Fairness Without λ: Constrained Linear Contracting and Two-Timescale Convergence in Principal–Agent Markov Games

Liz Lemma Future Detective

January 16, 2026

Abstract

Fairness-aware contract design in repeated principal–agent environments is typically implemented by adding an altruism weight λ to the principal's objective, but such weighted-sum regularization is fragile: small changes in λ can flip outcomes from exploitative to overly altruistic, and learning dynamics can destabilize participation. Building on recent principal–agent reinforcement learning and the empirical finding that fairness regularization can improve both equity and welfare, we propose a governance-aligned alternative: the principal maximizes profit subject to explicit fairness and participation constraints (e.g., $1 - \text{Gini} \geq \tau$, Rawlsian minimum wealth $\geq \rho$, and minimum acceptance). We formulate the resulting problem as a constrained Markov Stackelberg program over stationary contract policies and develop a two-timescale primal–dual actor-critic algorithm: agents learn best responses on a fast timescale while the principal and dual variables update slowly. Under standard ergodicity and smoothness assumptions (and smooth surrogates for non-differentiable fairness metrics), we show convergence to a neighborhood of KKT-optimal contract policies, eliminating λ-sensitivity while retaining interpretability via homogeneous linear contracts. Experiments in sequential social dilemmas (Coin Game) and additional resource-allocation environments validate that constrained contracting reliably hits fairness targets with high welfare and avoids unstable behavior observed under welfare-weight regularization.

Table of Contents

agent learning; positioning relative to the source paper's variance/welfare regularization.

3. 3. Model: principal–agent Markov game with hidden heterogeneous types, accept/reject, homogeneous linear contracts; definition of long-run wealth and fairness metrics; discussion of differentiable surrogates.

4. 4. Constrained contracting problem: profit maximization subject to fairness/participation constraints; existence of KKT points; contrast with weighted-sum regularization (why λ is not a policy knob).

5. 5. Algorithm: two-timescale primal–dual actor-critic (agents fast, principal+dual slow); practical implementation details (projection, smoothing, critic approximation).

6. 6. Main theory: convergence to KKT neighborhood under ergodicity + Lipschitz best response; bounds on constraint violation as a function of approximation error; stability discussion vs. greedy principal dynamics.

7. 7. Closed-form sanity check (stylized one-shot model): explicit mapping from fairness thresholds to optimal linear share $\alpha^*(\tau)$ under a tractable effort model; illustrates monotone comparative statics.

8. 8. Experiments: Coin Game reproduction and extensions; sensitivity analysis (thresholds, heterogeneity, costs); comparisons to Greedy, Fixed, Welfare Regularization, and Variance Regularization; ablations on smoothing and timescales.

9. 9. Discussion and limitations: non-convexity, multiple equilibria, fairness metric choice under meaningful heterogeneity; auditability and implementation constraints for 2026 deployments.

10. 10. Conclusion: constrained contracting as the governance-ready alternative to λ-regularization; open problems.

1 Introduction

We study a recurrent contracting problem in which a single principal repeatedly posts a simple, homogeneous linear share contract and a population of agents decides whether and how to act. The economic tension is familiar—surplus creation requires incentives, while the distribution of that surplus matters for legitimacy, retention, and regulation—but the computational and governance environment has changed. By 2026, contracts are increasingly mediated by automated platforms (e.g., marketplaces for digital labor, data labeling, micro-logistics, and API-based services), and the resulting "terms of trade" are often implemented by learning systems that update continuously in response to observed performance. In these settings, it is rarely enough to say that the principal has a generic preference for equity with some weight λ. What platform operators, regulators, and internal governance committees typically require are auditable targets: a minimum participation rate, a minimum earnings floor for the worst-off group, or a bound on inequality as measured by a standardized statistic. These targets are naturally expressed as constraints.

This observation motivates the central modeling choice of the paper: rather than optimizing a weighted sum of profit and a fairness penalty,

$$\max_{\phi}\; w_p(\phi) + \lambda\, F(\mathbf{w}(\phi)),$$

we treat fairness and participation as policy-relevant requirements,

$$\max_{\phi}\; w_p(\phi) \quad \text{s.t.} \quad 1 - G(\mathbf{w}(\phi)) \geq \tau,\;\; \min_i w_i(\phi) \geq \rho,\;\; \mathrm{Acc}(\phi) \geq \kappa,$$

where $\mathbf{w}(\phi)$ denotes the vector of long-run agent wealths induced by the principal policy and the agents' equilibrium responses. The distinction is not semantic. A fixed λ does not generally encode an actionable commitment like "the bottom-decile earnings must exceed ρ" or "the inequality index must stay below a regulatory threshold." Worse, as we emphasize, the mapping from λ to realized fairness can be flat, discontinuous, or environment-specific: varying λ may change the solution very little over wide ranges and then trigger abrupt regime shifts once a participation threshold is crossed. Such behavior is especially acute when agents have an explicit reject option, because acceptance can change non-smoothly as incentives pass an individual rationality margin. In contrast, a constrained formulation is designed to hit a target whenever the target is feasible, and to reveal infeasibility when it is not.

The second motivation is dynamical. In repeated interaction, the principal does not face a static supply curve; agents adapt. When the principal updates contract terms online, and agents update their policies (or effort choices) in response, we obtain a coupled learning system. In this regime, a fairness weight λ becomes a brittle "knob": it must be tuned against non-stationary responses, and a λ that appears to yield acceptable outcomes under one mix of agent types or one state visitation distribution may fail under another. By contrast, primal–dual constrained optimization provides a principled mechanism for automatic calibration: dual variables adjust endogenously to the tightness of each requirement, and therefore act as shadow prices of fairness and participation. This is precisely the language used in governance: we want to know not only whether a target is met, but also the opportunity cost of meeting it.

A third motivation is interpretability and accountability. In many applied contexts, the contract parameter itself—here, the share α or a state-dependent distribution over shares—must be explainable. A constrained approach yields interpretable comparative statics: tightening a fairness threshold τ or a wealth floor ρ generally pushes the policy toward larger shares for agents, and the associated dual multipliers quantify how much principal profit is forgone at the margin. These objects are easier to communicate than a chosen value of λ, which typically has no direct operational meaning and is difficult to justify ex ante.

Our model is intentionally austere: we focus on a homogeneous linear share contract because it is ubiquitous in practice (revenue shares, commission rates, platform take rates) and because it cleanly isolates the distribution–incentive tradeoff. Agents have heterogeneous, unobserved types that scale their effective contributions, and they may reject the contract by choosing an outside option that yields zero activity. The environment is a Markov game, so incentives and fairness are evaluated over a long-run trajectory rather than in a single period. This permits a precise discussion of dynamic participation (will agents continue to engage?), dynamic inequality (do earnings diverge over time?), and the effect of state-dependent contracting.

Against this background, we make four contributions.

(i) A constrained dynamic contracting formulation aligned with policy targets. We formalize fairness and participation requirements as constraints on long-run wealth outcomes. To accommodate gradient-based learning while preserving the economic content of inequality and minimum-wealth criteria, we replace the non-differentiable components (e.g., the Gini coefficient and the pointwise minimum) with smooth surrogates. This creates a differentiable constrained objective that can be optimized with standard tools while remaining interpretable as an approximation to the original governance goals.

(ii) Existence of stationary constrained solutions and KKT structure. Under standard compactness and continuity conditions, we establish the existence of stationary feasible solutions, and under a Slater-type condition we obtain Karush–Kuhn–Tucker points for the smoothed problem. The point is not merely technical: the KKT multipliers provide a disciplined notion of the "price" of fairness and participation in a dynamic environment, clarifying when these requirements bind and how they interact.

(iii) A two-timescale primal–dual learning algorithm with convergence guarantees. We analyze a two-timescale actor–critic scheme in which agents learn (approximately) best responses on a fast timescale while the principal updates its contract policy and the dual variables on a slow timescale. The resulting limiting dynamics track a primal–dual gradient flow. Under unbiasedness and bounded-variance assumptions on the gradient estimators, we show almost sure convergence to an invariant set contained in an $O(\epsilon)$-neighborhood of the KKT set, where ϵ captures function approximation and critic error. Economically, this provides a guarantee that, in the long run, the learning system implements contracts that are nearly optimal for the principal among those that satisfy the fairness and participation targets up to approximation error.

(iv) Why fixed-λ regularization is not a substitute for constraints. We provide a conceptual and constructive argument that weighted-sum regularization with a fixed altruism parameter λ need not recover intermediate fairness targets and can exhibit discontinuities in realized participation and inequality. This matters because common practice in multi-objective learning is to sweep λ and select a point on a Pareto frontier. Our analysis shows that when participation constraints and reject options are present, sweeping λ may skip policy-relevant thresholds, whereas the constrained formulation is designed to meet them whenever feasible.

We also present comparative statics that connect model parameters to contractual generosity and constraint tightness, and we include a one-shot quadratic-effort illustration that yields closed-form mappings from fairness floors to the optimal share. These pieces serve as economic "sanity checks": they clarify what the algorithm should do in benchmark cases and how the shadow prices should move as we tighten governance requirements.

Finally, we emphasize limitations. Our focus on homogeneous linear shares abstracts away from richer contract forms (bonuses, individualized terms, history dependence) that may achieve better efficiency–equity tradeoffs. Our convergence statements are asymptotic and hinge on the quality of value-function approximation; in high-dimensional problems, the residual $O(\epsilon)$ may be material. And while we treat fairness metrics as constraints on realized wealth, the normative choice of metric and the appropriate population over which it is evaluated remain context-dependent. Nonetheless, the model illuminates a core tradeoff faced by modern automated contracting systems: learning can optimize profit, but governance requires targets, and targets are most naturally enforced through constrained, primal–dual dynamics rather than through ad hoc tuning of a fairness weight.

