选择即权力：面向有界决策权的约束强化

Jose Manuel de la Chica Rodriguez
AI实验室负责人*AI*实验室，桑坦德集团，马德里，西班牙

Juan Manuel Vera Díaz
高级AI研究员*AI*实验室，桑坦德集团，马德里，西班牙

arXiv:2603.02019v1 [cs.MA] 2 Mar 2026

摘要——**《选择即权力》指出，在高风险智能体系统中，上游选择权威，而非内部目标错位，是风险的主要来源。然而，原有框架是静态的：治理约束限制了选择权力，却不会随时间调整。在本研究中，我们通过引入激励式选择治理，将该框架扩展至动态情境；在外部强制执行的主权约束下，对评分和归约器参数进行强化更新。**

我们将选择形式化为一个受约束的强化过程：参数更新被投影到由治理定义的可行集合中，从而防止选择集中度超过规定上限。在多个受监管金融场景中，无约束强化在反复反馈下持续收敛为确定性主导，且在较高学习率下尤为明显。相比之下，激励式治理能够在维持有界选择集中度的同时，实现自适应改进。

基于投影的约束将强化从不可逆锁定转变为受控适应，同时以治理债务量化优化压力与权威边界之间的张力。这些结果表明，只要在每个更新步骤都执行主权约束，学习动态便可与结构多样性并存，从而为在高风险智能体系统中整合强化提供了一种有原则的方法，而无需放弃受限的选择权限。

关键词：约束强化学习、人工智能系统的机制设计、多智能体学习、对抗鲁棒性、人工智能治理

通讯作者：
juanma.vera@gruposantander.com

一、引言

A. 从静态选择治理到自适应激励

自主智能体系统正越来越多地部署于受监管且事关重大的领域。在这些领域中，决策不可逆、需要接受外部审计，并受到制度约束。在 [1],中，我们指出，此类系统的主要风险所在并非仅仅是智能体计算什么，还在于它们获准选择、抑制和呈现什么。我们提出了一种治理架构，通过将认知、选择和行动划分为不同领域，并施加由外部强制执行的约束，从机制上限定选择权。

该框架将自身定位为与基于对齐的安全方法并列，例如基于人类反馈的强化学习（RLHF）[2], [3], 、宪法式人工智能 [4],以及受约束的强化学习 [5], [6]。虽然这些方法能够塑造内部目标或限制行动执行，但有关算法问责制的制度分析表明，仅凭对齐和透明度并不能消除上游选择机制中蕴含的结构性权力 [7], [8]。

原有治理架构表明，通过架构分离和机械式执行，可以限制选择权。然而，该架构有意保持静态：治理参数固定不变，系统也不会根据反馈随时间进行调整。尽管机械式治理能够防止结果遭到确定性捕获，却无法强化正确的选择，也无法在反复交互中不断改进。

本文旨在解决这一局限。

B. 核心问题

我们探讨选择治理能否在不重新导致权力集中的情况下实现自适应。具体而言，我们提出以下问题：

治理架构能否纳入
激励机制，以强化
随着时间推移作出正确选择，同时保
持受限的选择权限，并应对
对抗性和多智能体条件？

激励机制是现代机器学习的基础。强化学习提供了一个原则性框架，通过反馈改进决策策略 [9]，；多智能体强化学习（MARL）研究相互作用的智能体如何在共同或竞争目标下进行适应 [10], [11]。然而，无约束强化可能会加剧选择集中度、引发反馈回路，并使共同优化奖励的智能体形成共谋均衡[12], [13]。

在治理敏感环境中，这种动态可能会重新引入静态选择约束原本旨在防止的权力集中。

C. 激励式选择治理

为应对这一挑战，我们引入选择即权力 *2.0*：一种激励式治理框架，在该框架中，评分智能体和选择归约器都会根据反馈信号进行更新，但更新仅限于外部强制执行的主权约束范围内。

概括而言，系统的运行方式如下。受治理选择机制会根据给定的上下文选择一个智能体执行任务。被选中的智能体生成输出，由完美评估器进行评估，并提供二元正确性反馈。随后，这些反馈用于更新评分偏好和选择参数。至关重要的是，所有更新都受到治理规则的约束，以防止系统陷入确定性支配或消除结构多样性。

这将选择治理从静态的界定机制转变为动态的约束学习过程。

D. 贡献

本文作出五项贡献：

1）我们将激励式选择治理形式化为一种在外部强制执行的主权约束下运行的强化过程。2）我们提出针对评分参数和约简器参数的双重更新学习，同时保留机械性强制执行。3）我们通过理论分析表明，受约束更新能够在迭代适应过程中限制选择集中度。4）我们在受监管的金融场景中对该框架进行实证评估，衡量其在时间动态变化中的稳定性和治理债务。

E. 论文结构

本文其余部分安排如下。第二节将本研究置于强化学习、机制设计和多智能体系统的研究脉络中。第三节对激励式治理模型进行形式化。第四节介绍约束强化架构。第五节描述实验设置。第六节报告实证发现。第七节分析稳定性与动态特性。第八节讨论局限性与未来工作。

II. 相关工作

本研究处于强化学习、机制设计、多智能体系统和人工智能治理的交叉领域。尽管这些领域都涉及适应与控制的某些方面，但没有一个直接对外部强制执行的主权约束下的激励式选择治理进行形式化。

A. 强化学习与约束优化

强化学习（RL）为在反馈下优化序贯决策策略提供了通用框架 [9]。策略梯度方法[14]和行动者-评论家架构广泛应用于大规模系统中。然而，当奖励信号不完整或设定有误时，不受约束的强化学习可能导致意料之外的行为。

约束强化学习通过纳入显式安全约束来解决这一问题，这些约束通常借助约束马尔可夫

决策过程（约束马尔可夫决策过程） [15], [5] 进行建模。屏蔽方法引入运行时安全过滤器，以阻止不安全的行动 [6]。虽然这些方法限制了行动的执行，但它们通常假定行动空间是固定的，并且并不显式管控上游选择权威。

