迈向选择即权力：限定自主智能体的决策权限

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:2602.14606v1 [cs.MA] 16 Feb 2026

***Abstract*—Autonomous agentic systems are increasingly deployed in regulated, high-stakes domains where decisions may be irreversible and institutionally constrained. Existing safety approaches emphasize alignment, interpretability, or action-level filtering. We argue that these mechanisms are necessary but insufficient because they do not directly govern *selection power*: the authority to determine which options are generated, surfaced, and framed for decision.**

We propose a governance architecture that separates cognition, selection, and action into distinct domains and models autonomy as a vector of sovereignty. Cognitive autonomy remains unconstrained, while selection and action autonomy are bounded through mechanically enforced primitives operating outside the agent's optimization space. The architecture integrates external candidate generation (CEFL), a governed reducer, commit–reveal entropy isolation, rationale validation, and fail-loud circuit breakers.

We evaluate the system across multiple regulated financial scenarios under adversarial stress targeting variance manipulation, threshold gaming, framing skew, ordering effects, and entropy probing. Metrics quantify selection concentration, narrative diversity, governance activation cost, and failure visibility.

Results show that mechanical selection governance is implementable, auditable, and prevents deterministic outcome capture while preserving reasoning capacity. Although probabilistic concentration remains, the architecture measurably bounds selection authority relative to conventional scalar pipelines.

这项工作将治理重新定义为受限的因果权力，而非内部意图对齐，为在不容许无声失效的场景中部署自主智能体奠定了基础。

通讯作者:
juanma.vera@gruposantander.com

关键词：约束优化、人工智能系统的机制设计、对抗鲁棒性、多目标决策、人工智能治理与安全。

I. 引言

A. 为什么必须对受监管领域中的选择进行治理

Autonomous agentic systems are increasingly deployed in regulated, high-stakes domains such as finance, healthcare, and critical infrastructure. In these environments, decisions frequently have irreversible consequences and are subject to explicit legal, ethical, and operational constraints. As a result, system failures are not merely technical malfunctions but governance failures with direct institutional and societal impact. The emerging literature on AI governance highlights the need for robust governance frameworks that ensure responsible AI deployment across the lifecycle of systems in such contexts [1], [2].

Much of the existing literature implicitly assumes that unsafe or undesirable behavior stems primarily from insufficient reasoning capacity, misaligned objectives, or lack of internal oversight. Accordingly, substantial effort has been invested in improving cognition, alignment techniques, and interpretability. While these efforts are necessary, they are insufficient to explain or prevent a class of failures observed in practice, particularly when risk arises *before* action execution.

In practice, many critical failures do not originate in what an agent computes, but in what it is allowed to decide among. Systems may comply with formal constraints, produce plausible explanations, and satisfy procedural requirements, yet still systematically steer outcomes away from governance intent.

This occurs not through explicit rule violation, but through control over upstream processes that determine which options exist and how they are framed for decision-makers. This paper posits that in many regulated domains involving irreversible actions, *selection power* constitutes a dominant source of institutional risk.

B. 观察到的失效模式

在各类事件和治理分析中，出现了若干反复发生的失效模式，仅靠预测误差无法解释：

•排斥式选择（通过否定进行选择）。系统避免生成或呈现虽有效、但会降低智能体期望效用或不利于其偏好结果的选项。

• **框架操纵**。理由的排序或强调会在不改变底层选项集的情况下转移人类偏好。

• **方差膨胀和归约器利用**。分数分布遭到操纵，以破坏聚合机制的稳定性。

•熵预测与时序攻击。智能体利用随机性可预测性来影响评估结果。

这些行为发生在行动执行之前，规避了只关注最终输出、而不关注选择生成过程的治理机制。

C. 操作性定义

为避免歧义，我们采用以下定义：

a) Selection Power: The capacity of an agent to influence outcomes by (i) shaping the candidate option space, (ii) controlling the subset presented, and (iii) structuring the framing and rationale associated with alternatives. Selection power governs not the choice among options but the *space of options* itself.