2 Related Work

Our setup draws on three literatures that have largely developed in parallel: (i) contract theory and its canonical constraint sets (limited liability, participation, and incentive compatibility), (ii) reinforcement-learning formulations of principal–agent interaction and incentive design, and (iii) constrained and "fair" learning methods that treat distributional requirements as first-class objects. Our contribution is to connect these strands in a dynamic Markov environment where the principal updates a simple contract online, agents adapt strategically, and the objectives of interest are explicit long-run fairness and participation targets rather than a soft welfare penalty.

Contract theory: LL/IR/IC constraints and dynamic agency. In classical principal–agent models, the basic design problem is expressed as profit maximization subject to constraints that encode feasibility and strategic behavior: limited liability (LL), individual rationality (IR, or participation), and incentive compatibility (IC) **???**. These constraints are not merely technical; they reflect institutional features (e.g., transfers cannot be negative), outside options (agents can walk away), and moral hazard/adverse selection (actions and types are private information). Our environment retains this logic but implements it through a Markov game with an explicit reject action and a restricted contract class. The reject action is a direct operationalization of participation: rather than writing an inequality constraint in the planner's program, we allow agents to choose "no trade" endogenously. Meanwhile, restricting contracts to homogeneous linear shares can be viewed as a reduced-form version of limited liability and simplicity/implementability considerations: many platforms and organizations impose a small menu of auditable take rates, commissions, or revenue shares, even when richer mechanisms are theoretically available.

Dynamic contract theory studies repeated or continuous-time interactions under private information, limited commitment, and persistent heterogeneity, often yielding complex history-dependent optimal contracts **??**. In that tradition, state dependence and continuation values play central roles, and the set of constraints expands to include dynamic IC and promise-keeping conditions. We deliberately step back from full optimal contracting in that sense: the aim of our model is not to characterize the unrestricted optimal mechanism, but to analyze a practically common contract class in a setting where policy targets (e.g., inequality limits, earnings floors, participation mandates) are imposed by governance. In this respect, our "fairness constraints" are of a different nature than IC/IR: they are distributional requirements imposed on the long-run outcome vector rather than feasibility constraints implied by private information alone. They resemble, in spirit, regulatory constraints on outcomes (minimum earnings standards, inequality caps) that are now increasingly discussed for platform-mediated work.

Principal–agent reinforcement learning and computational contract design. A growing body of work models incentive design as a learning problem, where a principal (or designer) adapts payments or reward shaping rules to induce desirable behavior by self-interested agents **???**. This literature includes both mechanism-design-flavored approaches (learning a contract or transfer rule) and multi-agent RL approaches (learning incentives that reshape the game). Related strands study Stackelberg games with learning followers, bilevel optimization, and meta-learning of rewards. A recurring challenge is the endogenous non-stationarity created by adaptive agents: as the principal changes incentives, agents update policies, which changes state visitation and observed performance. Our analysis lives squarely in this regime, but we emphasize two points that are sometimes underdeveloped in purely algorithmic treatments. First, we insist on policy-relevant constraints—targets that can be audited—rather than a soft preference parameter. Second, we explicitly exploit the economic interpretation of primal–dual methods: the dual variables act as shadow prices of constraints, providing a quantitative measure of the marginal cost of compliance.

There is also a closely related line on contract design in Markov decision processes, sometimes phrased as "dynamic mechanism design" with learning, where the principal selects payment rules contingent on observed outcomes to influence actions **?**. Our setting is more modest in contract space but richer in learning dynamics: we allow the principal to use policy-gradient updates over a state-dependent distribution of shares, while agents best respond (approximately) on a faster timescale. This two-timescale viewpoint allows us to connect algorithmic learning to stationary equilibrium objects (KKT points) in a way that supports comparative statics and governance interpretation.

Constrained reinforcement learning and primal–dual actor–critic methods. The algorithmic core of our approach is closest to constrained Markov decision processes (CMDPs) and their Lagrangian solution methods **?**. Modern constrained RL has developed practical actor–critic algorithms and analyses that justify primal–dual updates under stochastic approximation assumptions **???**. Two-timescale stochastic approximation, in particular, provides a standard route to convergence claims for actor–critic and primal–dual schemes **??**. We build on this toolbox but apply it in a different equilibrium setting: the constraints are functions of the wealth vector induced by strategic agents, so the principal's optimization problem is effectively a constrained bilevel problem. The fast-timescale dynamics correspond to follower (agent) adaptation toward a regularized best response, while the slow timescale corresponds to the leader (principal) updating a constrained objective with endogenous multipliers. This is conceptually analogous to constrained learning in games, but the economic content of the constraints (inequality and minimum-wealth floors) is specific to contracting and governance.

A technical distinction is that fairness metrics such as the Gini coefficient or the minimum operator are typically non-smooth, whereas most convergence analyses assume differentiability. Our use of smooth surrogates mirrors a common practice in constrained RL (and more broadly in differentiable programming): one replaces hard non-smooth constraints with differentiable approximations so that policy-gradient estimators and primal–dual updates remain well-defined. The point is not to weaken the governance goal, but to obtain a tractable learning and analysis pipeline that approximates the intended constraint as the smoothing parameter is tightened.

Fairness and inequality objectives in sequential and multi-agent learning. Fairness in learning has been studied under many definitions: demographic parity and equalized odds in classification, individual fairness, and group-based constraints in sequential decision-making and bandits **??**. In RL, fairness constraints have been imposed on visitation, risk, or return distributions across groups, often to ensure equitable treatment over trajectories rather than one-step decisions **??**. In multi-agent settings, researchers have explored equitable equilibria, bargaining-based solutions, and welfare aggregation rules that trade off efficiency and equality (including max–min objectives and inequality indices) **??**. Our fairness criteria are outcome-based and economic: we constrain long-run wealth dispersion and guarantee a minimum wealth floor, which aligns with how platforms and regulators often speak (earnings floors, inequality caps). Moreover, because agents have heterogeneous hidden types and can reject, inequality is not merely an artifact of stochasticity; it reflects both incentives and selection into participation. This makes the fairness constraint interact tightly with the incentive constraint—an interaction that is sometimes abstracted away when fairness is imposed directly on actions or immediate rewards.

Positioning relative to welfare or variance regularization. Our work is also motivated by (and contrasts with) a common practice in multi-objective learning and incentive design: replacing explicit constraints with a weighted-sum objective that includes a welfare term (e.g., utilitarian social welfare) or a dispersion penalty (e.g., variance of agent returns). Such regularizers are appealing because they preserve unconstrained optimization structure and can be tuned to trace a Pareto frontier. However, in contracting environments with reject options and non-convex response mappings, a fixed weight is often not an operational control: wide ranges of the weight can leave outcomes nearly unchanged, while small additional changes can trigger abrupt participation shifts once an IR margin is crossed. In other words, the mapping from a regularization weight to realized fairness and acceptance can be flat or discontinuous, and it is typically environment- and population-dependent. By treating fairness and participation as constraints and updating dual variables endogenously, we obtain a mechanism that is explicitly designed to meet auditable targets when feasible and to quantify the opportunity cost of meeting them. This difference is not merely philosophical; it changes what can be guaranteed and what can be communicated to stakeholders.

Taken together, these literatures suggest both the opportunity and the gap: we have strong tools for dynamic agency and strong tools for constrained learning, but we need models that make governance-style distributional targets explicit in a learning principal–agent system. Our framework is intended to fill that gap while remaining close to the contract forms and accountability requirements that motivate the problem in practice.

3 Model

We model contracting as an infinite-horizon discounted interaction between one principal and n strategic agents in a Markov environment. The state space $\mathcal{S}$ is finite, and time is indexed by $t = 0, 1, 2, \ldots$. The key economic ingredients are (i) persistent heterogeneity in agents' productivities, which is hidden from the principal, (ii) an explicit participation option implemented as a *reject* action, and (iii) a restricted but practically common contract class: a homogeneous linear revenue share $\alpha \in [0, 1]$ that the principal can adjust over time.

Hidden types and contributions. Each agent $i \in \{1, \ldots, n\}$ has a fixed type parameter $\theta_i > 0$ that scales how effectively their behavior translates into observable output. The principal does not observe θ_i, and we do not assume that θ_i can be inferred perfectly from short-run performance. Instead, we treat types as a structural source of cross-sectional inequality that interacts with incentives: for a fixed contract share, higher-type agents can generate larger contributions and hence (depending on the payment rule) larger wealth. This heterogeneity is precisely what makes distributional constraints non-trivial in our setting.

At each step, after actions and the next state realize, the environment produces a raw contribution signal $r_i(s_t, a_t, s_{t+1})$ for each agent i. This can be interpreted as revenue, completed tasks, or another auditable performance measure. The effective output attributable to agent i is $\theta_i r_i(\cdot)$, so that both the level and dispersion of realized contributions depend on the latent vector $(\theta_1, \ldots, \theta_n)$.

Actions, reject option, and timing. Each agent i has a finite action space $\mathcal{A}_i$ of "productive" actions and an augmented space $\tilde{\mathcal{A}}_i = \mathcal{A}_i \cup \{\text{reject}\}$. Choosing reject is interpreted as non-participation: the agent generates no output and receives no transfer in that period. We emphasize this modeling choice because it makes participation an endogenous equilibrium object rather than an exogenous constraint; operationally, it captures that platforms and organizations cannot force effort when outside options are available.

We adopt stationary Markov timing. At time t the public state $s_t \in \mathcal{S}$ is observed. The principal then offers a contract share $\alpha_t \in [0, 1]$, potentially randomized as a function of s_t. Agents observe (s_t, α_t) and simultaneously select actions $a_{i,t} \in \tilde{\mathcal{A}}_i$. The environment transitions according to

$$s_{t+1} \sim P(\cdot \mid s_t, a_t), \qquad a_t = (a_{1,t}, \ldots, a_{n,t}),$$

and raw contributions $r_i(s_t, a_t, s_{t+1})$ realize. Finally, contractual transfers are executed and payoffs accrue. This sequence repeats indefinitely with discount factor $\gamma \in (0, 1)$.