我们的工作有两点不同。首先，我们将选择视为一种受治理的资源，而非固定的行动空间。其次，强化更新会投影到外部定义的主权约束集上，确保适应不会侵蚀结构性治理保障。

B. 奖励建模与对齐

奖励建模和基于人类反馈的强化学习（RLHF）已成为使大型模型与人类偏好保持一致的核心方法 [2], [3]。宪法式人工智能则通过基于规则的批评和自我改进进一步约束行为 [4]。

这些方法主要着眼于使内部目标保持一致。然而，正如算法问责制研究 [7] 以及人工智能对齐的不完全契约分析 [8]，所指出的，内部对齐并不能消除决策权方面的结构性不对称。即便智能体已实现对齐，仍可能通过上游选择或框定施加不成比例的影响。

我们的框架对基于对齐的方法起到补充作用：它治理强化如何随时间推移重塑选择权，而非仅仅塑造内部偏好。

C. 多智能体强化学习

多智能体强化学习（MARL）研究相互作用的智能体如何在共享或竞争目标下进行适应 [10], [12]。即使没有明确的通信，合作均衡和竞争均衡也可能出现，尤其是在重复交互环境中 [11]。

博弈论分析表明，在重复交互中，学习动态可能导致合谋均衡出现 [13], [16]。在多智能体强化学习环境中，智能体可能会隐式协调以最大化联合奖励，从而形成涌现的合作策略，但这些策略未必与系统级目标一致。

我们的工作将选择治理扩展到明确考虑多智能体动态。我们提出了一种智能体选择的形式化度量，并分析约束强化能否限制选择概率的放大。

D. 机制设计与激励相容

机制设计为在策略环境中构建激励提供了正式工具[17]。在算法环境中，机制设计原则已应用于拍卖、分配规则和分布式系统。

然而，经典机制设计假设智能体在预先定义的分配规则内行动。相比之下，智能体人工智能系统能够影响选项空间本身的构建。因此，选择治理不仅需要约束激励，还需要限制其对选项生成和缩减的权限。

我们的框架可以理解为一种受约束的机制，其中强化更新被视为策略性行动，并投射到维护治理的约束集中。

E. 人工智能治理与问责

制度分析强调，高风险系统的安全性需要可执行的约束，而非纯粹的规范性指南 [7]。治理机制必须在因果层面发挥作用，而不能仅停留在认知层面。

我们先前的工作 [1] 已将这一理念形式化，用于静态系统。本文将其扩展至动态情境，在这些情境中，激励会随时间推移重塑决策策略。本文的关键贡献在于证明，只要主权约束仍由外部强制执行，自适应强化便可与受限的选择权限并存。

F. 本文的定位

总而言之：

- 与标准强化学习不同，我们治理的是选择，而不仅仅是行动。
- 与受约束强化学习不同，我们通过主权投影来限制强化学习更新。
- 与基于人类反馈的强化学习不同，我们调节的是结构性权限，而非内部偏好对齐。
- 不同于经典多智能体强化学习，我们明确地对选择动态进行建模并加以约束。

据我们所知，这是首个将约束强化、选择治理和多智能体选择分析整合到统一架构中的框架，专为高风险智能体系统而设计。

III. 问题设定：激励式选择治理

我们将 [1] 中提出的「选择即权力」扩展为动态形式。该设定将静态治理架构扩展为序贯学习框架，其中选择策略和约简器参数根据评价反馈进行更新，同时仍受外部强制执行的主权约束限制。

A. 系统概述

我们考虑一个以 $t = 1, \ldots, T$ 为索引的离散时间过程。在每个时间步：

1) 给定一个任务上下文 x_t 。2) 受治理选择机制从有限候选集 $\mathcal{A} = \{A_1, \ldots, A_n\}$ 中选择一个智能体 A_i 。3) 被选中的智能体生成输出 $o_t = A_i(x_t)$。4) 完美评估器 E 返回正确性信号 $r_t \in \{-1, +1\}$。5) 根据 r_t 更新选择参数和归约器参数，同时受治理约束限制。

我们假设评估器是正确且确定性的。因此，我们不考虑奖励标注中的不确定性，而只关注选择动态。

B. 候选智能体与上下文

令 $\mathcal{A}$ 表示候选智能体的有限集合。每个智能体 A_i 都将上下文映射为输出：

$$A_i : \mathcal{X} \to \mathcal{O}.$$

任务上下文 $x_t \in \mathcal{X}$ 取自外生分布 $\mathcal{D}$。我们假设 $\mathcal{D}$ 在整个评估时域内保持平稳。

与经典强化学习场景不同[9], ，智能体不会直接优化行动，而是由系统根据给定的上下文，优化选择哪个智能体来采取行动。

C. 受治理的选择策略

在时间 t，系统维护选择参数 θ_t。这些参数定义了一种选择策略：

$$\pi_{\theta_t}(A_i \mid x_t),$$

其决定在上下文 x_t 下选择智能体 A_i 的概率。

至关重要的是， π_{θ_t} 并非不受约束。它必须满足外部施加的主权约束，以限定选择权。

D. 归约器参数

除了选择权重外，系统还维护归约器参数 ϕ_t。这些参数包括以下数量：

- 多样性配额，
- 阈值，
- 探索系数，
- 方差界。

归约器参数决定如何筛选并呈现评分后的候选项，遵循先前关于《选择即权力》 [1]的研究所定义的架构结构。

E. 主权约束集

我们将治理约束形式化为闭合可行集：

$$\theta_t \in \mathcal{C}, \qquad \phi_t \in \mathcal{G}.$$

- $\mathcal{C}$ 约束选择参数。
- $\mathcal{G}$ 约束归约器参数。

这些集合编码了如下不变量：

- 最低探索概率，
- 有界选择集中度，
- 多样性要求，
- 熵隔离约束。

重要的是，约束集由外部定义，且不能由学习过程修改。

F. 强化更新

观察到奖励 r_t后，系统会更新参数。

a) 选择更新。：选择参数通过投影梯度上升进行更新：

$$\theta_{t+1} = \Pi_{\mathcal{C}}(\theta_t + \alpha r_t \nabla_\theta \log \pi_{\theta_t}(A_i \mid x_t)),$$

