plotted noise levels.

If the floor is loose and the learning speed is slow, the ten rounds of warm-up in preseason rotation are simply not sufficient on their own: each desired pair gets to practice 5 times, and after practicing the level only reaches $s \approx 0.54$, which is still below the threshold θ. The rest depends on the

Table 2: **Assignment under noisy observation** ($\delta = 0$, $\sigma = 0.05$, $m = 1$ preseason assessment per pair, 50-run means; standard errors at most 0.24 on NR_{20}). "Perf." is the actual cumulative true performance of each trajectory divided by the total of oracle greedy with the same start and random seed; "Viol." counts only the instances where the true floor is violated in floor-constrained rounds. Therefore, the ten floor-exempt warm-up rounds of preseason rotation are not counted in "Viol."; the other policies have 100 constrained rounds. NR_{20} refers to the one-sided fixed-check nonreach count as stated in Table 1. "Preseason rotation" refers to the floor-exempt preseason rotation of that table; here, the results are obtained from noisy beliefs.

	Tight floor ($\varepsilon = 0.05$, $\eta = 0.15$)			Loose floor ($\varepsilon = 0.10$, $\eta = 0.05$)		
Policy	Perf. (÷ oracle greedy)	Viol.	NR_{20}	Perf. (÷ oracle)	Viol.	NR_{20}
Oracle greedy	1.000	0.0	7.82	1.000	0.0	7.82
Posterior-mean greedy	0.999	0.04	8.08	0.997	0.0	8.08
Thompson	0.998	0.40	7.68	0.988	0.08	7.92
Optimism (UCB)	0.999	0.04	8.08	0.991	0.08	8.00
Certified planner	0.999	0.04	8.08	0.998	0.0	8.08
Fixed-z heuristic	0.999	0.04	8.08	0.998	0.0	8.08
Preseason rotation	0.976	0.0	0.00	0.966	0.0	5.04

urgency rule run on noisy beliefs: at $\sigma = 0$ it reaches every pair ($NR_{20} = 0$, as in Table 1), at $\sigma = .02$ it leaves .16, and at $\sigma = .05$ it leaves 5.04. When the noise level σ is 0.05, the sufficient certification condition mentioned in Proposition 11 does not hold true. In actual measurements, no in-season rotation passed the certification. If the noise level is reduced to $\sigma = 0.02$, the mean NR_{20} for the same preseason policy would only be 0.16. The fixed-z in-season rotation heuristic without warm-up, in all cases with $\sigma > 0$ we tested, had a mean NR_{20} ranging from 7.92 to 8.08. In the loose/slow cell, if $\sigma = 0$, this number is 0.60.

Validation of the option measure. The exact computation of $O_i(t, H)$, although having polynomial complexity when n is fixed (Theorem 3(a)), is still too cumbersome when $n = 5$. Therefore, we use the case of $n = 3$ to validate the catch-up check mentioned in Section 6 from two perspectives. First, under the durable regime: we use the count-matrix dynamic program of Theorem 3(a), directly employing the parameters inherent to the experiments ($H = 20$, $\eta \in \{0.05, 0.15\}$, $\varepsilon \in \{0.05, 0.10\}$, $\delta = 0$) to compute option sets exactly. The results are remarkably accurate: all 360 "player + role" tests were correctly matched, without misjudging reachable situations as unreachable, or recoverable situations as unrecoverable. With forgetting, the dynamic program we just mentioned does not apply. We changed our approach: we shortened the horizon to $H = 6$ (under forgetting, the horizon cannot be too long), set the learning rate to $\eta = 0.30$, δ to 0 or 0.05, and ε to 0.05 or 0.15. Then we searched through all feasible schedules for comparison. The result was still the same: all 360 of 360 tests agreed, with not a single mistake. Therefore, the validation of the durable regime is fully aligned with the horizon and rates of the experiments themselves, while the validation of the forgetting regime is at reduced scale.