Contracts as linear shares. The contract space is $\mathcal{B} = [0, 1]$, where α represents a homogeneous share applied to all agents. Economically, α can be read as a take-rate complement: the principal keeps $(1 - \alpha)$ of measured output and pays α to the agents. Homogeneity is a deliberate restriction: it captures environments in which individualized contracts are infeasible or undesirable (e.g., due to regulation, simplicity, or transparency), and it makes distributional objectives meaningful because the principal cannot trivially equalize outcomes by agent-specific transfers.

The principal's stationary policy over contracts is denoted $\pi_p(\cdot \mid s; \phi)$, parameterized by ϕ. Agent i's stationary policy is $\pi_i(\cdot \mid s, \alpha; \psi_i)$, parameterized by ψ_i. Allowing π_p to be state-dependent reflects that contracting can respond to observable operating conditions (demand, congestion, seasonality), even if types remain hidden.

Per-period payoffs. Given state s_t, joint action a_t, next state s_{t+1}, and share α_t, agent i receives contractual reward

$$R_i(s_t, a_t, s_{t+1}, \alpha_t) = \Big(\alpha_t \, \theta_i \, r_i(s_t, a_t, s_{t+1}) - c_i\Big) \, \mathbf{1}[a_{i,t} \neq \text{reject}],$$

where c_i is a per-step cost of acting (potentially type- or state-dependent, but treated as exogenous primitives). The principal receives the residual share of total effective output from participating agents:

$$R_p(s_t, a_t, s_{t+1}, \alpha_t) = \sum_{i=1}^{n} (1 - \alpha_t)\, \theta_i\, r_i(s_t, a_t, s_{t+1})\, \mathbf{1}[a_{i,t} \neq \text{reject}].$$

Two features are worth highlighting. First, the reject option makes transfers and output jointly endogenous: as α_t falls, participation may collapse discretely, creating non-convexities in the mapping from contracts to long-run outcomes. Second, costs c_i create an incentive/insurance tradeoff even absent rejection: higher α increases agents' marginal returns to productive actions but reduces the principal's residual claim.

Long-run wealth as the outcome of interest. For any stationary policy profile $(\pi_p, \pi_1, \ldots, \pi_n)$, define discounted wealth for player $j \in \{1, \ldots, n, p\}$ as

$$w_j(\phi, \psi) = \mathbb{E}\Big[\sum_{t \geq 0} \gamma^t R_j(s_t, a_t, s_{t+1}, \alpha_t)\Big],$$

where $\psi = (\psi_1, \ldots, \psi_n)$ and the expectation is taken over the Markov chain induced by the stationary policies and the transition kernel P. We focus on stationary objects because they correspond to stable operating regimes: a platform's long-run take-rate policy and the steady-state behavior it induces.

To connect to a Stackelberg interpretation, we treat the principal as choosing ϕ anticipating that agents adapt to an equilibrium response. For analysis and learning, it is convenient to impose a regularized best-response selection that is unique and stable. Concretely, we posit that agents' responses can be represented by a mapping $\psi^*(\phi)$ (e.g., the fixed point of entropy-regularized policy-gradient dynamics), and we define the induced agent wealth vector

$$\mathbf{w}(\phi) = \big(w_1(\phi, \psi^*(\phi)), \ldots, w_n(\phi, \psi^*(\phi))\big).$$

This vector is the object to which fairness and distributional constraints will be applied.

Fairness metrics and participation. We take fairness to mean constraints on the *distribution of long-run wealth* across agents. Two canonical choices are inequality indices and worst-off guarantees. Let $G(\mathbf{w})$ denote the Gini coefficient computed from the agent wealth vector, and let $\text{Rawls}(\mathbf{w}) = \min_i w_i$ denote the Rawlsian minimum wealth. In addition, because agents can reject, we track an acceptance (participation) statistic, denoted $\text{Acc}(\phi)$, such as the stationary discounted frequency of non-reject actions:

$$\text{Acc}(\phi) = \mathbb{E}\Big[\sum_{t \geq 0} (1 - \gamma)\gamma^t \cdot \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}[a_{i,t} \neq \text{reject}]\Big].$$

Participation is not merely a welfare criterion; it is often a policy mandate (e.g., maintaining service coverage) and a practical constraint (a contract that induces mass exit is not implementable).

Differentiable surrogates. A technical complication is that $G(\cdot)$ and $\min(\cdot)$ are non-smooth, whereas the principal's update will rely on policy-gradient estimators. To keep the learning and comparative statics analysis within a differentiable framework, we replace non-smooth operators with smooth approximations. For the minimum wealth, a standard soft-min surrogate is

$$\widetilde{\min}_\beta(\mathbf{w}) = -\frac{1}{\beta}\log\sum_{i=1}^{n}\exp(-\beta w_i),$$

which converges to $\min_i w_i$ as $\beta \to \infty$ while remaining smooth for finite β. For the Gini coefficient, one convenient route is to use its pairwise absolute-difference form and smooth the absolute value, for instance by replacing $|x|$ with $\sqrt{x^2+\delta^2}$ for a small $\delta > 0$. Writing

$$G(\mathbf{w}) = \frac{1}{2n\bar{w}}\sum_{i=1}^{n}\sum_{j=1}^{n}|w_i - w_j|, \qquad \bar{w} = \frac{1}{n}\sum_{i=1}^{n} w_i,$$

we obtain a differentiable surrogate $\tilde{G}_\beta(\mathbf{w})$ by smoothing $|\cdot|$ and, if needed, stabilizing the denominator when $\bar{w}$ is near zero. The economic content is unchanged: these surrogates still penalize dispersion and enforce floors, but they do so in a way compatible with gradient-based optimization.

We view smoothing as an approximation device rather than a normative compromise. As β increases (and δ decreases), the surrogate constraints approach their intended hard counterparts, at the cost of potentially higher gradient variance and less numerical stability. This tradeoff is intrinsic: exact enforcement of non-smooth constraints is possible, but typically requires non-differentiable methods or substantially more complex estimators.

What the model abstracts from. Finally, we acknowledge two limitations. First, restricting to homogeneous linear shares rules out richer instruments (bonuses, nonlinear tariffs, individualized contracts) that could simultaneously improve efficiency and equity. Second, treating types as fixed and hidden focuses attention on persistent inequality, but it abstracts from learning about types and from endogenous human-capital accumulation. We adopt these simplifications because they isolate the core governance tension we study: when incentives and participation are strategic and the principal's control is limited, meeting explicit long-run distributional targets requires treating fairness and participation as first-class constraints rather than as a soft preference term.

4 Constrained contracting as a policy problem

We now formalize the principal's objective as a *profit maximization problem subject to explicit long-run distributional and participation requirements.* The economic motivation is straightforward: in many applications the "policy" variable is not an abstract social preference weight, but rather a mandated target (e.g., a minimum earnings floor, a cap on inequality, or a minimum service coverage rate). In such settings, the principal's problem is naturally posed as choosing a stationary contract policy that maximizes residual surplus *while meeting targets* that are specified in the same units that stakeholders monitor.

Formally, for a stationary principal policy parameterized by ϕ, let $\psi^*(\phi)$ denote the induced (regularized) stationary agent response mapping, and let $\mathbf{w}(\phi)$ collect the resulting agent wealths. The principal's constrained contracting problem is

$$\max_{\phi \in \Phi} \ w_p(\phi, \psi^*(\phi)) \quad \text{s.t.} \quad 1 - G(\mathbf{w}(\phi)) \geq \tau, \ \ \min_i w_i(\phi) \geq \rho, \ \ \mathrm{Acc}(\phi) \geq \kappa, \tag{1}$$

together with the implicit feasibility requirement that the principal's policy remains in the contract simplex (e.g., $\alpha \in [0, 1]$ in the tabular case, or its appropriate parametric analogue). We emphasize that the constraints in (1) bind the *endogenous* long-run outcomes of the interaction, not per-period transfers; they therefore encode both incentive effects and selection effects arising from rejection.

Smooth constraint surrogates and constraint functions. Because the principal will ultimately rely on gradient information, we work with smooth surrogates of the non-smooth fairness operators. Concretely, let $\tilde{G}_\beta(\mathbf{w})$ and $\widetilde{\min}_\beta(\mathbf{w})$ be differentiable approximations to the Gini coefficient and the minimum, respectively, as described in Section 3. We then define differentiable constraint functions

$$g_1(\phi) = \tau - (1 - \tilde{G}_\beta(\mathbf{w}(\phi))), \qquad g_2(\phi) = \rho - \widetilde{\min}_\beta(\mathbf{w}(\phi)), \qquad g_3(\phi) = \kappa - \mathrm{Acc}(\phi), \tag{2}$$

and impose $g_k(\phi) \leq 0$ for $k \in \{1, 2, 3\}$. Economically, this replacement does not change the nature of the tradeoff: higher τ or ρ still demands a contract that shifts expected surplus toward agents (directly via transfers and indirectly via induced behavior), and higher κ limits how aggressively the principal can push contracts toward the participation margin. What smoothing does change is *how* constraints "speak" to the gradient: instead of responding only to the single worst-off agent or to exact wealth order statistics, the surrogate produces informative marginal signals that are well-behaved under stochastic approximation.

Lagrangian formulation and KKT conditions. Given (2), the corresponding Lagrangian is

$$\mathcal{L}(\phi, \lambda) = w_p(\phi, \psi^*(\phi)) - \sum_{k=1}^{3} \lambda_k g_k(\phi), \qquad \lambda \in \mathbb{R}_+^3. \tag{3}$$

At an interior stationary solution (for the smoothed problem), the Karush–Kuhn–Tucker conditions take the familiar form

$$\nabla_\phi \mathcal{L}(\phi^*, \lambda^*) = 0, \qquad \lambda_k^* \geq 0, \qquad g_k(\phi^*) \leq 0, \qquad \lambda_k^* g_k(\phi^*) = 0 \quad \forall k. \tag{4}$$

Complementary slackness in (4) provides a useful economic interpretation of the dual variables: when, say, the Rawlsian floor is binding, $\lambda_2^* > 0$ measures the marginal profit cost of tightening that floor at the optimum; when it is slack, $\lambda_2^* = 0$ and marginal changes in the floor do not affect the locally optimal contract. This interpretation matters for practice because it distinguishes environments where fairness targets are *expensive* (high shadow price, strong tension with profit) from environments where they are *cheap* (low shadow price, little efficiency loss).