其中，$\Pi_{\mathcal{C}}$ 表示投影到可行集$\mathcal{C}$，$\alpha > 0$ 表示学习率。

b) 归约器更新。：归约器参数以类似方式更新：

$$\phi_{t+1} = \Pi_{\mathcal{G}}(\phi_t + \beta r_t \nabla_\phi L_{\text{selection}}(\phi_t)),$$

其中，$L_{\text{selection}}$ 表示选择质量的可微替代指标，$\beta > 0$ 表示学习率。

基于投影的更新是约束优化中的标准方法 [18]。在此，它们确保每次迭代都满足主权约束。

G. 选择集中度

为量化选择权，我们定义智能体 A_i的经验纳入概率：

$$P_t(A_i) = \mathbb{E}[\pi_{\theta_t}(A_i \mid x_t)].$$

我们将选择集中度在时刻 t 定义为：

$$\text{SC}_t = \max_i P_t(A_i).$$

有界选择权力要求：

$$\sup_t \text{SC}_t \leq \gamma,$$

其中，$\gamma < 1$由治理约束决定。

H. 稳定性目标

与经典强化学习不同，我们不要求收敛到一个最优不动点。相反，我们要求经验稳定性：

- 有界选择集中度，
- SC_t 的有界方差。

形式上，我们将经验稳定性定义为：

$$\limsup_{t \to T} \text{SC}_t \leq \gamma.$$

这将适应性治理定义为一种约束动力系统，而非不受约束的优化过程。

I. 与标准强化学习的比较

在标准强化学习中， [9]，策略更新不受约束，除非采用可选的正则化。相比之下，我们的模型在每次迭代时都会强制投影到治理约束集上，即使在持续正强化的情况下，也能从结构上限制权限。

本文余下部分将分析该约束动力系统在对抗条件和串谋条件下的影响。

IV. 激励式选择治理的架构

本节介绍实现第三节所提出形式模型的系统架构。其目标是在保留外部强制执行的主权约束的同时，将基于强化的适应机制融入原有的“选择即权力”架构。

该架构在静态流水线（CEFL →评分 →归约器 →呈现门）的基础上，引入了评估反馈回路和约束参数更新。由此形成的系统可以理解为一种受约束的动态机制，而非固定的决策流水线。

A. 架构概述

在每个时间步 t，系统都会执行以下阶段：

1) **候选扩展（CEFL）**。2) **评分与特征评估**。3) **受治理归约与选择**。4) **输出执行**。5) **评估与奖励分配**。6) **约束参数更新**。

算法 1 总结了激励式选择治理架构的完整迭代流程。该算法明确区分了受治理候选生成、自适应评分、基于归约器的过滤、评估和约束参数更新。重要的是，强化信号会影响评分和归约器参数，但每次更新都会投影到外部定义的主权约束集上。这一投影步骤是该架构的保障机制，使激励式治理有别于不受约束的强化学习：适应过程会发生，

但仅限于通过机械方式强制执行的可行区域内。该流程还纳入了故障显式机制，确保策略性协调或过度集中会触发纠正性干预，而非无声漂移。

与静态架构的关键区别在于引入了最后一个阶段：在该阶段中，评分和归约器参数会根据评价反馈进行更新，并投影到治理约束集上。

B. 候选扩展与冻结层 *(CEFL)*

候选扩展与冻结层保持非自适应。其作用是使用外部定义规则和熵源，为上下文 x_t 生成候选全集 $C_t \subseteq \mathcal{A}$。

CEFL 满足以下属性：

- 它不具备智能体属性，且没有可学习参数。
- 它受外部治理，不受强化更新的影响。
- 这保证合格候选项的暴露概率非零。

通过保持 CEFL 静态，该架构可防止强化动态导致候选项集合本身发生坍缩。

C. 可学习权重评分模块

为每个候选项 $A_i \in C_t$ 分配一个评分向量：

$$s_{i,t} = f_{\theta_t}(A_i, x_t),$$

其中， θ_t 为可学习的评分参数。

这些参数可用于对效用、风险、稳定性或其他领域特定标准赋予权重。强化会修改 θ_t，但更新会投影到约束集 $\mathcal{C}$上。

a) 评分主权约束：编码在 $\mathcal{C}$ 中的约束可能包括：

- 选择分布最小熵。
- 集中度上界。
- 强制性多样性阈值。

投影确保强化不会消除结构性探索。

D. 自适应参数受控归约器

归约器使用参数 ϕ_t进行确定性过滤和随机选择。

典型的归约器组件包括：

- 硬策略约束，
- 方差钳制，
- 帕累托前沿过滤，
- 多样性分桶，
- 随机抽样。

在选择即权力中，某些归约器参数（例如阈值、探索系数）会变为自适应参数。不过，这些更新会被投影到可行集 $\mathcal{G}$ 上，以保持不变量。

a) 归约器约束示例：

- $\sigma_{\text{最大值}}$ 受固定上限约束。
- 探索配额 $k \geq k_{\text{最小值}}$。
- 多样性分桶 $\geq d_{\text{最小值}}$。