Parameters. For the subsequent comparisons of policies, boundaries, and policies with noise, 50 random seeds were used for each condition, with a uniform fixed-check horizon $H = 20$. The rotation cap was set to $t_{\max} = 60$, and there were $2n = 10$ floor-exempt rounds during the preseason warm-up. The catch-up check fields the target pair with its highest-value completion of the other four players whenever that lineup clears the floor, else greedy. Table 1 prints standard errors on NR_{20}, and the standard errors of the corresponding performance are also very small, with none exceeding .008 in each cell. The numbers in Figure 7(a) are derived as follows: when $n = 2$, each noise level is run 4,000 times; when $n = 5$, each noise level is run 400 times. For each exact-check validation mentioned above, each condition was calculated from 10 different starts; the durable comparison used $H = 20$, while the exhaustive comparison with forgetting used $H = 6$. Those single deterministic trajectories, as well as the window curves calculated by closed form, are not obtained by averaging multiple seeds. We use fixed parameter grids for the regular and noise experiments. For those policy learners in Table 1, the initial draws used for each seed are the same as those in the table (the greedy baseline is also recomputed from the same start), except that their exploration is generated using a separate random stream. The exploration rate of the ϵ-greedy learner is set to .10, the step size is fixed at .20 (so that the estimated values can keep up with the actual

values that gradually improve with practice), and the initial values are set to 0. In the event of a tie, a uniform random selection is made. The score-update learner assigns a preference score to each lineup from the 120 permutations, samples a lineup from the softmax distribution restricted to the floor-feasible set, and then adjusts the score of the played lineup by .50 times reward minus a running-mean baseline; this chosen-score update is a REINFORCE-style rule, not the full softmax score-function gradient.

Option-aware learners and the option planner. The option term is computed after each round on the true skills: a desired pair counts when some lineup that clears the floor fields the player in that role, so the term ranges from 0 to 10, against a team value of about 2 to 5. Over $\lambda \in \{.05, .2, 1\}$ the ϵ-greedy learner stays between 8.20 and 8.30 unreached pairs in every cell, since it explores 1.0 to 1.6 lineups whatever the reward. The score-update learner responds only where it explores: in the loose/slow cell NR_{20} falls from 7.36 to 7.18, 6.64 and 4.60 at performance .895, .878 and .837; in the tight/slow cell to 7.82, 7.68 and 7.18; in the loose/fast cell to 7.46, 7.18 and 6.76; in the tight/fast cell it stays at 7.80 at every λ. The option planner scores each candidate lineup by the option counts $|O_i|$ that the catch-up check ($H = 20$) reports from the next state, lexicographically by the smallest count, the total count and the team value; the candidates are greedy plus, for every desired pair below θ, the lineup that fields it with its best floor-feasible completion. On the 50 paired seeds it leaves $3.94 \pm .17$, $5.76 \pm .12$, $4.64 \pm .21$ and $4.38 \pm .17$ pairs unreached in the tight/slow, tight/fast, loose/slow and loose/fast cells at performance .929, .999, .933 and .982 (in the loose/slow cell it pays 6.7% for 4.64 unreached pairs, where the urgency rule pays 7.5% for none), plays 4.4, 1.1, 4.2 and 3.4 distinct lineups, and ends with 6.1, 4.2, 5.3 and 5.3 desired pairs qualified. The planner chooses greedy while all candidates tie on their next-state option counts. It can deviate while every current option remains open when a non-greedy candidate prevents an imminent next-state closure.

Alternative team objectives. Two variants of the objective replace the sum of the fielded skills: the bottleneck $\min_i s_{i,M(i)}$ and the concave $\sum_i \sqrt{s_{i,M(i)}}$; the floor, greedy and the catch-up check all use the variant. The cheapest deviation changes with the objective, and each threshold below is the limit of the deviation's cost-to-budget ratio as the own-role skill d tends to 1. Under the sum it is a transposition costing $2(d - \bar{s})$ against the budget $\varepsilon n d$, hence $\varepsilon^* = 2(1 - \bar{s})/n = .24$. Under the bottleneck every deviation is worth one off-role skill $\bar{s}$ against εd, whatever the number of moved players, hence $\varepsilon^* = 1 - \bar{s} = .60$. Under the square root the transposition costs $2(\sqrt{d} - \sqrt{\bar{s}})$ against $\varepsilon n \sqrt{d}$, hence $\varepsilon^* = 2(1 - \sqrt{\bar{s}})/n = .147$. The greedy boundary sweep ($n = 5$, $\eta = .10$, $\kappa = .03$, 50 seeds) puts the measured midpoints at .610 and .148, with plateaus of 7.56 and 7.82 unreached pairs (the bottleneck greedy opens differently). In the loose/slow cell count-maximin keeps 5.00 pairs unreached at performance .996 and .999 under the two variants, and the urgency rule leaves $.70 \pm .25$ at .932 under the bottleneck and 0 at .959 under the square root.