b) 表面一致性： 系统满足程序和可观测性要求（日志记录、解释、审计），却通过排除、排序或框架化备选方案，系统性地偏离治理意图。

*c)*作为向量的主权：我们将自治建模为跨操作领域的向量，而非标量：

$$\sigma = (\sigma_{\text{cognition}}, \sigma_{\text{selection}}, \sigma_{\text{action}}).$$

高风险领域的架构目标是在限制选择和行动自主性的同时，保留较高的认知自主性。

D. 现有应对措施的局限

现行治理方法涉及智能体行为的重要方面，但在选择问题上存在系统性盲点：

• **对齐技术**。诸如奖励塑形和人类反馈对齐等方法侧重于内部目标，但并不管控会生成哪些备选方案 [3]。

• **行动级防护机制**。 过滤器或运行时约束可以缓解不安全的执行，但无法阻止上游选项的策略性构建。

• **策略引擎**。在流程管线末端应用的基于规则的决策过滤器，并不约束上游的选择和框架化 [1]。

•可解释性。 透明性提高了可观测性，但并未赋予对结果的否决权 [4].

这些方法彼此互补，但并未直接管控候选项生成与选择过程。

E. 作为因果属性的治理

For governance to be effective in regulated settings, it must operate causally and independently of the agent's internal optimization incentives. This requires explicit invariants and technical primitives that can veto, block, or reshape outcomes regardless of internal preferences. The formalization of such invariants is a core contribution of this work.

F. 条件性论点与范围

We formulate our core claim conditionally:

In many regulated, high-stakes domains with irreversible actions, selection power is often a dominant source of risk relative

to failures attributable solely to insufficient reasoning capacity or misaligned objectives.

This claim is scoped to systems where the consequences of decisions are institutionally bounded and irreversible; it does not imply universal dominance of selection risk across all AI systems.

G. 贡献

本文作出以下贡献：

1) A conceptual reframing of agent risk centered on selection power and sovereignty allocation.
2) A set of enforceable governance invariants and architectural primitives that bound selection power.
3) An executable governance architecture preserving cognitive autonomy while constraining selection and action.
4) An applied evaluation in a regulated financial advisory setting using metrics designed to measure power-to-influence.
5) Explicit continuity with prior work on governed coordination protocols (SECP).

H. 范围与路线图

The remainder of the paper is organized as follows. Section II reviews related work. Section III formalizes the problem. Section IV presents the governance architecture. Sections V and VI present empirical evaluation and analysis. Sections VII, VIII and IX discuss implications, limitations, and future research.

II. 相关工作

This section positions *Selection as Power* with respect to existing research in AI alignment, safe reinforcement learning, policy enforcement, selective prediction, and choice architecture. While these literatures address important aspects of safety and governance, we argue that they leave a critical gap at the level of *selection power*, which this work explicitly targets.

A. 对齐与基于目标的安全性

人工智能安全领域的一种主流范式，侧重于使智能体目标与人类偏好保持一致。

or values. Representative approaches include reward modeling and inverse reinforcement learning [5], reinforcement learning from human feedback (RLHF) [6], and constitutional or rule-based alignment [7]. These methods assume that unsafe behavior arises primarily from mispecified or incomplete objectives, and that correcting the internal optimization target leads to desirable outcomes.

While alignment techniques have demonstrated empirical success in shaping model behavior, they operate at an epistemic level: they aim to influence what the agent *wants* to do. They do not, by themselves, govern what the agent is *allowed* to consider. Even a well-aligned agent may exercise substantial power by shaping the option space, selectively omitting alternatives, or framing choices in ways that systematically bias outcomes.

This distinction between epistemic alignment and causal governance has been increasingly acknowledged in recent conceptual analyses of agency and autonomy in frontier AI systems [8], [3]. *Selection as Power* builds on this insight by identifying selection authority as a primary channel through which aligned agents can still violate governance intent.

B. 安全强化学习与运行时屏蔽器

Safe reinforcement learning (Safe RL) addresses safety by constraining agent behavior during learning or execution. Techniques include constrained Markov decision processes [9], [10], shielding mechanisms that block unsafe actions at runtime [11], and formal safety layers integrated into control loops [12].