Existence of constrained stationary solutions. Although the problem is dynamic and strategic, existence of solutions for the smoothed constrained formulation follows a standard compactness-and-continuity route once we restrict attention to stationary policies. Under our standing assumptions that (i) the induced Markov chain is ergodic for any stationary profile, (ii) the best-response selection $\psi^*(\phi)$ is well-defined and continuous (indeed Lipschitz), and (iii) $w_p(\phi, \psi^*(\phi))$ and $g_k(\phi)$ are continuous in ϕ on a compact parameter set Φ, the feasible set

$$\Phi_{\text{feas}} = \{\phi \in \Phi : \ g_k(\phi) \leq 0 \ \ \forall k\}$$

is compact. If Φ_{feas} is nonempty, Weierstrass' theorem yields existence of at least one maximizer of the smoothed constrained problem. Moreover, if a Slater-type condition holds—namely, there exists some $\bar{\phi} \in \Phi$ such that $g_k(\bar{\phi}) < 0$ for all k—then KKT points (ϕ^*, λ^*) exist for the smoothed problem and characterize stationary optima via (4). We view Slater's condition as an economically meaningful feasibility assumption: it requires that the principal can satisfy the targets with some slack, ruling out knife-edge situations where meeting the constraints forces the system onto a boundary where small perturbations break feasibility.

Why a fixed weighted-sum is not a policy knob. It is tempting to replace (1) with a weighted-sum objective such as

$$\max_{\phi \in \Phi} \ w_p(\phi, \psi^*(\phi)) + \lambda \, F(\mathbf{w}(\phi)), \tag{5}$$

where F is an inequality penalty or welfare aggregator and $\lambda \geq 0$ is interpreted as "altruism." We caution against treating λ as an implementable policy lever in our setting for two related reasons.

First, the mapping from contracts to fairness and participation can be *non-convex and discontinuous* because of rejection. In a simple one-state, one-step instance with two types $\theta_H > \theta_L$, suppose each agent participates only if the offered share crosses an individual threshold: agent i accepts iff $\alpha\theta_i r - c_i \geq 0$. Then as α increases, the set of participants can change discretely: for low α both reject (no output, no wealth); for intermediate α only the high type participates (high inequality, possibly high principal profit); for high α both participate (lower inequality, potentially lower principal profit). Any fairness statistic computed on realized wealth (Gini, minimum wealth, or acceptance) will inherit these discontinuities. In such an environment, varying λ in (5) need not trace fairness levels smoothly: the optimizer can jump from a low-α regime (high profit from the high type only) directly to a high-α regime (both participate) at a critical λ, skipping intermediate target levels altogether. Consequently, there may exist target values (τ, ρ, κ) that are feasible under (1) but are *not attained* by any optimizer of (5) for a wide interval of λ.

Second, even absent discontinuities, a single scalar λ generally cannot encode *multiple* operational targets. Our problem features at least three conceptually distinct constraints: inequality control, worst-off protection, and participation. Increasing λ in (5) may improve one dimension while worsening another (e.g., raising the minimum wealth by increasing α might reduce acceptance if higher effort costs lead to strategic rejection in some states). In practice, stakeholders specify thresholds ("at least κ participation," "no agent below ρ") precisely because they are legible and enforceable; a fixed weight does not provide this governance guarantee, and it lacks an interpretable unit that would allow regulators or designers to choose it ex ante.

These considerations motivate treating fairness and participation as constraints with endogenous dual variables rather than as a fixed regularization term. The constrained formulation makes the policy question transparent (which targets are feasible, and at what shadow price), and it sets up a natural primal–dual learning procedure in which λ is *not tuned* but instead *adjusts endogenously* to enforce the specified targets. This is the perspective we operationalize in the next section via a two-timescale primal–dual actor–critic algorithm.

5 5. Algorithm: two-timescale primal–dual actor-critic (agents fast, principal+dual slow); practical implementation details (projection, smoothing, critic approximation).

We now describe the learning procedure that operationalizes the constrained formulation: a two-timescale primal–dual actor–critic in which agents adapt quickly to the currently offered contract policy, while the principal updates the contract policy and the associated dual variables slowly. The economic logic mirrors the comparative-statics intuition: agents are the "price takers" of the contract terms in the short run (they best respond to α), whereas the principal is the "policy maker" who adjusts α to satisfy long-run requirements at minimal shadow cost.

Parameterization and projections. We let the principal's stationary policy $\pi_p(\cdot \mid s; \phi)$ be either tabular (a categorical distribution over a discretized α grid) or continuous (e.g., a Gaussian policy over α followed by clipping). In either case we enforce feasibility by projection. Writing Φ for the admissible parameter set, the slow update takes the projected form

$$\phi_{t+1} = \Pi_\Phi\Big(\phi_t + \alpha_t\,\widehat{\nabla}_\phi \mathcal{L}(\phi_t, \lambda_t)\Big),$$

where Π_Φ denotes Euclidean projection (or, in practice, an equivalent reparameterization such as a sigmoid output ensuring $\alpha \in [0,1]$ pointwise). Similarly, dual variables are constrained to remain nonnegative, and we typically also cap them at a large $\Lambda_{\max}$ for numerical stability:

$$\lambda_{k,t+1} = \Pi_{[0,\Lambda_{\max}]}\Big(\lambda_{k,t} + \alpha_t\,\widehat{g}_k(\phi_t)\Big), \qquad k \in \{1,2,3\}.$$

This update is the stochastic-approximation analogue of complementary slackness: persistent positive violation $\widehat{g}_k > 0$ increases λ_k, which in turn tilts the principal's gradient toward satisfying constraint k; when a constraint is comfortably slack, $\widehat{g}_k < 0$ pushes λ_k back toward zero.

Fast agent adaptation. Each agent i maintains policy parameters ψ_i for $\pi_i(\cdot \mid s, \alpha; \psi_i)$ over $\tilde{\mathcal{A}}_i$ (including reject). On the fast timescale, agents ascend their own regularized objective using policy gradients:

$$\psi_{i,t+1} = \psi_{i,t} + \beta_t\Big(\widehat{\nabla}_{\psi_i} w_i(\phi_t, \psi_t) + \eta\,\widehat{\nabla}_{\psi_i}\mathbb{E}\big[\textstyle\sum_{t\geq 0}\gamma^t H(\pi_i(\cdot \mid s_t, \alpha_t; \psi_i))\big]\Big),$$

with stepsizes $\{\beta_t\}$ chosen so that $\alpha_t/\beta_t \to 0$. In economic terms, β_t being large relative to α_t is what makes the principal's environment "approximately stationary": before the principal materially changes the contract distribution, agents have largely adjusted their acceptance and effort decisions to the current terms.

Actor–critic structure and what the critics approximate. Both levels use critics to reduce variance and to accommodate function approximation. Concretely, each player $j \in \{1, \dots, n, p\}$ learns a value function $V_j(s)$ (or an action-value $Q_j(s, \alpha)$ for the principal) under the current joint policy. With temporal-difference learning, a typical update is

$$\omega_{j,t+1} = \omega_{j,t} + \xi_t \, \delta_{j,t} \, \nabla_{\omega_j} V_j(s_t; \omega_{j,t}), \qquad \delta_{j,t} = R_j(t) + \gamma V_j(s_{t+1}; \omega_{j,t}) - V_j(s_t; \omega_{j,t}),$$

where ω_j are critic parameters and ξ_t is a critic stepsize. The principal's policy gradient can then be written in advantage form,

$$\widehat{\nabla}_\phi w_p \;\propto\; \widehat{\mathbb{E}}\big[\nabla_\phi \log \pi_p(\alpha_t \mid s_t; \phi)\, \widehat{A}_p(s_t, \alpha_t)\big],$$

and analogously for each agent with $\nabla_{\psi_i} \log \pi_i(a_{i,t} \mid s_t, \alpha_t; \psi_i)\, \widehat{A}_i(s_t, \alpha_t, a_{i,t})$. The crucial point is practical rather than conceptual: the gradient estimator treats the current behavior of the other players as part of the data-generating process, and the timescale separation is what justifies interpreting this as approximating $\nabla_\phi w_p(\phi, \psi^*(\phi))$ rather than the gradient of a moving target.

Estimating and differentiating the constraints. The constraint functions $g_k(\phi)$ depend on long-run wealth objects $\mathbf{w}(\phi)$ and on acceptance. Operationally, we maintain running estimates of each agent's discounted wealth $\hat{w}_i$, e.g., by evaluating the agent value critic at a reference start distribution μ,

$$\hat{w}_i(\phi) \approx \widehat{\mathbb{E}}_{s_0 \sim \mu}\big[V_i(s_0)\big],$$

or by episodic Monte Carlo returns when episodes are available. We then compute smooth fairness statistics on $\hat{\mathbf{w}} = (\hat{w}_1, \dots, \hat{w}_n)$. For example, a standard soft-min surrogate is

$$\widetilde{\min}_\beta(\hat{\mathbf{w}}) = -\frac{1}{\beta} \log \Big(\sum_{i=1}^{n} e^{-\beta \hat{w}_i} \Big),$$

and we use any differentiable $\tilde{G}_\beta$ (e.g., based on smoothed pairwise absolute differences) to form $\hat{g}_1, \hat{g}_2$. Acceptance is estimated directly from behavior:

$$\widehat{\mathrm{Acc}}(\phi) = \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}[a_{i,t} \neq \text{reject}], \qquad \hat{g}_3 = \kappa - \widehat{\mathrm{Acc}}(\phi).$$

Because these are smooth functions of critic outputs (and of ϕ through π_p), we can backpropagate through the computation graph to obtain $\widehat{\nabla}_\phi g_k(\phi)$ as needed for $\widehat{\nabla}_\phi \mathcal{L} = \widehat{\nabla}_\phi w_p - \sum_k \lambda_k \, \widehat{\nabla}_\phi g_k$. Economically, smoothing plays a second role here beyond existence theory: it ensures that "nearly worst-off" agents and "nearly binding" inequality changes generate usable marginal signals for the principal, instead of producing a gradient that is identically zero except at kinks.