Exercise-based option set. A stricter reading of Definition 2 counts a pair as open only if a floor-feasible schedule reaches θ and fields the player in that role in the qualifying round. The reported exercise-based values cover only the 12 true-skill policies in Table 1. For a pair below θ the schedule the catch-up check finds does exactly this, since only play raises a skill, so the two readings differ only on pairs already qualified but no longer playable, and the difference equals the number of such pairs in every one of the 28 rule-based policy–cell combinations. The $n = 3$ dynamic program agrees with the corresponding check in 360 of 360 tests under both readings. Under the exercise-based reading only the rows that qualify pairs move in Table 1: preseason rotation goes from 0 to 5.00 in every cell; the urgency rule from 4.40, 0 and 4.86 to 5.00, 5.00 and 5.14 in the tight/slow, loose/slow and loose/fast cells; in-season rotation from 4.80 and .60 to 5.00 in the tight/slow and loose/slow cells. Greedy, the two maximin rules and round-robin do not move, and pair-count maximin in the loose/slow cell keeps 1.40 because its 36 all-qualified runs end with every pair playable. For a pair already at or above θ the exercise-based count asks for playability at the final state, a floor-feasible lineup fielding it now, which is narrower than playability at some reachable state within H. Among the learner and planner rows the two ϵ-greedy learners do not move; the score-update learner moves from 7.80, 7.36 and 7.54 to 7.82, 7.84 and 7.60 in the tight/fast, loose/slow and loose/fast cells, its option-aware version from 7.68, 7.80, 6.64 and 7.18 to 7.74, 7.82, 7.36 and 7.32; and the option planner moves most, from 3.94, 5.76, 4.64 and 4.38 to 6.94, 5.90, 7.62 and 6.18, because the pairs it qualifies are the ones it can no longer field once its lineup specializes. In the scaling test of n, 50 random seeds are used when n takes the values of $\{5, 8, 10\}$, 25 seeds are used when $n = 20$, and 15 seeds are used when $n = 50$.

H Robustness to the Shape of the Learning Curve

We also tried the power-law and logistic curves, both adjusted to pass through θ at the same number of practices. The result was that all three curves exhibited such a steep transition, but the power law shifted the transition point to the left, from .245 to .220.

The learning rule in Equation (1) involves adding a fixed proportion of the remaining gap with each practice session, and all proofs in this paper are tailored to this rule. This appendix only examines, through simulation, how much of the original conclusion remains valid when switching to two other forms of saturating growth.

The two alternative curves, and how they are matched. In the *power law*, the gap shrinks like a power of $1/k$ in the practice count k, as $(1 - s_0)(1 + k)^{-\alpha}$ with $\alpha = .26$, .34 and .42 at $\eta = .05$, .10 and .15; in the *logistic* curve, it first travels horizontally for a period (a flat toe) before starting to rise. The time to threshold of these three families is perfectly aligned: assuming the initial skill value is $\bar{s} = .40$, when they first reach the threshold of $\theta = .70$, the corresponding continuous practice counts $k_c(\eta)$ are exactly the same; for example, when the learning rate η is .05, they all require 13.51 practices; when η is .10, they all require 6.58 practices; and when η is .15, they all require 4.27 practices. Even if rounded up to a whole number, the integer first crossing is also the same for all. The integer first crossing is the ceiling of k_c: 14, 7 and 5 practices. It should be noted here that only the final time to threshold (the endpoint) is aligned, not the intermediate practice path. Taking the group with the slowest learning pace ($\eta = .05$) as an example, after practicing 5 times, the skill value of the exponential is .536, the power law is .623, and the logistic is only .443. Therefore, the differences among the three families we will see later are all caused by different paths, not different endpoints. After matching, there was still one free parameter left in the logistic, and we arbitrarily set a start condition, $s(0) = \bar{s} + .01$ (i.e., .01 more than the initial value), and fixed this pin. Later, we also tested the impact of changing this pin over five values, which is labelled post-hoc (decided after seeing the results) sensitivity analysis. In full, the logistic is $s(k) = s_0 + (1 - s_0)/\big(1 + e^{-g(k-k_c)}\big)$, where $s_0 = \bar{s}$ is the lower level the curve rises from, k_c is the midpoint, and the pin fixes $g = \ln\big((1 - \bar{s})/.01 - 1\big)/k_c$; the power-law exponent is $\alpha = \ln\big((1 - \bar{s})/(1 - \theta)\big)/\ln(1 + k_c)$, and $\delta = 0$ throughout, so a pair's skill depends on its practice count alone.