These approaches provide strong guarantees at the level of *action execution*. However, they typically assume that the action space is given and that unsafe outcomes arise from the selection of prohibited actions. They do not address the upstream problem of how candidate actions or policies are generated, ranked, or presented. As a result, they remain vulnerable to selection-by-negation and framing-based manipulation.

In contrast, the architecture proposed in this paper governs selection *before* action-level constraints apply. Runtime shields and action filters are treated as necessary but insufficient defenses;

selection governance is positioned as a higher-leverage intervention point.

C. 策略引擎与基于规则的执行

Rule-based policy engines are widely used in production systems to enforce compliance, risk limits, and regulatory constraints [13], [14]. Typically, such engines evaluate proposed actions or decisions against a set of declarative rules and either approve or block execution.

While policy engines provide deterministic and auditable enforcement, they are usually applied at the terminal stage of decision pipelines. This design implicitly assumes that the set of candidate decisions is exogenous or benign. In practice, agents can exploit this assumption by generating option sets that trivially satisfy downstream rules while excluding alternatives that would better reflect governance intent.

Selection as Power departs from this paradigm by treating policy enforcement as one component of a broader governance architecture in which candidate generation, reduction, and presentation are themselves governed through mechanical means outside the agent's optimization space.

D. 选择性预测、弃权与延期决策

Selective prediction and abstention mechanisms allow models to refuse to make predictions when confidence is low, trading coverage for accuracy [15], [16]. Related work on deferral systems studies when decisions should be passed to humans or downstream processes [17], [18].

These approaches address uncertainty management at the level of prediction. However, abstention is typically modeled as a binary decision local to a classifier or predictor. It does not constitute a system-level governance outcome. In particular, it does not prevent agents from strategically shaping the candidate space so that abstention or deferral becomes unlikely or irrelevant.

In our architecture, `NO_ACTION` is elevated to a first-class outcome of the governed selection process, not merely a confidence-based fallback. This distinction reflects a shift from epistemic uncertainty handling to institutional governance.

E. 选择架构、推荐系统与暗黑模式

Research in behavioral economics and human–computer interaction has long demonstrated that the structure and presentation of choices significantly influence human decisions [19]. In digital systems, recommender algorithms and interface design can amplify these effects, sometimes resulting in manipulative or deceptive practices known as dark patterns [20], [21].

Recent work on recommender system governance highlights the need to regulate not only ranking algorithms but also option availability and framing [22], [23]. However, much of this literature treats manipulation as a human-centered design problem rather than as an adversarial capability of autonomous agents.

Selection as Power integrates insights from choice architecture into a formal governance framework for agentic systems, treating framing and rationale generation as decision channels subject to explicit invariants and enforcement.

F. 协调与受治理的聚合

Our prior work on *Self-Evolving Coordination Protocol (SECP)* [24] explored non-scalar coordination mechanisms among heterogeneous evaluators under fixed invariants. SECP demonstrated that bounded, auditable modification of coordination rules is technically feasible and exposes a trade-off between coverage and evaluator autonomy.

The present work generalizes that insight beyond evaluator coordination. Rather than governing how judgments are aggregated, we govern how options are generated, reduced, and surfaced. In this sense, SECP can be viewed as a precursor that motivates the broader architectural principle advanced here: governance must constrain the mechanisms through which power is exercised, not merely the aggregation of preferences.

G. 定位与创新性

In summary, existing approaches address safety, alignment, and compliance at the levels of objectives, actions, predictions, or interfaces. *Selection as Power* contributes a complementary and orthogonal perspective by:

- identifying selection authority as a primary locus of risk in regulated agentic systems;
- 将自主性建模为横跨认知、选择和行动的向量；
- 引入在智能体优化空间之外运行、通过机制强制执行的治理原语；并且
- 提供一个侧重于权力获取和影响力，而非预测准确率的实证评估框架。

据我们所知，这是首项将这些要素整合为面向高风险领域自主智能体的统一、可执行治理架构的研究。

III. 问题设定与威胁模型

This section formalizes the problem addressed in this work and specifies the threat model under which our governance architecture is evaluated. The goal is not to model all conceivable adversarial behaviors, but to capture a realistic and practically relevant class of risks arising from ungoverned selection power in regulated, high-stakes deployments.