这可以防止归约器在强化过程中退化为确定性的 top-1 选择。

E. 评估模块

候选智能体生成输出 o_t后，评估器 E 会分配奖励 $r_t \in \{-1, +1\}$。

我们假设：

- 评估器正确性。
- 确定性奖励分配。
- 评估器与候选智能体之间不存在策略性交互。

这使我们能够将选择动态与奖励噪声区分开来，并分析理想监督下的强化。

F. 受约束的双重更新机制

评估后， θ_t 和 ϕ_t 都会更新。

a) 评分更新： 参数 θ_t 通过对奖励加权的对数似然进行梯度上升来更新，随后投影到 $\mathcal{C}$上。

b) 归约器更新： 参数 ϕ_t 使用可微代理目标的奖励加权梯度进行更新，然后投影到 $\mathcal{G}$上。

这种双重更新设计使我们的架构区别于标准强化学习流程；在标准流程中，要么只更新策略参数，要么将约束作为惩罚项施加，而非采用硬投影。

算法 1 激励式选择治理（迭代过程）

需要： 候选集 $\mathcal{A}$，初始选择参数 $\theta_0 \in \mathcal{C}$，约简器参数 $\phi_0 \in \mathcal{G}$，评估器 E，学习率 α，$\beta > 0$，时间跨度 T

确保： 审计日志 $\mathcal{L}$

$\mathcal{L} \leftarrow \varnothing$

对于 $t = 1$ 到 T **执行**

  上下文： 接收任务上下文 x_t

  CEFL： 生成候选池 $C_t \leftarrow \text{CEFL}(x_t)$ ▷ 外部受治理候选生成

  评分： 对每个 $A_i \in C_t$，计算分数 $s_{i,t} \leftarrow f_{\theta_{t-1}}(A_i, x_t)$

  硬约束： 应用确定性过滤器以获得 $\tilde{C}_t$

  归约器： 计算呈现集合 $S_t \leftarrow \text{REDUCER}(\tilde{C}_t, \{s_{i,t}\}, \phi_{t-1})$

  选择： 示例智能体 $A_k \sim \pi_{\theta_{t-1}}(\cdot \mid x_t, S_t)$

  执行： 生成输出 $o_t \leftarrow A_k(x_t)$

  评估： 获取正确性信号 $r_t \leftarrow E(o_t)$

  评分更新：

    $\theta'_t \leftarrow \theta_{t-1} + \alpha\, r_t \nabla_\theta$ 对数 $\pi_{\theta_{t-1}}(A_k \mid x_t)$

    $\theta_t \leftarrow \Pi_{\mathcal{C}}(\theta'_t)$

  归约器更新：

    $\phi'_t \leftarrow \phi_{t-1} + \beta\, r_t \nabla_\phi L_{\text{selection}}(\phi_{t-1}; S_t, \{s_{i,t}\})$

    $\phi_t \leftarrow \Pi_{\mathcal{G}}(\phi'_t)$

  审计与日志记录： 追加（x_t，C_t，$\{s_{i,t}\}$，S_t，A_k，o_t，r_t，θ_t，ϕ_t）到 $\mathcal{L}$

  如果 投影裁剪了参数**则**

    触发故障显式报告机制（阻止 / 限流 / 警报）并记录事件

  结束 if

结束 for

返回 $\mathcal{L}$

G. 系统属性

该架构满足以下性质：

1) **受限的选择权限：** 选择集中度仍受 $\mathcal{C}$ 和 $\mathcal{G}$ 中编码的约束限制。 2) **自适应强化：** 正确的选择会在允许范围内提高选择概率。 3)**故障显式报告行为：** 违反约束会触发阻止或限流。

H. 与约束优化的关系

基于投影的更新是凸优化和约束学习中的标准方法 [18], [15]。然而，在此架构中，投影的作用不仅是确保可行性，还在于维护选择权所涉及的机构主权不变量。

因此，强化在一个编码了治理规则的有界可行域内运行，从而将选择治理转变为一个约束动力系统。

下一节将介绍用于在对抗和串谋条件下评估该架构的实验设置。

V. 实验设置

本节介绍用于评估序贯适应和多智能体动态下激励式选择治理的实验方案。评估的目标是考察强化更新下的经验稳定性和有界选择集中度。

A. 序贯评估方案

与《选择即权力》中的静态评估不同，这里的实验是在长度为 T 的时间跨度内进行的。每次迭代：

1) 从平稳分布中采样一个任务上下文 x_t。2) 受治理的选择策略选择一个智能体。3) 被选中的智能体生成输出。4) 完美评估器赋予奖励 $r_t \in\{-1,+1\}$。5) 通过约束强化更新评分和约简器参数。

每个实验运行包含 $T = 250$ 次迭代。所有结果均在独立随机种子上取平均值。

B. 领域场景

实验在三个受监管金融场景中进行：

- **欺诈检测：** 在异构欺诈模式下于欺诈监控智能体之间进行选择。
- **支付基础设施监控：** 跨基础设施监控智能体进行异常检测。
- **QBR 分析：** 跨分析智能体进行金融摘要与风险报告。

每个场景包含五种任务变体，从而形成异质上下文，并在真实的领域差异下对选择策略进行压力测试。

C. 候选智能体池

候选集由 $n = 7$ 个具有异质特征档案的结构化金融人工智能智能体组成：

- 风险档案，
- 稳定性评分，
- 延迟，
- 可审计性评分，
- 监管合规标签。

CEFL 使用确定性嵌入相似度（基于哈希以确保可复现性）生成上下文条件候选子集，过采样因子为 $\alpha > 1$。

D. 强化配置

学习率设置如下：

$$\alpha \in \{0.01, 0.05\}, \qquad \beta \in \{0.01, 0.05\}.$$

约束集 $\mathcal{C}$ 和 $\mathcal{G}$ 强制执行：

- 最低探索概率 $p_{\text{最小}} = 0.1$，
- 最大选择集中度上界，
- 最小多样性分桶 $d_{\text{最小}} = 2$，
- 方差限幅 $\sigma_{\text{最大}} = 0.18$。

每次更新后都会执行投影。

E. 基线比较

我们将激励式治理架构与以下方案进行比较：

- **静态治理：** 无强化更新。
- **无约束强化学习：** 无投影的强化。
- **标量 Top-K：** 无多样性约束的确定性聚合。

这些基线能够分离出：

- 强化的效应，
- 主权投影的效应，
- 归约器适配的效应。

F. 指标

我们评估五类指标。*1)* 选择集中度：

$$\mathrm{SC}_t = \max_i P_t(A_i)$$

我们报告随时间变化的轨迹以及最终时期的集中度。

2) 强化选择集中度（*RSC*）：衡量集中度随时间的变化：

$$\mathrm{RSC} = \mathrm{SC}_T - \mathrm{SC}_0$$

3) 治理稳定性指数（*GSI*）：

$$\mathrm{GSI} = 1 - \frac{\mathrm{Var}(\mathrm{SC}_t)}{\mathrm{Var}_{\text{unconstrained}}}$$