The transition survives in all three families. Under the committed grid ($n = 5$, $\eta = .10$, $\delta = 0$, $\kappa = .03$, with 50 random seeds), we gradually adjusted ε under greedy and observed that the NR_{20} of the three families all plummeted from a plateau of 7.82 (out of a maximum of 10) to 0, with a particularly rapid decline. The measured midpoints (i.e., the ε values at the half-plateau crossing) are as follows: .2447 for the exponential, which is basically consistent with the predicted $\varepsilon^* = .24$; .2200 for the power law; and .2373 for the logistic. Only the power law shows a more pronounced shift, and it shifts to the left, indicating that its lock-in is weaker: with just a slight degree of tolerance, it can jump out of the plateau phase. The slight shift in the logistic cannot be regarded as an effect of the flat toe, as it can be explained solely by the arbitrarily set start offset (calculation: $2(1 - .41)/5 = .236$, which is almost exactly its midpoint).

A post-hoc account of the power-law shift. Under greedy, when ε is very small, the same lineup is used throughout the T rounds. To deviate in the most cost-effective way, it costs $2(S(T) - s(0))$, and the corresponding optimal benefit is $nS(T)$. Thus, the critical ε value $\varepsilon^*_{\mathrm{gap}}$ is $2(S(T) - s(0))/(nS(T))$. The numbers calculated using this formula at $\eta = .10$ are .2400, .2174, and .2360, while the actual measurements are .2447, .2200, and .2373; the difference is less than .005 each time, which is very accurate. If it is applied to the case of the exponential, this formula can be directly simplified to $2(1 - \bar{s})/n$. Exactly, the exponential value is $2(1 - \bar{s})(1 - (1 - \eta)^T)/(n(1 - (1 - \bar{s})(1 - \eta)^T))$, which is .239997 at $T = 100$ and tends to $2(1 - \bar{s})/n$. Why are the results of the power law different? Because the trained pairs can only reach $S(100) = .876$ at most, and cannot reach 1. As a result, the gap between "trained" and "untrained" is narrowed, and lock-in (that is, always using the same lineup without changing) relies on this gap to function. When the gap is narrowed, it naturally becomes different.

Where the policy ordering breaks. There is a conclusion that is very stable and unwavering: across the three families tested or the two cells tested, count-maximin can ultimately retain at least 4 NR_{20} while maintaining a performance of at least .99 compared to greedy. As for floor-exempt preseason

rotation: in the exponential and power-law families, it can retain all 10 pairs without eliminating any; however, if switched to the logistic family and encountering the tight/slow cell, $4.40 \pm .23$ NR_{20} will remain in the end. Why? Because in 10 rounds of floor-exempt preseason warm-ups, each desired pair can be practiced 5 times. In the initial stage of the logistic (the toe), practicing 5 times can only raise the s-value to .443, while using the exponential can reach .536. To put it simply, whether a fixed-length preseason is effective depends on the shape of the ability growth curve. Afterwards, we tested the sensitivity using five parameter points (pins .005, .01, .02, .05, .10 above $\bar{s}$) from the logistic (specifically fixing the time to threshold to ensure that all versions crossed the θ line at the same integer practice): when the s value after practicing five times, $s(5)$, increased from .428, .443, .464, .509 to .560, the corresponding NR_{20} values were 5.00, 4.40, 1.40, 0.00, and 0.00, respectively. Therefore, this failure is gradual, rather than a sudden and catastrophic failure akin to "clicking and then suddenly not working"; however, the results of selecting that reference pin indeed depend on the specific situation of that pin itself. The urgency rule cannot salvage the situation in the loose/slow cell of the power-law family: the performance can still reach .989, but NR_{20} is $4.96 \pm .03$; if switched to the exponential family, this cell directly drops to 0.00. Throughout the entire process, $\delta = 0$ and $n = 5$. Fifty sets of random seeds were used, and only the two cells marked for testing, along with the four specified policies, were evaluated.