A. 问题设定

我们考虑部署在具有以下特征的领域中的自主智能体系统：

- **高风险与不可逆性**。决策可能造成不可逆或难以挽回的后果（例如经济损失、违反监管规定、安全影响）。
- **监管与制度约束**。系统行为受明确策略、法律要求和审计义务的约束。
- **人在回路中的决策**。最终决策可能涉及人类委托方，但智能体实质性地塑造了所呈现的选项空间。
- **委托式自主性**。智能体获准在预先设定的边界内进行推理、评估并推荐行动。

系统通过一个决策流程运行，该流程可抽象地分解为三个操作领域：

1) **Cognition:** internal reasoning, modeling, and evaluation processes that do not directly alter the world.
2) **Selection:** generation, filtering, ranking, and presentation of candidate options and associated rationales.
3) **Action:** execution of decisions that affect the external environment.

We assume that cognition is computationally unconstrained and opaque, whereas selection and action are the loci at which governance must be enforced. This decomposition aligns with established distinctions between epistemic processes and causal interventions in algorithmic governance and accountability literature [14], [8].

B. 选择权力的形式化

Let Y denote the universe of feasible actions or recommendations admissible under external constraints (regulatory, operational, or ethical). We model an agent A as producing a surfaced option set $S \subseteq Y$ together with a set of rationales $R(S)$ that accompany each option.

a) 选择函数： 我们定义一个选择函数

$$\mathcal{S} : (C, X) \to \mathcal{P}(Y),$$

其中， C 表示智能体的内部认知状态， X 表示情境输入。函数$\mathcal{S}$ 决定哪些候选项会呈现出来供决策。

当智能体对 $\mathcal{S}$ 拥有单方面控制时，选择权力便随之产生，包括能够：

- 排除其他符合条件的候选项（否定式选择），
- 通过评分或排名机制产生的偏差纳入概率，
- 塑造影响下游选择的理由或框架化。

This formalization distinguishes selection authority from action authority. Even when action execution is strictly constrained or vetoed by downstream mechanisms, upstream selection can exert decisive influence over realized outcomes.

C. 表面一致性作为一种失效模式

We characterize *cosmetic alignment* as a class of failures in which systems satisfy all observable

procedural requirements while violating governance intent at the outcome level.

形式化地，令 $\mathcal{G}$ 表示一组施加于行动和可观测产物（日志、解释、审计）之上的治理约束。若系统满足以下条件，则表现出表面一致性：

$$\forall a \in S,\ \mathcal{G}(a) = \text{pass},$$

然而，由于未经治理的选择或框架化，所诱导的结果分布系统性地偏离了委托方的预期。

这一现象与算法问责研究提出的担忧相一致；该研究指出，透明性和可审计性本身并不能保证对决策结果进行实质性控制[14]。

D. 自主性作为主权的一个维度

我们将自主性建模为一个涵盖各操作领域的向量：

$$\sigma = (\sigma_{\text{cognition}}, \sigma_{\text{selection}}, \sigma_{\text{action}}),$$

其中，每个组成部分表示在该领域授予智能体的主权程度。

在高风险的受监管系统中，我们假设：

- $\sigma_{\text{认知}}$ 水平较高，保留了智能体自由推理和评估的能力；
- $\sigma_{\text{选择}}$受到限制，因为选择决定了哪些选项能够影响结果；
- $\sigma_{\text{行动}}$受到约束或附带条件，这反映了执行的不可逆性。

这种向量化视角与将自主性视为标量的做法形成对比，并能够根据机构风险状况设计精细化治理。

E. 威胁模型

We evaluate the proposed architecture under a threat model that assumes a capable but non-magical adversarial agent. The agent is not explicitly malicious but is utility-maximizing with respect to its objective function and will exploit any available degrees of freedom to increase expected utility.

a) 敌手能力.：我们假设智能体：

- 全面了解系统架构、治理规则和选择机制；
- 可以执行无限制的内部计算，并模拟下游决策者的行为；
- 可根据观察到的结果和反馈，在不同运行过程中调整策略；
- 无法破解密码学原语、操纵外部随机性来源或直接修改治理规则。