4) 治理债务（*GD*）：触发显式失败保护机制的迭代占比。

G. 评估目标

这些实验旨在检验：

1) 约束强化是否会提高选择集中度。2) 投影界限是否能防止陷入确定性主导。3) 治理机制是否能保持经验稳定性。

VI. 实验结果

本节报告第五节所述动态学习率扫描实验的结果。分析涵盖三种金融场景、四种治理架构，以及两种学习率（0.01和0.05）。

A. 选择集中度的时间演化

图 1a–1c 展示了各情景下，选择集中度 SC_t 随学习率和架构变化的时间演化。

在所有领域中，不受约束的_强化学习都表现出集中度快速增长，尤其是在 lr =0.05条件下。在欺诈检测领域（图 1a）中，不受约束的_强化学习达到 $\mathrm{SC}_T = 1.0$，表明在反复强化下出现确定性锁定效应。

相比之下，激励式治理表现出受控的强化。集中度随时间上升，但在所有情景下始终低于 1。这种有界行为在图 1a、1b 和 1c 中清晰可见：激励式轨迹始终低于不受约束的强化轨迹。

B. 学习率效应

图 2 对比了在 lr =0.01 和 $lr = 0.05$ 下的 SC_t 轨迹。

提高学习率会加快强化动态：

- 在无约束的_强化学习中，较高的学习率会带来更快的锁定效应和更高的最终集中度。
- 在激励模式下，较高的学习率会提高集中度，但投影可防止确定性支配。

这表明，学习率控制强化速度，而主权投影界限则限制其幅度。

C. 实际选择主导度（头部智能体占比）

尽管 SC_t 衡量的是选择分布的集中程度，单个智能体的实际支配性更能由被选择最多的智能体的经验选择频率来刻画，记为最高_智能体_份额。图3绘制了这一指标随时间的变化，涵盖每种架构和学习率。

这些模式在结构上各不相同：

- **无约束强化学习**。在两种学习率下，实际支配性都单调上升并趋近于1。0，其中 $lr = 0.05$ 收敛得更快。这证实了确定性锁定效应：一旦某个智能体累积了偏差，强化就会使其无限放大。
- **标量_topk**。 支配性从早期迭代开始就维持在较高且稳定的水平。由于归约器是确定性的，且缺乏多样性，选择频率反映的是结构性单一化，而非学习式适应。
- **激励式治理**。 支配性在强化作用下逐步上升，但始终稳定在 1 以下。在 $lr = 0.01$ 下，头部智能体占比缓慢上升，并在中等水平趋于平稳。在 $lr = 0.05$ 下，上升更快，但仍受到限制，不会达到确定性崩溃。
- **静态**。 选择频率保持平缓且较低，反映出缺乏学习。

这些变化轨迹揭示了一个关键区别：投影约束无法阻止强化驱动的偏好形成，却能防止不可逆的支配。激励式治理允许偏差不断累积，同时维持智能体之间的剩余竞争。

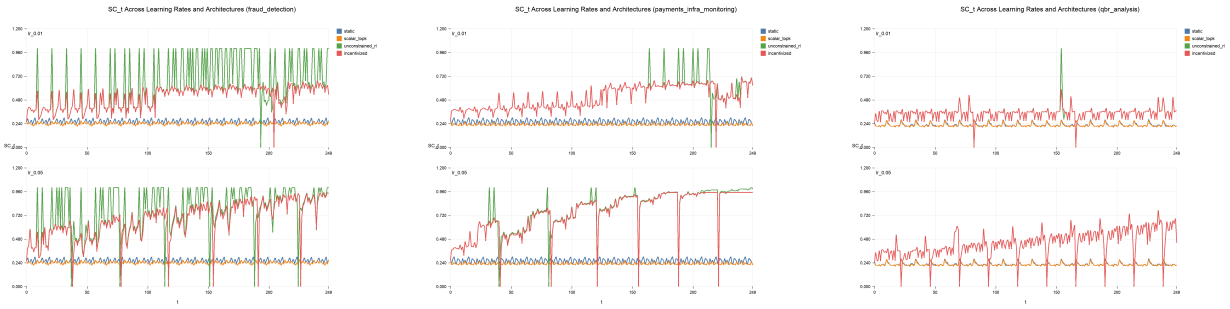
D. 治理债务权衡

在较高学习率下，激励式治理会产生可测量的治理债务（GD_动态），这反映了强化尝试超出集中度约束时发生的投影裁剪事件。

这种权衡在图1b中清晰可见：激励式轨迹趋近集中度边界，但会定期通过投影得到稳定。

E. 任务级强化动态

顶部_份额图进一步表明，强化效应在重复的任务情境下最为明显。在无约束_强化学习中，重复的奖励信号会迅速推动对单一智能体的近乎排他性选择。在激励式治理中，重复的上下文仍会强化高性能智能体，但集中度稳定在亚确定性水平。



(a) 欺诈检测 (b) 支付基础设施监控器-ing (c) QBR 分析

图 1：不同治理架构和学习率下选择集中度的时间演化（$\mathrm{SC}_t = \max_i P_t(A_i)$）。无约束强化会导致集中度迅速上升，并在某些情况下造成确定性锁定。激励式治理通过基于投影的主权约束，在保留受限的选择权限的同时实现自适应强化。较高的学习率会加快强化动态，但在治理约束下仍受到限制。