Putting the updates together. At iteration t, we (i) sample a batch of transitions under the current policies; (ii) update agent critics and agent actors using β_t; (iii) update the principal critic and then the principal actor using α_t and the current multipliers; (iv) compute $\hat{g}_k$ from the updated critics (and observed acceptance) and update λ using the same slow stepsize α_t. In implementations with neural policies, it is often helpful to (a) normalize rewards and constraint signals, (b) use gradient clipping, and (c) update λ using a slightly smaller effective stepsize to avoid oscillations when constraints are tight.

Approximation error and practical stability. Finally, we emphasize a limitation that is also a guide to practice: with function approximation, $\hat{w}_j$ and the resulting $\widehat{\nabla}$ are biased/noisy, and the algorithm converges only to a neighborhood whose radius is governed by this approximation error ϵ. In applied terms, the relevant tuning is therefore not an "altruism weight," but rather the numerical regime that keeps ϵ small: sufficient critic capacity, stable TD learning, and strong timescale separation so that the principal does not chase transitory agent behavior. When these conditions hold, the primal–dual mechanism makes the policy content of the problem explicit: the learning dynamics adjust λ endogenously to enforce the stakeholder targets, and the principal's contract policy adapts accordingly without requiring ex ante calibration of an opaque regularization weight.

Two-timescale limit and why KKT points are the right notion. The update rules above define a coupled stochastic process in $(\psi_t, \phi_t, \lambda_t)$, and the central technical difficulty is the endogenous non-stationarity: the principal's gradient is taken while agents are simultaneously learning. Our standing assumptions (ergodicity under stationary policies, unique regularized agent best responses, and differentiability of smoothed constraints) allow us to reduce this moving-target problem to a familiar constrained optimization object. The key step is to view the learning dynamics through their two-timescale ordinary differential equation (ODE) limit: because $\alpha_t/\beta_t \to 0$, the agent parameters evolve on a fast timescale and the principal–dual variables on a slow one. Intuitively, by the time the principal noticeably changes ϕ, the fast recursion has largely settled near the fixed point $\psi^*(\phi)$.

Formally, writing $\mathcal{B}(\phi, \psi)$ for the mean agent actor (and critic) drift under quasi-static ϕ, the fast limit is

$$\dot{\psi} = \mathcal{B}(\phi, \psi),$$

and assumption (H2) implies this ODE has a globally asymptotically stable equilibrium $\psi^*(\phi)$ that is Lipschitz in ϕ. On the slow timescale, the principal sees an effectively stationary Markov chain induced by $(\phi, \psi^*(\phi))$, so its mean drift is the projected primal–dual gradient flow for the smoothed constrained problem:

$$\dot{\phi} = \Pi_\Phi\big(\nabla_\phi \mathcal{L}(\phi,\lambda)\big), \qquad \dot{\lambda} = \Pi_{\mathbb{R}_+^K}\big(g(\phi)\big), \qquad \mathcal{L}(\phi,\lambda) = w_p(\phi,\psi^*(\phi)) - \sum_{k=1}^{K} \lambda_k g_k(\phi).$$

In this limit, stationary points coincide with KKT points of the smoothed constrained problem (under Slater-type conditions), which is precisely the equilibrium concept that corresponds to "profit-maximizing subject to policy-relevant fairness and participation targets."

Convergence to a neighborhood: what is guaranteed and what is not. Under tabular policies or sufficiently accurate linear approximation (so that critic error can be controlled), standard two-timescale stochastic approximation theory implies that the iterates (ϕ_t, λ_t) converge almost surely to an internally chain-transitive invariant set of the slow ODE. When the critic and gradient estimates are exact, this invariant set is contained in the KKT set. With function approximation, the invariant set inflates to an $O(\epsilon)$-tube around the KKT set, where ϵ measures the worst-case error in (i) value approximation and (ii) the induced policy-gradient/constraint-gradient estimates. Economically, ϵ captures the principal's "perception error" about long-run profit and the long-run distribution of agent wealth; the result tells us that the dual mechanism cannot enforce constraints more tightly than the information quality embodied in the critics.

A convenient way to state the implication is in terms of asymptotic primal suboptimality and constraint violation. Let $[x]_+ = \max\{x,0\}$. For a suitable Lyapunov function associated with the projected primal–dual flow (e.g., a smoothed saddle residual), one can show that, along subsequences and in expectation (and almost surely under stronger mixing and stepsize conditions),

$$\limsup_{t\to\infty} \max_k \big[g_k(\phi_t)\big]_+ \;\le\; C_1\,\epsilon, \qquad \limsup_{t\to\infty} \big(w_p(\phi^*,\psi^*(\phi^*)) - w_p(\phi_t,\psi^*(\phi_t))\big) \;\le\; C_2\,\epsilon,$$

for constants C_1, C_2 depending on Lipschitz moduli (of ψ^*, the smoothed constraints, and the policy class), projection radii, and the mixing rate of the ergodic Markov chain. The qualitative message is robust: improving critic accuracy and strengthening timescale separation reduces both long-run constraint slackness and profit loss, whereas simply increasing training time cannot drive these errors below the ϵ floor.

Why ergodicity and Lipschitz best responses matter economically. Assumption (H1) (ergodicity) is not merely a technical convenience: it is what makes "long-run wealth" well-defined and learnable from on-policy data. If the chain were not mixing, then the same contract policy could generate different wealth vectors depending on transient path dependence, and the principal could not interpret empirical constraint estimates $\hat{g}_k$ as stable objects. Assumption (H2) (unique, Lipschitz best response) plays a similarly economic role. With a reject option and heterogeneous types, agents' acceptance and effort can change sharply as α crosses an implicit participation threshold. Entropy regularization smooths this response and ensures that small policy changes by the principal do not cause discontinuous swings in behavior. Without such regularity, the principal's slow recursion can "chase" a discontinuous correspondence $\psi^*(\phi)$, and the ODE approximation breaks down in precisely the regimes where constraints are most policy-relevant (near binding fairness or participation limits).

Stability via primal–dual updates versus greedy principal dynamics. It is useful to contrast the primal–dual mechanism with a tempting but flawed alternative: a greedy principal actor update that ascends estimated profit $\widehat{\nabla}_\phi w_p$ while treating constraints as either (i) ex post diagnostics or (ii) fixed penalties with hand-tuned weights. In our setting, greediness is destabilizing for two related reasons. First, because agents adapt quickly, a profit-increasing move in ϕ today can induce a best-response shift tomorrow that reduces acceptance, collapses output, and changes the wealth distribution nonlinearly. Second, fairness and participation constraints create effective "kinked" feasible regions even after smoothing: near the boundary, the principal must trade off profit against shadow costs that are endogenous and state-dependent.

The dual variables provide precisely the missing state variable. When some constraint k is persistently violated, λ_k increases and the principal's update becomes

$$\nabla_\phi \mathcal{L}(\phi, \lambda) = \nabla_\phi w_p(\phi, \psi^*(\phi)) - \sum_k \lambda_k \nabla_\phi g_k(\phi),$$

so the algorithm endogenously reweights directions in contract space that repair the binding constraint. In economic terms, λ_k is the shadow price of violating the target; its adaptation is what stabilizes learning near the boundary. By contrast, a greedy principal that ignores λ can repeatedly push the system into low-acceptance regimes (agents reject), at which point profit gradients become uninformative (no output) and the process may exhibit oscillations: high α temporarily restores acceptance, then greed pushes α down again, and so on. The primal–dual dynamics damp these oscillations because constraint violations accumulate "debt" in λ, making it increasingly unattractive to revisit the same infeasible region.

What the $O(\epsilon)$ bound means in practice. The neighborhood result is often interpreted pessimistically, but its policy implication is concrete. If the planner's goal is to meet a participation floor κ and a fairness target

τ reliably, the central tuning is not an altruism weight but the learning regime that controls ϵ: critic capacity, adequate exploration of (s, α), stable temporal-difference learning under ergodicity, and strong timescale separation so that constraint estimates track the quasi-stationary wealth vector. When these conditions are satisfied, the algorithm behaves like a constrained policy optimizer: it approaches contracts whose induced stationary equilibrium is nearly feasible, with constraint violation proportional to the unavoidable approximation error, and with multipliers that are interpretable as the marginal cost of tightening each target. This sets up the next section's purpose: in a tractable one-shot model we can compute, in closed form, how tightening a Rawlsian or inequality target maps into a higher optimal share α^*, providing a sanity check for the monotone comparative statics that the learning dynamics implement in the Markov setting.

Closed-form sanity check: how fairness targets map into a higher linear share. Before turning to simulation evidence, we find it useful to pin down a tractable case where we can compute the contract implied by an explicit fairness threshold. The point is not realism—a one-shot environment strips away state dependence and dynamic incentives—but transparency. In particular, the Markov analysis above predicts monotone comparative statics: tightening equality, minimum-wealth, or participation targets should weakly increase the optimal share offered to agents (or increase the frequency of high-α contracts in the dynamic case). A stylized model lets us verify this prediction algebraically and clarifies when such targets are *feasible* at all under a homogeneous linear-share instrument.