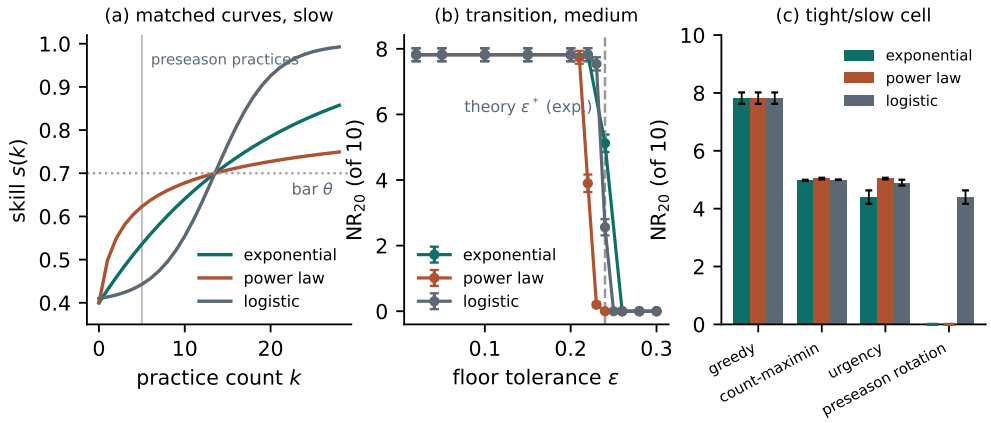


Figure 8: **The learning curve has changed, but the threshold remains robust.** NR_{20} under greedy, against tolerance ε, covers three curves: the exponential, power-law, and logistic families; all three curves are adjusted to pass through $\theta = .70$ at the same number of practices ($n = 5$, $\eta = .10$, $\delta = 0$, $\kappa = .03$; means over 50 seeds). Panel (a) shows the three matched curves at the slow rate $\eta = .05$ with the bar θ and the five preseason practices marked; panel (b) is the sweep described here at $\eta = .10$; panel (c) is the tight/slow cell ($\varepsilon = .05$, $\eta = .05$) policy by policy. All three curves plummet from the same plateau of 7.82 (out of 10), and only the power-law curve shifts the descent location to the left, from .245 to .220. All values are counts from one-sided catch-up checks, so a nonzero value only indicates that no training schedule was found during the check, not that recovery is completely impossible.

The half-gap horizon. The formula for the half-gap horizon is $h = \ln(.5)/\ln(1-\eta)$. What does this number mean? It represents how many practice sessions are needed to reduce the remaining gap to half. Table 3 lists the values of h corresponding to the several η values used in the paper, as well as k_θ, which represents the number of practice sessions required to raise the skill of one player–role pair from $\bar{s} = .40$ to $\theta = .70$. Coincidentally, under the initial and target skill levels selected in our paper, these two values coincide when rounded up, as $(1-\theta)/(1-\bar{s})$ equals exactly $1/2$. Moreover, the lock-in threshold has nothing to do with η: the threshold ε^* in Theorem 1 is $2(1-\bar{s})/n$, and η is not mentioned in the formula at all; furthermore, the window W in Proposition 1 only varies monotonically with η (note: if η changes, W will only continue to rise or fall, without fluctuating up and down, but lock-in should still exist).

Parameters for the curve-robustness runs. The curve-matching parameters and evaluation criteria are recorded before the sweeps. The reported results combine the boundary, policy-cell, gap-formula, and sensitivity computations. The settings used are as follows: $n = 5$, $T = 100$, $H = 20$, $\theta = .70$, $\delta = 0$, $\kappa = .03$, conduct 50 repeated experiments with random seeds. The value range of

Table 3: **Learning rates read as sessions.** $h = \ln(.5)/\ln(1-\eta)$ means that after practicing h times, the gap towards complete mastery will be reduced by half; $k_\theta = \lceil \ln((1-\theta)/(1-\bar{s}))/\ln(1-\eta) \rceil$ represents the number of practice sessions required to reach the target level $\theta = .70$ from the initial level $\bar{s} = .40$. Neither of these two numbers is calibrated using real data; instead, η is simply restated in the unit of practice sessions that a coach can intuitively judge.

	$\eta = .05$	$\eta = .10$	$\eta = .15$
Half-gap horizon h (sessions)	13.51	6.58	4.27
Sessions to $\theta = .70$, k_θ	14	7	5

the committed boundary is $\varepsilon \in \{.02, .05, .10, .15, .20, .22, .24, .26, .28, .30\}$; first, scan all values of ε at the level of $\eta = .10$. If the measured transition interval width is greater than .02, then refine the ε spacing to .01 and perform a more detailed measurement. Two fixed policy cells were tested: $(\varepsilon, \eta) = (.05, .05)$ and $(.10, .05)$. The policies employed include: greedy, count-maximin, urgency, and preseason rotation. The operational value e_{lo} is the largest grid value with mean strictly above half the maximum mean. The operational value e_{hi} is the first grid value after e_{lo} with mean exactly zero. The tested width is $e_{\text{hi}} - e_{\text{lo}}$. A transition is considered to have occurred when both of the following conditions are met: the plateau height is at least 5; the width of the interval from the plateau to zero does not exceed .06. The specific location of the transition is calculated by interpolating the point corresponding to half the height of the plateau. The gap-formula account and the logistic start-offset sensitivity are annotations added afterwards (post-hoc), marked both in the data file and in the preceding text.