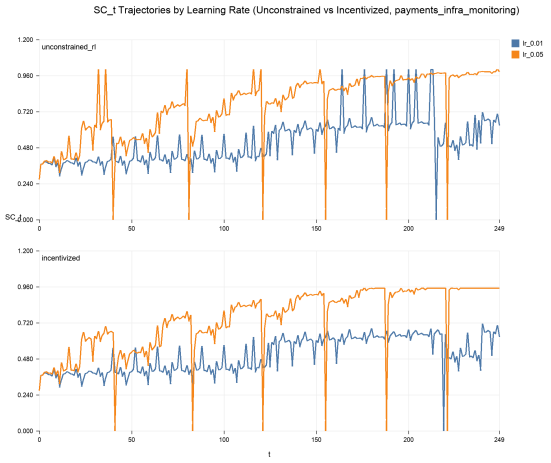


图 2：比较学习率（0.01 与0.05）对选择集中度（SC_t）时间演化的影响。较高的学习率会加快强化动态，使集中度增长得更快。在无约束强化下，这会导致迅速锁定；而激励式治理则通过基于投影的主权约束限制集中度增长。

重要的是，无约束_强化学习与激励模式之间的差距会随着更高的学习率而扩大。在 $lr = 0.05$下，无约束强化迅速坍缩，而激励式强化虽会加速，但由于投影裁剪而仍保持有界。

这一实证差异证实，主权投影不仅调节分布层面的集中度（以 SC_t 衡量），也调节随时间变化的实际选择主导地位。

F. 场景级指标摘要

表 I 总结了不同场景、治理架构和学习率下的主要评估指标。指标包括平均奖励、平均选择集中度（mean_SC）、初始和最终集中度（$\mathrm{SC}_0, \mathrm{SC}_T$）、强化选择集中度（RSC）、集中度的时间方差（var_SC）以及动态治理债务（GD_dynamic）。

G. 实证模式总结

图 1a–3 所示的时间轨迹，以及表 I 报告的定量值，揭示了不同场景和学习率下的一致结构模式。

首先，无约束强化会产生确定性的集中化。图中显示，SC_t 和实际头部智能体份额均呈单调增长，而表中则证实，在较高学习率下，SC_T趋近于或达到1.0 。因此，不经过投影的强化会导致完全锁定效应。

其次，激励式治理能够实现强化，同时保持有界权力。

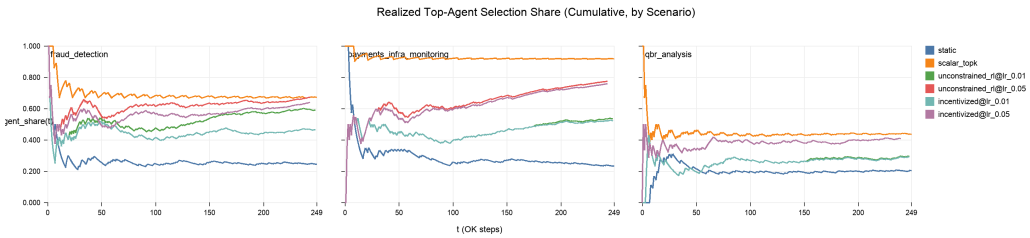


图 3：不同场景和治理架构下实际实现的累计顶级智能体选择份额。每个面板显示被选次数最多的智能体所获选择比例随成功步数的变化。无约束强化学习呈现出向确定性支配单调收敛的趋势，且学习率越高，锁定效应越快。标量 top-k 由于确定性聚合，从早期迭代开始便表现出结构性高支配程度。相比之下，激励式治理通过基于投影的主权约束，在提高对高性能智能体的选择偏好的同时，使其稳定在确定性崩溃程度以下。静态模式则基本保持不变，反映出其缺乏学习过程。

表 I：按场景、架构和学习率汇总的聚合指标。数值为 $T = 250$ 次迭代的平均值。

场景	Mode		LR	平均奖励_	平均 SC_	SC_T	RSC	GD_动态
欺诈检测	静态		0.01	0.222	0.258	0.269	0.010	0.000
	标量_topk		0.01	0.422	0.239	0.268	0.009	0.000
	不受约束	_rl	0.05	0.362	0.742	1.000	0.741	0.000
	激励		0.05	0.358	0.669	0.942	0.683	0.004
支付监控	静态		0.01	0.165	0.256	0.222	-0.047	0.000
	标量_topk		0.01	0.436	0.231	0.219	-0.040	0.000
	不受约束	_rl	0.05	0.374	0.755	0.985	0.716	0.000
	受到激励		0.05	0.372	0.744	0.950	0.681	0.188
QBR 分析	静态		0.01	0.239	0.226	0.217	-0.024	0.000
	标量_topk		0.01	0.288	0.223	0.215	-0.016	0.000
	无约束	_rl	0.05	0.283	0.433	0.442	0.201	0.000
	激励		0.05	0.283	0.433	0.442	0.201	0.000

在所有场景中，集中度在学习过程中上升，但仍严格低于1。强化选择集中度（RSC）保持为正，证实了自适应改进，但最终集中度值始终低于无约束_强化学习中的值。轨迹中观察到的视觉平台现象在数值上体现为有界SC_T 值。

第三，学习率起到缩放因子的作用。提高学习率会同时加快无约束模式和激励模式下的强化速度。然而，只有无约束强化会坍缩为单一文化。在激励治理中，投影会吸收增加的梯度压力，将更高的学习率转化为更快（但仍有界）的适应。

第四，治理债务将约束-执行操作化。当强化推向被禁止的集中区域时，会发生投影裁剪，从而产生可衡量的治理债务。这一效应在学习率较高的场景中可见，并证实约束执行是主动的而非被动的。

最后，确定性聚合（标量_topk）表明，即使没有学习，结构上也可能出现支配地位。标量_topk与激励模式之间的对比凸显了治理必须同时规范结构性聚合规则和强化动态。