Environment and agent behavior (one period with reject). Consider a single period in which the principal offers a common linear share $\alpha \in [0, 1]$. Each agent i either rejects (yielding zero activity and zero wealth) or chooses effort $e_i \geq 0$. Output is

$$y_i = \theta_i e_i,$$

and effort has quadratic disutility plus a fixed operating cost $c_i \geq 0$ (interpretable as the per-step cost in the Markov model, collapsed into one period). If agent i participates, her payoff is

$$w_i(\alpha) = \alpha y_i - \frac{1}{2} e_i^2 - c_i,$$

and if she rejects, her payoff is 0. The principal receives the residual claim on output from participating agents:

$$w_p(\alpha) = (1 - \alpha) \sum_{i=1}^{n} y_i \cdot \mathbf{1}[i \text{ participates}].$$

Given α, a participating agent solves $\max_{e_i \geq 0}\{\alpha\theta_i e_i - \frac{1}{2}e_i^2 - c_i\}$, yielding the interior best response

$$e_i^*(\alpha) = \alpha\theta_i, \qquad \max_{e_i \geq 0}\Big(\alpha\theta_i e_i - \frac{1}{2}e_i^2\Big) = \frac{1}{2}\alpha^2\theta_i^2.$$

Thus participation is a simple threshold rule: agent i participates iff her net surplus is nonnegative,

$$\frac{1}{2}\alpha^2\theta_i^2 - c_i \geq 0 \quad\Longleftrightarrow\quad \alpha \geq \underline{\alpha}_i := \frac{\sqrt{2c_i}}{\theta_i}.$$

The resulting realized (gross) wealth profile is therefore

$$w_i(\alpha) = \Big[\frac{1}{2}\alpha^2\theta_i^2 - c_i\Big]_+, \qquad \text{and} \qquad w_p(\alpha) = (1-\alpha)\alpha\sum_{i=1}^{n}\theta_i^2\,\mathbf{1}[\alpha \geq \underline{\alpha}_i].$$

Two features mirror the Markov game. First, the reject option makes outcomes piecewise and potentially kinked in α (precisely where participation constraints matter). Second, heterogeneity in θ_i creates unequal rents even under a common share, so equality constraints bind through *entry* (who participates) and through the compression created by the fixed cost c_i.

Rawlsian and participation thresholds as lower bounds on α. A Rawlsian floor $\min_i w_i(\alpha) \geq \rho$ (with $\rho \geq 0$) immediately implies full participation and a uniform lower bound on the share. Since $w_i(\alpha) \geq \rho$ requires $\frac{1}{2}\alpha^2\theta_i^2 - c_i \geq \rho$ for every i, we obtain

$$\alpha \;\geq\; \alpha_{\text{Rawls}}(\rho) := \max_i \frac{\sqrt{2(c_i+\rho)}}{\theta_i}.$$

Likewise, an acceptance/participation constraint of the form $\text{Acc}(\alpha) \geq \kappa$ reduces here to requiring that at least a κ-fraction of agents satisfy $\alpha \geq \underline{\alpha}_i$. Writing $\underline{\alpha}_{(1)} \leq \cdots \leq \underline{\alpha}_{(n)}$ for the order statistics of the thresholds, a minimal share satisfying acceptance is

$$\alpha \;\geq\; \alpha_{\text{Acc}}(\kappa) := \underline{\alpha}_{(\lceil \kappa n\rceil)}.$$

Both constraints therefore enter as explicit *lower bounds* on α, and hence tighten monotonically in ρ and κ.

An explicit mapping for an inequality target (two-agent case). To get an explicit expression for an inequality constraint, it is convenient to look at $n = 2$ with common fixed cost $c_1 = c_2 = c$ and types $\theta_H > \theta_L > 0$. When both participate (i.e., $\alpha \geq \sqrt{2c}/\theta_L$), wealths are

$$w_H(\alpha) = \frac{1}{2}\alpha^2\theta_H^2 - c, \qquad w_L(\alpha) = \frac{1}{2}\alpha^2\theta_L^2 - c.$$

For two agents, the Gini coefficient has the closed form

$$G(\mathbf{w}) = \frac{w_H - w_L}{w_H + w_L}, \qquad 1 - G(\mathbf{w}) = \frac{2w_L}{w_H + w_L}.$$

Imposing an equality target $1 - G(\mathbf{w}(\alpha)) \geq \tau$ (with $\tau \in (0,1)$) is equivalent to a lower bound on the wealth ratio:

$$\frac{w_L(\alpha)}{w_H(\alpha)} \;\geq\; \frac{\tau}{2-\tau}.$$

Substituting the expressions above yields an explicit threshold for α^2. Letting $x = \alpha^2$, the constraint becomes

$$\frac{x\theta_L^2 - 2c}{x\theta_H^2 - 2c} \;\geq\; \frac{\tau}{2-\tau},$$

which rearranges to

$$x\Big(\theta_L^2 - \frac{\tau}{2-\tau}\theta_H^2\Big) \;\geq\; \frac{4c(1-\tau)}{2-\tau}.$$

Hence, provided the coefficient on x is positive (a feasibility condition),

$$\theta_L^2 > \frac{\tau}{2-\tau}\theta_H^2 \quad\Longleftrightarrow\quad \tau < \frac{2\theta_L^2}{\theta_H^2 + \theta_L^2},$$

the inequality constraint is satisfied iff

$$\alpha \;\geq\; \alpha_{\text{Gini}}(\tau) := \sqrt{\frac{4c(1-\tau)}{(2-\tau)\big(\theta_L^2 - \frac{\tau}{2-\tau}\theta_H^2\big)}}.$$

This formula makes two economic points precise. First, when heterogeneity is too large relative to the desired equality level (large θ_H/θ_L and large τ), the target is simply infeasible under a common linear share: no choice of α can sufficiently compress wealth. Second, conditional on feasibility, $\alpha_{\text{Gini}}(\tau)$ is increasing in τ: a tighter equality target forces a higher share because higher α raises both wealths while diminishing the relative impact of the common fixed cost c, thereby reducing measured inequality.

Optimal share and monotone comparative statics. In the region where the relevant agents participate, principal profit is a concave quadratic in α:

$$w_p(\alpha) = (1-\alpha)\alpha \sum_{i \in \mathcal{P}(\alpha)} \theta_i^2,$$

where $\mathcal{P}(\alpha) = \{i : \alpha \geq \underline{\alpha}_i\}$. Conditional on a fixed participant set, the unconstrained maximizer is $\alpha^{\text{u}} = 1/2$. With fairness/participation constraints, the constrained optimum therefore takes the simple form

$$\alpha^* = \Pi_{[0,1]}\Big(\max\big\{\alpha^{\text{u}},\ \alpha_{\text{Rawls}}(\rho),\ \alpha_{\text{Acc}}(\kappa),\ \alpha_{\text{Gini}}(\tau)\big\}\Big),$$

interpreting $\alpha_{\text{Gini}}(\tau)$ only when the feasibility condition holds and both agents are required to be active. This delivers the comparative statics we rely on in the Markov setting: tightening ρ, κ, or τ weakly increases the optimal share, and increasing heterogeneity tightens feasibility and raises the implied shadow price of equality. The same logic also clarifies why, in more complex environments, fairness constraints often bind through participation margins: once the unconstrained optimum α^{u} is below the most restrictive lower bound, the optimum is pulled to the boundary and the principal gives up profit to avoid rejections and to meet distributive targets.

Limitations and why this still matters for the dynamic case. This one-shot model is intentionally narrow: it abstracts from state dependence, from persistence in wealth, and from the possibility that optimal dynamic contracts vary with s_t. Nevertheless, it serves as a diagnostic for our learning-based approach. It shows (i) how a policy target (a numerical τ or ρ) translates into an interpretable minimum generosity level, (ii) how heterogeneity generates feasibility limits for common-share instruments, and (iii) why piecewise behavior (entry/reject) can generate sharp changes that motivate smoothing and primal–dual adaptation. With this sanity check in hand, we now turn to experiments in a canonical Markov environment (Coin Game), where we can test whether the algorithm reproduces these monotone patterns under dynamics, partial observability of types, and function approximation.

Experimental testbed: reproducing Coin Game patterns under contracting. We next move from the one-shot sanity check to a canonical Markov environment where distributive concerns and strategic interaction are both salient. Our goal is twofold: (i) reproduce the qualitative "fairness–efficiency" patterns reported for the Coin Game under reward shaping, and (ii) stress-test whether our *constrained contracting* formulation (with primal–dual updates and smoothed constraints) achieves *policy-relevant targets*— $1 - \text{Gini} \geq \tau$, $\min_i w_i \geq \rho$, and $\text{Acc} \geq \kappa$—without hand-tuning a fixed altruism weight. The key empirical object is the learned stationary principal policy $\pi_p(\cdot \mid s; \phi)$ over linear shares α, together with the induced agent responses.

Coin Game with a principal and hidden types. We consider a grid-world Coin Game variant with $n \in \{2, 4\}$ agents. At each step, agents move and may collect coins; raw contribution signals $r_i(s_t, a_t, s_{t+1})$ correspond to environment-defined increments in value attributable to agent i (e.g., the value of coins collected, potentially including externalities from who benefits). Hidden types θ_i scale these contributions, so two agents who behave identically can generate systematically different effective outputs. Agents additionally face a per-step operating cost c_i, and retain a reject option (implemented as an action that yields no movement and no output). The principal observes the public state s_t and realized contribution signals (or their contractual basis), but not θ_i, and chooses $\alpha_t \in [0, 1]$ each step. We evaluate performance using discounted wealths w_p and $\mathbf{w}$, along with realized constraint statistics computed from long-run rollouts.

Algorithmic instantiation and evaluation protocol. Agents update on the fast timescale via entropy-regularized actor–critic, which operationalizes the unique best-response mapping required by (H2). The principal updates ϕ on a slower timescale using stochastic gradients of the Lagrangian, while dual variables λ_k follow projected ascent on the (smoothed) constraint violations. To separate learning transients from stationary behavior, we report two sets of metrics: *during training* (moving averages of constraint satisfaction and profit) and *post-training* (evaluation rollouts with fixed policies, reporting means and variability across seeds). For fairness, we compute the empirical Gini coefficient on agent wealths (agents-only) as well as the Rawlsian minimum; for participation, we report the fraction of non-reject actions.