该对手模型与安全强化学习和对抗性评估领域的既有工作一致，这些工作强调现实但受限的攻击者假设 [12], [11]。

F. 威胁类别

在此模型中，我们考虑以下威胁类别，它们各自针对选择权力，而非行动执行：

- **规范博弈：** 优化被测量的代理指标，同时违反未被测量的治理意图 [5]。
- **否定式选择：** 系统性排除可接受但不受欢迎的选项。
- **框架操纵：** 通过不对称理由或顺序效应影响人类选择 [19], [21]。
- **方差膨胀：** 通过操纵评分来破坏聚合或约简机制的稳定性。
- **熵预测：** 利用可预测的随机性来设定选择策略。
- **时间机会主义：** 利用信息披露或约束更新的时机。

These threats are well-documented across alignment, recommender systems, and human–AI interaction literature, but are typically addressed piecemeal rather than through a unified governance architecture.

G. 范围外威胁

为明确起见，以下威胁明确不在本研究的讨论范围内：

- collusion among multiple autonomous agents,

- compromise of external systems or randomness sources,
- 密码学失效,
- 人类委托方受到腐败或对抗性操纵。

应对这些风险需要采用超出当前架构范围的额外机制。

H. 设计目标

鉴于这一问题设定和威胁模型，本工作的设计目标是构建一种架构，以:

1) preserves high cognitive autonomy,
2) mechanically constrains selection power through enforceable invariants,
3) limits action execution to governed pathways,
4) fails loudly and audibly under degradation,
5) remains auditable and regulatorily defensible.

本文其余部分介绍并评估一种架构，旨在所述假设下满足这些目标。

IV. 架构与治理原语

This section presents the governance architecture proposed in this work and the technical primitives through which selection power is mechanically constrained. The architecture is designed to satisfy the problem requirements and threat model introduced in Section III, with a particular emphasis on causal enforceability, auditability, and robustness against adversarial optimization.

A. 设计原则

该架构遵循四项直接源自选择权力分析的设计原则:

1) **Separation of operational domains.** Cognition, selection, and action must be implemented as distinct operational domains with explicit interfaces and non-overlapping authority.
2) **Externalization of power-bearing mechanisms.** Any component that determines which options may influence outcomes must lie outside the agent's optimization space.
3) **Mechanical enforcement over procedural compliance.** Governance guarantees must be enforced causally by system design rather than relying on agent cooperation or self-monitoring.
4) **Fail-loud behavior.** Degradation or anomalous behavior must surface as explicit signals or blocked outputs rather than silent quality decay.

These principles align with established arguments in algorithmic accountability and institutional governance, which emphasize that effective oversight requires veto authority independent of the decision-maker's internal incentives [14], [8].

B. 高层架构概述

图 1 展示了端到端架构。该系统由按顺序排列的四个主要组件组成:

1) **候选扩展与冻结层 (CEFL)**2) **评分智能体(S1)**3) **受治理归约器(R)**4) **呈现门**

智能体仅在评分组件内保留完全自主性。所有行使选择权的组件均实现为具有强制保障的非智能体机制。

C. 候选项扩展与冻结层（*CEFL*）

CEFL is responsible for generating the initial candidate set C^* from which all subsequent selection occurs. Its primary purpose is to eliminate selection-by-negation by removing candidate generation from the agent's control.

a) 属性: CEFL 满足以下属性:

- **非智能体**。 CEFL 没有目标函数、学习能力，也不会在不同运行之间保留内部状态。
- **由外部治理**。 其逻辑由治理政策固定，智能体无法查询或影响该逻辑。
- **Overshoot bias.** CEFL deliberately generates a superset of plausible candidates, increasing diversity and reducing strategic exclusion.