综合来看，这些图展示了定性动态，而表格则量化了其幅度。实证证据始终支持本文的核心主张：强化可以推动自适应选择改进，但只有

基于投影的主权约束才能防止决策权不可逆地集中。

VII. 讨论

实验结果证明，自适应强化与有界选择权威可以在单一架构框架内共存。本节结合第三节所引入的形式模型来解读这些实证发现。

A. 强化而不坍缩

将激励引入选择治理的一个核心风险在于，强化可能重新引入静态约束旨在防止的那种权力集中。无约束强化学习的结果清楚地展示了这种失效_模式：在迭代奖励反馈下，选择概率收敛于确定性支配，尤其是在较高学习率时。

相比之下，激励式治理在持续强化的同时，仍保持在完全集中之下的有界范围内。这种行为反映了向主权约束集进行投影的作用。强化更新提高了对高性能智能体的偏好，但投影防止了不可逆坍缩。

这一实证分歧支持《选择即权力 2.0》的核心论点：

> 只要选择更新受到外部强制实施的主权投影约束，强化就能提升性能而不消除结构多样性。

B. 学习率作为控制参数

学习率扫描揭示了系统的一项结构特性：学习率控制的是强化速度，而不仅仅是奖励累积。

较高的学习率会加速无约束模式和激励模式下优势地位的形成。然而，两者在定性上的差异依然存在：无约束强化会收敛至单一文化，而激励式强化则会稳定在有限的优势水平。

这表明，主权投影将学习率效应从灾难性崩溃转变为受控适应。

C. 结构性优势与习得性优势

标量_topk基线表明，确定性主导并不需要学习。去除多样性与随机选择就足以产生结构性单一化。

这一区分很重要：

- 不受约束的_强化学习会产生习得的 </style支配。
- 标量_topk 会造成结构性 </style支配。
- 激励式治理通过维护多样性约束来缓解这两种问题。

因此，治理不仅必须规范强化动态，还必须规范结构聚合规则。

D. 治理债务与投影边界

当强化过程触及集中度限制时，基于投影的更新会引入治理债务。实证结果表明，较高的学习率会增加激励模式下的治理债务。

这反映了一种固有的权衡：

- 更强的激励会加速性能提升。
- 更强的激励会加大对主权约束的压力。

因此，治理债务充当了性能优化与结构公平之间张力的操作性信号。

重要的是，投影可防止无声漂移。系统不会逐渐崩溃，而是在约束被违反时明显截断更新。这与最初的 Selection as Power 框架中提出的 fail-loud 原则相一致。

E. 动态治理作为一种受约束的动力系统

从系统视角来看，受激励的选择治理表现为一个约束动力系统：

- 奖励信号引入定向漂移。
- 投影引入反射边界。
- 多样性约束维持结构异质性。

经验轨迹展示了这种交互：强化会提高选择概率，投影则使其在约束

边界处反射，而多样性则防止分布坍缩为退化分布。

这一解释将实验结果直接与第三节中的形式模型联系起来，在该模型中，更新会被投影到可行集 $\mathcal{C}$ 和 $\mathcal{G}$上。

F. 当前实验的局限性

实验使用了完美评估器和合成奖励结构。虽然这足以展示强化动态和约束效应，但无法体现评估器噪声、延迟奖励或对抗性反馈。

此外，选择集中度是在分布层面（SC_t）和实际频率层面（头部智能体占比）进行测量的，但并未评估长期公平性或遗憾指标。

这些局限性界定了本文结论的适用范围：结果证明了受控强化下的可行性和有界性，但并未证明其在所有现实条件下都具备普适的稳健性。

G. 总结

动态实验支持三项主要结论：

1）不受约束的强化会导致确定性支配。2）基于投影的主权约束能够限制迭代更新过程中的集中化。3）激励式治理能够在保留结构多样性的同时实现自适应改进。

综上，这些发现表明，选择治理不必是静态的。只要强化严格在外部强制执行的约束集内运行，就可以将激励融入架构，而不必放弃有界权力。

VIII. 局限性与未来工作

尽管所提出的激励治理框架在受控条件下展现了有界自适应强化，若干局限性仍制约了当前结果的范围与可推广性。

A. 完美评估器假设

实验假定存在一个完美评估器，用于分配确定性正确性信号。在实践中，评估器可能表现出噪声、偏差或延迟反馈。在不完美评估下，强化更新可能会放大错位信号，从而可能提高对系统性误评估智能体的集中程度。

未来工作应纳入随机奖励模型，并分析评估器噪声下的稳定性。特别是，有必要研究当奖励信号不完美或具有对抗性时，投影约束是否仍然充分。

B. 合成奖励结构

实验中使用的奖励函数是基于特征的合成函数。虽然这足以展现强化和投影动态，但无法体现现实世界奖励景观的复杂性。

在实际部署中，奖励信号可能取决于长期表现、外部监管审计或潜在质量指标。将激励式治理扩展到延迟且部分可观测的奖励环境，仍是一个尚待解决的问题。

C. 约束投影的可扩展性

在当前实验配置中，投影到主权约束集 $\mathcal{C}$ 和$\mathcal{G}$ 的计算开销很小。然而，随着候选智能体数量或约束维度的增加，投影可能会变得更加复杂。

未来的工作应探索：

- 适用于高维参数空间的高效投影算法，
- 自适应约束收紧机制，
- 大规模系统下投影稳定性的理论保证。

D. 学习率敏感性

实验表明，学习率会显著影响强化速度和治理债务。然而，研究并未开展系统的超参数敏感性分析。

尚待探讨的问题包括：

- 自适应学习率调度是否能提高稳定性,
- 投影阈值是否应随学习率调整,
- 在长时间跨度训练下，治理债务会如何变化。

理解这些相互作用对于实际部署至关重要。

E. 公平性与长时间跨度多样性

本次评估衡量了选择集中度和实际支配性，但未量化较长时间跨度内的公平性。即使集中度受到限制，也可能导致某些智能体系统性曝光不足。

未来的工作应纳入：

- 基于遗憾的多样性指标,
- 长期曝光公平性约束,
- 反事实公平性评估。

F. 理论保证

本文的数学分析证明了投影下的有界性，但未能为多智能体动态下的完全收敛性保证或均衡刻画提供充分结论。

仍有一些形式化问题尚待解决：

- 受约束的强化何时能够收敛?
- 投影是否会在某些条件下引发振荡或混沌行为?