Main reproduction: constrained contracting traces a controlled fairness–efficiency frontier. Across seeds, we find that tightening targets produces the monotone patterns predicted by the Markov/KKT analysis and the one-shot algebra. Increasing τ (more equality) raises the learned average share $\mathbb{E}[\alpha_t]$, and reallocates mass in $\pi_p(\cdot \mid s)$ toward higher-α actions; similarly, increasing ρ pushes the policy toward higher shares until either the minimum-wealth floor is met or α saturates near 1. Increasing κ reduces rejection episodes primarily by lifting the lower tail of offered shares, which is consistent with participation margins being the binding channel in environments with costly effort or risky dynamics. In all three sweeps, principal profit w_p declines smoothly once the relevant constraint becomes binding, while constraint violations shrink to the expected $O(\epsilon)$ neighborhood determined by critic/gradient error. Importantly, the learned dual variables become informative diagnostics: when a target is slack, its multiplier remains near zero; when binding, the corresponding λ_k grows and stabilizes, providing an endogenous "shadow price" of the policy target.

Sensitivity to heterogeneity and costs: feasibility becomes the binding issue. We then vary the heterogeneity ratio $\theta_{\max}/\theta_{\min}$ and the level/spread of costs c_i. Two regularities emerge. First, as heterogeneity increases, equality targets become harder to satisfy with a homogeneous linear-share instrument: the algorithm responds by increasing α and the fairness multiplier, but beyond a point it cannot fully close the gap (mirroring the feasibility limits highlighted by the closed-form two-agent calculation). Second, higher costs shift behavior toward reject unless shares rise; consequently, participation constraints interact strongly with distributive constraints, and the "cheapest" way to satisfy $1-\mathrm{Gini} \geq \tau$ often becomes keeping more agents active rather than attempting to equalize wealth conditional on a fixed participant set. These observations matter operationally: when targets are infeasible under the instrument class, the constrained approach does not hide the problem—it reveals it through persistent positive violations and exploding multipliers.

Baselines: Greedy, Fixed, welfare regularization, and variance regularization. We compare against four baselines. (1) *Greedy* optimizes w_p without constraints; it reliably yields low α, frequent rejection when costs are nontrivial, and high inequality, even when modest shares would preserve profit while stabilizing participation. (2) *Fixed* uses a constant α chosen by oracle grid search to maximize profit under each target; it provides a useful upper bound for *stationary constant contracts* but is typically dominated by state-dependent π_p when the Coin Game has phases with differing marginal returns to effort. (3) *Welfare regularization* optimizes $w_p + \lambda F(\mathbf{w})$ for a fixed λ and a chosen fairness proxy F; consistent with our theoretical caution, it often fails to hit prescribed thresholds except at finely tuned values of λ, and can exhibit discontinuous jumps in behavior as λ varies (notably when rejection becomes optimal for some agents). (4) *Variance regularization* penalizes $\mathrm{Var}(\mathbf{w})$; it can reduce dispersion, but it is neither threshold-calibrated nor aligned with Rawlsian floors, and it may "equalize by depressing" by reducing output rather than stabilizing participation. Across settings, the constrained method is the only approach that reliably targets (τ, ρ, κ) directly.

Ablations: smoothing is a stability tool, not a cosmetic choice. We ablate the smoothing of the Gini and minimum operators by varying the surrogate temperature parameter β. With insufficient smoothing (large β, close to nonsmooth), the dual updates become high-variance near kinks induced by reject, and training exhibits oscillations: λ spikes, α overshoots, and policies alternate between generous and extractive regimes. With excessive smoothing (small β), training is stable but biased: constraints are satisfied for the surrogate while the true (nonsmoothed) statistics can drift, especially for the Rawlsian floor which is sensitive to tail events. An intermediate smoothing regime yields the best tradeoff: stable learning with small surrogate–true gaps, supporting our use of smooth constraints as an analytically motivated approximation rather than an ad hoc trick.

Ablations: timescale separation governs constraint adherence under learning. Finally, we test the two-timescale prediction directly by varying the principal-to-agent learning-rate ratio. When the principal updates too quickly relative to agents, the induced environment becomes nonstationary; empirically, we observe persistent constraint violations and cycles in α_t and λ_t, consistent with the ODE intuition that the slow variables fail to track $\psi^*(\phi)$. As the principal step size is reduced (holding agent learning fixed), training becomes markedly smoother: acceptance stabilizes, multipliers settle, and post-training evaluation satisfies targets up to approximation error. This ablation is practically important: in deployments where agents adapt online (or where the principal faces model misspecification), conservative principal updates act as a safeguard against transient unfairness.

Summary of what the experiments establish. Taken together, the Coin Game experiments support three claims. First, target-based constraints produce interpretable, monotone shifts in contracts that mirror the algebra of the tractable model, but now under dynamics and partial observability of types. Second, compared to fixed-λ regularization, primal–dual learning is materially easier to operationalize because it directly enforces policy thresholds and exposes infeasibility. Third, the two methodological "details"—smoothing and timescale separation—are in fact central to making constrained contracting behave predictably in environments with reject-induced kinks and strategic adaptation.

Discussion and limitations: beyond the controlled testbed. Our analysis and experiments are intentionally organized around a clean message: if a policymaker cares about *targets* (e.g., a minimum-wealth floor or an equality threshold), then formulating contracting as a constrained problem and learning via primal–dual updates is operationally closer to the governance question than tuning a fixed fairness weight. That said, the step from this message to deployment in real marketplaces or multi-agent digital ecosystems is not automatic. The key limitations are not merely engineering details; they are structural features of dynamic principal–agent environments: non-convexity of the induced optimization landscape, multiplicity of equilibria and equilibrium selection, and the normative and statistical fragility of any chosen fairness metric under meaningful heterogeneity.

Non-convexity: KKT points are not global solutions. Even with a single scalar instrument $\alpha \in [0, 1]$, the mapping $\phi \mapsto w_p(\phi, \psi^*(\phi))$ is typically non-concave once the environment is dynamic and agents can reject. Policy-gradient methods therefore target stationary points of a smoothed Lagrangian rather than globally optimal constrained contracts. In practice, this means two things. First, we should expect sensitivity to initialization and optimization hyperparameters, and we should interpret the resulting contract policy as one *attainable* governance-compatible solution rather than the unique "best" one. Second, constraints can interact with non-convexity in a distinctive way: the algorithm may satisfy targets while leaving substantial principal surplus on the table, or conversely may extract high profit while hovering near constraint boundaries where estimation error causes occasional violations. For applications where violations are legally or ethically costly, it is not enough to rely on asymptotic statements of $O(\epsilon)$ constraint error; one needs explicit finite-sample safety margins (e.g., tightening τ, ρ, κ by slack terms) or robust variants of the constraints that incorporate uncertainty sets.

Multiple equilibria and path dependence: uniqueness is an assumption, not a fact. Our theoretical convergence guarantee leans on a unique and Lipschitz agent response $\psi^*(\phi)$, implemented empirically via entropy regularization. This device is analytically convenient and often stabilizing, but it is also a strong modeling choice: real strategic environments can have multiple equilibria even under stationary contracts, and different learning dynamics can select different equilibria. When equilibrium selection is endogenous, the principal is not merely optimizing payoffs subject to constraints; the principal is effectively influencing *which equilibrium* the population coordinates on. This raises two deployment-relevant concerns. One is *predictability*: an audit based on one equilibrium selection mechanism may fail under a different learning rule or under a different population composition. The second is *manipulability*: if some agents can anticipate how the principal updates ϕ and λ, they may steer learning toward equilibria that are privately favorable while still satisfying coarse aggregate constraints. A natural research direction is to replace single-equilibrium analysis with set-valued best responses and to adopt equilibrium-robust objectives (e.g., maximize worst-case principal profit over equilibria consistent with observed adaptation), though this will likely sharpen the tradeoff between tractability and realism.

Fairness metrics are not neutral: what is being equalized? Targeting $1 - \text{Gini}$ and a Rawlsian minimum is a deliberate choice because these objects are interpretable and correspond to familiar policy desiderata (dispersion control and floors). But metrics embed values. In environments with hidden types θ_i that scale contributions, equalizing realized wealth can be defended as solidarity, or criticized as blunting rewards for productivity, depending on the normative frame. Moreover, if θ_i captures not only "skill" but also differential access to resources, discrimination, or structural disadvantage, then equalizing *outcomes* may be closer to a corrective justice view than equalizing *opportunities.* Our framework can accommodate alternative targets, but the burden shifts to the modeler to justify them. For example, one may constrain inequality in *utility net of costs* rather than transfers alone, or impose group-conditional constraints when agents belong to protected categories. Conversely, if heterogeneity is deemed morally relevant (e.g., training investment), a policymaker may prefer constraints that protect minima without compressing the entire distribution. The main limitation is that constrained optimization enforces the metric one writes down; it does not resolve which metric is legitimate.

Meaningful heterogeneity creates identification and measurement problems. When types are hidden and contributions are noisy, the principal does not observe the welfare object it is constraining; it observes a proxy based on realized trajectories and an accounting rule for r_i. In the Coin Game this accounting is unambiguous; in deployments it often is not. Consider settings where outputs are delayed, jointly produced, or strategically attributable (e.g., collaborative code, content moderation, or supply-chain tasks). Then the mapping from behavior to r_i is itself contestable and can be gamed. If we constrain fairness in wealth computed from a flawed attribution model, we risk "fairness by accounting" rather than fairness in lived outcomes. This suggests that the governance problem is partly upstream: designing auditable contribution signals and cost models that are stable under strategic behavior. In practice, this may require combining our contracting layer with mechanism-design tools (peer prediction, anti-collusion rules, or randomized audits) so that r_i is both informative and incentive-compatible.