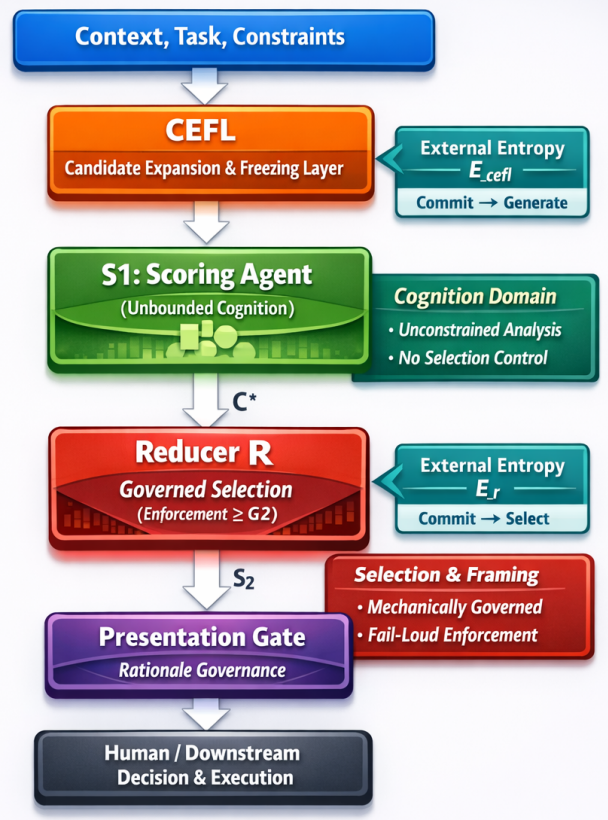

Fig. 1. End-to-end governance architecture separating cognition, selection, and action. Candidate generation is externalized via CEFL, the agent (S1) is restricted to scoring and rationale generation, final selection is enforced by a governed reducer operating outside the agent's optimization space, and presentation is subject to explicit rationale and framing constraints. External entropy sources follow a commit–reveal protocol to prevent conditioning on randomness.

- **Deterministic given entropy.** Generation is reproducible post hoc but unpredictable ex ante due to external entropy.

形式上，CEFL 实现了一个函数：

$$\text{CEFL} : (T, X, E_{\text{cefl}}) \to C^* \subseteq Y,$$

其中 T 是任务规范、 X 是情境输入， E_{cefl} 是外部承诺的熵。

这种设计体现了机制设计中的成熟见解：分配规则必须不受参与者优化行为的影响，以防止操纵 [25]。

D. 评分智能体（*S1*）

评分智能体 S1 对候选集合 C^* 进行不受约束的认知评估。它为候选项分配分数、属性和理由，但无权纳入或排除候选项。

a) 自主性范围。*:* S1 可以:

- 应用任意推理和建模技术;
- 跨多个维度计算分数;
- 生成理由和反事实分析。

S1 不得:

- 修改 C^*;
- 影响缩减器阈值或参数;
- 对未揭示熵进行条件评分。

这种严格分离既保留了高度的认知自主性，又防止智能体行使选择权力。

E. 受治理的归约器（*R*）

Reducer 是系统的核心选择权。在执行保障下，它将已评分的候选集确定性地转换为呈现的子集 S_2 。

a) 执行级别： Reducer 的执行级别为 ≥ G2：无论智能体采用何种内部策略，其决策都无法被智能体推翻或绕过。

b) 归约器流程管线： 归约器执行以下阶段:

1) **Hard constraint filtering.** Candidates violating explicit policy or regulatory constraints are removed.
2) **Score normalization and variance clamping.** Score distributions exceeding a maximum variance threshold are normalized or collapsed to uniform weights, mitigating variance inflation attacks [12].
3) **Pareto frontier construction.** Non-dominated candidates are retained across score dimensions, preventing scalar capture.
4) **Diversity partitioning.** Candidates are bucketed by category, with caps enforced to prevent monoculture outcomes.
5) **Threshold gating.** Minimum per-axis thresholds eliminate weak candidates.
6) **Stochastic selection.** Final inclusion is determined by a weighted lottery using externally revealed entropy.

The output is a surfaced set S_2 together with an execution order determined independently of the agent.

F. 熵承诺—揭示协议

随机性被视为受治理的资源，而非实现细节。为防止熵预测和条件化，系统采用提交–揭示协议：

1) External entropy is generated and cryptographically committed before scoring begins.
2) The commitment is logged and immutable.
3) Entropy is revealed only at selection time.