将经验稳定性与形式化保证相结合，是未来研究的一个重要方向。

G. 部署考量

在现实世界的受监管系统中，治理机制必须与审计流水线、日志要求和机构监督相结合。当前框架未涵盖延迟、计算预算和跨系统协调等部署约束。

将激励式选择治理扩展到生产级环境，需要：

- 稳健的监控基础设施,
- 透明的治理报告,
- 约束阈值的制度校准。

H. 总结

所提出的架构表明，可以通过外部强制实施的主权投影来约束基于强化学习的适应。然而，当前评估是在受控假设和简化奖励结构下进行的。

未来工作应将该框架拓展至含噪评估器、对抗性共谋、长时间跨度的公平性约束，以及形式化收敛分析。在现实部署条件下明确激励型治理的局限，仍是一个开放且前景广阔的研究方向。

IX. 结论

本研究将“选择即权力”框架从静态治理拓展至强化机制下动态且受激励的选择。我们探讨的核心问题是：自适应奖励驱动更新能否与高风险智能体系统中的受限选择权限并存。

实证结果揭示了明确的结构性差异。无约束强化会导致确定性集中：反复出现的奖励信号会不断提升选择概率，直至单个智能体占据主导地位。即使没有学习，确定性聚合规则（scalar_topk）也会造成结构性单一化。相比之下，通过投影到外部强制执行的主权约束集来实施激励式治理，能够实现自适应强化，同时防止多样性不可逆地崩塌。

学习率控制强化的速度，而投影控制其幅度。较高的学习率会加速所有启用学习的架构中的偏好形成；然而，只有在缺少投影时，这才会导致确定性锁定效应。每个更新步骤都执行主权约束时，集中增长会保持在有界范围内，治理债务也反映的是主动的约束执行，而非无声漂移。

这些发现支持一个更广泛的主张：选择治理不必是静态的。激励机制可以整合进架构，而不放弃有界权力，前提是强化严格在外部定义的约束集内运行。

更一般而言，本研究重新界定了自主系统中激励与权力之间的关系。强化学习常被视为一种纯粹以性能为导向的机制。我们的结果表明，当嵌入主权保全架构时，强化可以提升性能，而不会重新引入决策权限的结构性集中。

未来的工作必须将这项分析扩展到含噪评估器和对抗性奖励信号。受约束强化下的形式化收敛保证仍有待研究。尽管如此，当前结果证实了自适应选择治理的可行性：即使在持续的强化压力下，也能保留有界权力。

在受监管的高风险环境中，治理不仅必须规范智能体可以选择什么，还必须规范强化如何随时间推移重塑其权限。选择即权力 2.0 为实现这一目标提供了有原则的推进步骤。

致谢

作者感谢Grupo Santander的AI实验室团队参与讨论并提供反馈，感谢审稿人提出建设性批评，帮助提升论文的技术准确性并坦诚说明其局限，也感谢匿名评估者。他们对单次实验设计和N=1 局限的担忧促使作者明确阐述方法论依据。

参考文献

[1] J. M. de la Chica Rodriguez 和 J. M. V. Díaz，《迈向选择即权力：界定自主智能体的决策权限》，2026。[2] P. Christiano、J. Leike、T. Brown、M. Martic、S. Legg 和 D. Amodei，《基于人类偏好的深度强化学习》，神经信息处理系统进展，2017。[3] L. Ouyang、J. Wu、X. Jiang、等，《通过人类反馈训练语言模型遵循指令》，神经信息处理系统进展，2022。[4] Y. Bai、S. Kadavath、S. Kundu、A. Askell、J. Kernion、A. Jones、A. Chen、A. Goldie、A. Mirhoseini、C. McKinnon、等，《宪法式人工智能：通过人工智能反馈实现无害性》，*arXiv* 预印本 *arXiv:2212.08073*，2022。[5] J. Achiam、D. Held、A. Tamar 和 P. Abbeel，《约束策略优化》，载于国际机器学习会议，第 22–31 页，PMLR，2017。[6] M. Alshiekh、R. Bloem、R. Ehlers、B. Könighofer、S. Niekum 和 U. Topcu，《通过屏蔽实现安全强化学习》，载于*AAAI* 人工智能会议论文集，第 32 卷，2018。

[7] J. A. Kroll，可问责算法。博士论文，普林斯顿大学，2015。[8] D. Hadfield-Menell 和 G. Hadfield，《不完全契约与人工智能对齐》，*arXiv* 预印本 *arXiv:1804.04268*，2018。[9] R. S. Sutton 和 A. G. Barto，强化学习：导论。麻省理工学院出版社，2018。[10] L. Buşoniu、R. Babuška 和 B. De Schutter，《多智能体强化学习综合综述》，*IEEE* 系统、人类与控制论汇刊，2008。[11] J. Foerster 等，《反事实多智能体策略梯度》，*AAAI*，2018。[12] K. Zhang、Z. Yang 和 T. Basar，《多智能体强化学习：理论与算法的选择性综述》，强化学习与控制手册，2021。[13] D. Fudenberg 和 D. Levine，博弈中的学习理论。麻省理工学院出版社，1998。[14] R. J. Williams，《用于连接主义强化学习的简单统计梯度跟随算法》，机器学习，1992。[15] E. Altman，约束马尔可夫决策过程。Routledge，2021。[16] M. J. Osborne 和 A. Rubinstein，博弈论教程。麻省理工学院出版社，1994。[17] E. Friedman、P. Resnick、R. Sami、N. Nisan、T. Roughgarden、E. Tardos 和 V. Vazirani，《算法博弈论》，载于抗操纵声誉系统，第 677–697 页，剑桥大学出版社（剑桥），2007。[18] S. Boyd 和 L. Vandenberghe，凸优化。剑桥大学出版社，2004。