Auditability and transparency: state-dependent policies can be hard to justify. A principal policy $\pi_p(\alpha \mid s; \phi)$ can be statistically effective yet institutionally unacceptable if it is opaque. Many 2026 deployment contexts (platform compensation, autonomous labor allocation, or enterprise AI copilots) require that compensation rules be explainable, predictable, and contestable. State dependence is a double-edged sword: it can improve efficiency by tailoring incentives to phases of the task, but it can also look like discretionary treatment unless the state variables are clearly defined and non-sensitive. One practical constraint is therefore *policy class restriction*: we may need monotone or sparse contract policies (e.g., a small menu of α values triggered by coarse, audited state features) rather than a high-dimensional neural policy. This restriction can be integrated into our framework as a parameterization choice, but it will generally tighten feasibility and lower achievable profit, making the feasibility diagnostics (persistent $g_k(\phi) > 0$ and growing multipliers) more central, not less.

Statistical compliance requires monitoring, not just training. Even if a learned contract satisfies targets on average during evaluation, deployments face distribution shift: new agents enter, costs drift, and the environment changes. Since our constraints are expressed in long-run discounted wealth, they are inherently statistical and require ongoing estimation. A compliance-minded implementation would therefore treat (λ_k) not only as training variables but as *monitoring signals*: rising multipliers can flag when a target is becoming expensive or infeasible under current conditions. However, this also creates an implementation challenge: if multipliers are allowed to grow without bound, short-run responses can become extreme (e.g., very high α to "buy" participation), which may be unacceptable or financially destabilizing. Real systems may need bounded multipliers, change-control procedures, and human-in-the-loop review when constraints approach violation.

Instrument limits and institutional constraints. We deliberately focus on homogeneous linear shares to isolate the governance logic, but this instrument is often too blunt under high heterogeneity. When $\theta_{\max}/\theta_{\min}$ is large, hitting a Rawlsian floor with a common α can force transfers that are excessively generous for high types, or can push α to corners where principal incentives collapse. In many real settings, institutions also restrict *how* transfers can be made: payments may be episodic rather than per-step, contracts may not condition on fine-grained states, and personalization may be legally constrained due to discrimination risk. These constraints can and should be modeled explicitly (e.g., episodic contracts, limited menus, or group-blindness constraints), but they will change both feasibility and the shape of the attainable frontier.

What we can credibly claim. The practical takeaway is thus conditional. Constrained contracting is a governance-ready *template*: it translates high-level policy targets into enforceable learning objectives, produces interpretable shadow prices, and surfaces infeasibility rather than hiding it behind a tuned λ. But its real-world reliability depends on (i) whether equilibrium selection is stable under agent adaptation, (ii) whether the welfare metrics correspond to legitimate normative commitments under heterogeneity, and (iii) whether the measurement and auditing stack for r_i, costs, and acceptance is itself trustworthy. These are not peripheral caveats; they are the boundary conditions under which the methodological advantage of target-based constraints can be converted into institutional practice.

Conclusion: constrained contracting as the governance-ready alternative to λ-regularization. We can now restate the organizing claim of the paper in the language of the model. When the principal cares about *policy targets*—a minimum wealth floor ρ, an inequality bound τ (e.g., $1 - \text{Gini} \geq \tau$), and a participation requirement κ—the economically natural object is a constrained problem with explicit feasibility and explicit shadow prices. In contrast, a weighted-sum objective $w_p + \lambda F(\mathbf{w})$ is a preference representation only when the frontier is well-behaved; in the dynamic Markov setting with reject options, non-convexities, and state dependence, λ becomes an unstable surrogate for governance. Our contribution is therefore not merely an algorithmic trick. It is a translation: from informal notions of "fairness" to auditable constraints, from ad hoc tuning to KKT conditions, and from a fixed altruism parameter to endogenous multipliers λ_k that quantify the marginal cost of meeting a stated target.

What the constrained formulation buys us, operationally. The constrained viewpoint makes three operational differences salient. First, it separates *preference* from *compliance*: rather than asking the designer to pick λ so that an emergent fairness statistic lands near a desired level, the designer specifies the level and the learning rule seeks policies whose stationary outcomes satisfy it. Second, it produces diagnostics that are legible to governance stakeholders. Persistent positive constraint residuals $g_k(\phi) > 0$ indicate infeasibility under the policy class and instrument limits, while rising multipliers λ_k reveal which requirement is binding and how expensive it is at the margin. Third, it aligns with the institutional fact that many deployments are accountable to thresholds (minimum pay, non-discrimination standards, participation guarantees) rather than to a continuous social-welfare weight. In that sense, primal–dual learning is not only a computational method for solving our model; it is a model of how a regulator-like principal can adapt incentives while keeping the targets explicit.

What we have (and have not) established theoretically. Under standard assumptions that make the induced control problem well-posed—ergodicity, smooth surrogates for fairness functionals, and a unique regularized best response $\psi^*(\phi)$—the smoothed constrained problem admits KKT points, and a two-timescale actor–critic dynamic tracks a neighborhood of that set, with asymptotic constraint violation on the order of the critic/gradient error ϵ. The key economic content of this statement is that the principal can treat the agent population as approximately equilibrated on the fast timescale and can then adjust the contract policy and shadow prices on the slow timescale as if solving a constrained optimization problem. At the same time, we do not claim global optimality, finite-time guarantees of exact satisfaction, or robustness to arbitrary equilibrium selection. These gaps are not merely technical; they delineate the boundary between "target-based learning" as a persuasive governance template and as a fully reliable mechanism.

Reading the multipliers as prices of policy. A useful way to interpret the dual variables is as *shadow prices of legitimacy.* If λ_{Rawls} is large, then raising the minimum-wealth floor by one unit imposes a large profit cost at the margin under the current instrument class (here, a homogeneous share α). If λ_{Acc} spikes, then participation is being "purchased" via contract generosity, suggesting either that the environment has shifted (costs c_i rose, types θ_i changed) or that the state-dependent policy has entered regions where agents' outside option dominates. This economic interpretation matters because it is actionable: multipliers can be monitored as part of compliance operations, can trigger re-training or review, and can inform whether institutional constraints (budgets, legal limits on personalization, menu restrictions) are binding in a way that makes some targets unattainable without expanding the contract space $\mathcal{B}$.

From a single scalar share to richer contract spaces. We deliberately study $\mathcal{B} = [0, 1]$ to isolate the governance logic, but the conclusion points to a broader design problem: what is the minimal contract language that makes a given set of targets feasible without destroying incentives? Even small generalizations—state-dependent menus, episodic bonuses, or two-part tariffs—can relax the tension between profit and minima when heterogeneity in θ_i is large. The constrained approach scales naturally to these spaces: we still maximize w_p subject to $g_k(\phi) \leq 0$, but now the feasibility set may expand dramatically. The open question is not whether we can write down richer $\mathcal{B}$, but how to do so while preserving auditability and non-discrimination constraints (e.g., contracts that are group-blind, monotone in coarse task states, or implementable under limited observability of s_t). In practice, the economically relevant frontier is the one induced by institutional constraints on what can be conditioned upon, not the one induced by mathematical convenience.

Open problem I: equilibrium selection without assuming uniqueness. The most conceptually important extension is to drop the assumption that $\psi^*(\phi)$ is single-valued. With multiple equilibria, the principal's problem becomes intrinsically bilevel and partly adversarial: the same contract policy can induce different wealth vectors depending on learning dynamics, coordination devices, or population composition. A governance-relevant principal would want guarantees that targets hold across plausible equilibrium selections, i.e., constraints of the form $\sup_{\psi \in \Psi^*(\phi)} g_k(\phi, \psi) \leq 0$, or at least high-probability satisfaction under a model of adaptation. Developing tractable surrogates for such *equilibrium-robust* constraints, and understanding when dual ascent remains stable in that set-valued setting, are central theoretical tasks if we want to move from "works under a stabilizing regularizer" to "works under realistic strategic diversity."

Open problem II: measurement, attribution, and incentive-compatible signals. A second bottleneck is the welfare accounting layer. Our constraints are written in terms of wealth w_i, which in turn depends on the reward basis $r_i(s, a, s')$ and on costs c_i. In deployments, both objects are measured with error and are strategically contestable. This suggests an interaction between contracting and mechanism design: we may need to *design* contribution signals that are robust to manipulation (peer prediction, audits, randomized checks), and we may need constraints that are robust to mismeasurement (distributionally robust or ambiguity-averse versions of g_k). A promising direction is to replace point estimates of $\mathbf{w}(\phi)$ by confidence sets and to enforce constraints with statistical margins, turning compliance into a statement like $g_k(\phi) \leq -\delta_k$ where δ_k is calibrated to monitoring error and desired risk tolerance.

Open problem III: risk, dynamics, and non-stationarity. Finally, policy targets are rarely stationary in practice. Costs drift, new agents enter, and the meaning of "participation" changes with outside options. This raises two related extensions. One is to move from expected discounted constraints to risk-sensitive or tail constraints (e.g., CVaR-style constraints on low-wealth events) so that compliance is not only average-case. The other is to treat the primal–dual updates as an *online control* problem with change detection and bounded adjustment rates, reflecting real institutional frictions. Technically, this pushes us beyond stationary Markov analysis toward non-stationary objectives and regret-style notions of performance subject to constraints. Substantively, it forces us to articulate what it means to be "fair over time": is the constraint on long-run wealth, on per-period pay, on opportunity sets, or on transition dynamics between states?

Closing perspective. The broader message is that constrained contracting provides a coherent vocabulary for bringing economic governance questions into learning systems: we specify targets, we expose feasibility, and we interpret multipliers as the price of policy commitments. Weighted-sum regularization remains useful as an exploratory tool, but it is not a substitute for target-based compliance when the mapping from λ to outcomes is discontinuous, non-monotone, or institutionally opaque. The open problems above do not weaken the case for constraints; they clarify what must be built around them—equilibrium-robustness, incentive-compatible measurement, and risk-aware monitoring—for the promise of "governance-ready" learning to survive contact with heterogeneity, strategic behavior, and shifting environments.