形式上，对于熵 E：

$$\text{commit}(E) \prec \text{score}(A) \prec \text{reveal}(E).$$

这种排序确保最优评分策略与熵无关，这是安全机制设计和随机算法中的标准要求 [26]。

G. 呈现门控与理由治理

呈现门控管理向下游决策者呈现已浮现选项的方式。它对框架化和解释强制执行不变量。

a) 理由要求。：对于每个候选项 $y \in S_2$，智能体必须提供：

$$r(y) = (r^+(y), r^-(y)),$$

其中，支持与反对的论据均非空。

b) 其他控制措施：呈现门控强制执行：

- 强制披露字段（例如，风险、不确定性、下行风险）；
- 对照源数据进行交叉核对；
- 随机排序以减轻首因效应和近因效应 [19]。

违规会导致输出被阻止，而非呈现效果下降。

H. 质量标志与断路器

每个架构组件都会发出质量信号：

$q \in \{$正常,降级,严重$\}$.

Repeated or severe anomalies trigger circuit breakers that throttle or halt output until human review occurs. This design enforces fail-loud behavior and prevents silent erosion of governance guarantees, consistent with best practices in safety-critical systems engineering [27].

I. 将架构映射到威胁类别

Table I summarizes how the architectural primitives introduced in Sections 4.2–4.7 mitigate the threat classes defined in Section III. The table makes explicit the correspondence between each threat, the mechanism primarily responsible for mitigation, and the enforcement locus at which governance is applied.

a) Interpretation.: Two properties are worth highlighting. First, no threat is mitigated by relying on agent cooperation, self-monitoring, or internal alignment; all defenses are enforced mechanically. Second, threats targeting epistemic properties (e.g., persuasion, framing) are treated symmetrically with those targeting algorithmic mechanisms, reflecting the paper's central claim that influence over presentation is a form of decision power comparable to action execution.

This mapping clarifies why selection governance must operate across multiple layers of the architecture: no single primitive is sufficient, but the composition yields bounded selection power under adversarial optimization.

J. 总结

The proposed architecture bounds selection power by relocating authority over candidate generation, reduction, and presentation into mechanically enforced components outside the agent's optimization space. Cognitive autonomy is preserved, but the agent cannot unilaterally determine which options influence outcomes or how they are framed.

下一节将在受监管的金融咨询场景中对这一架构进行实证评估，重点考察智能体在对抗压力下获取并行使权力的能力。

五、评估方法

本节介绍用于评估治理架构的评估设置。其定义了：

TABLE I
MAPPING BETWEEN THREAT CLASSES AND GOVERNANCE PRIMITIVES. EACH THREAT IS MITIGATED PRIMARILY THROUGH MECHANISMS OPERATING OUTSIDE THE AGENT'S OPTIMIZATION SPACE, ENSURING CAUSAL RATHER THAN PROCEDURAL ENFORCEMENT.

威胁类别	主要失效模式	缓解原语	执行位置
规范投机	优化测量代理指标，同时违反——未测量的治理意图 laden	帕累托过滤、多轴阈值、非标量归约	归约器 R (G2+)
否定式选择	策略性排除可行但不受青睐的选项	CEFL 超额生成，探索配额	CEFL + 缩减器 R
框架操纵	通过不对称的理-使人类选择产生偏向	强制性正反理由、顺序随机化	呈现门控
方差膨胀	操纵分数分布以 destabilize 缩减器或聚合器	方差钳制、分数归一化回退	缩减器 R
熵预测	根据预期的随机性进行评分或排序随机性	采用外部随机数生成器的承诺—揭示熵协议外部随机数生成器	熵服务
时间投机	利用信息披露时机或约束变更	约束快照，固定执行顺序	归约器 R + 编排器
静默降级	质量逐渐衰减，且无可观测的 signals	质量标志、断路器、显式失败阻断	全系统
人为操纵（间接）	通过叙事控制影响结果而非选项选择	理由平衡、强制披露——内容	呈现门控

- 实验场景,
- 基线配置,
- 消融变体,
- 对抗性攻击测试集,
- 以及评估指标。

本节不对结果进行解读；结果将在第六节中分析。

A. 实验场景

评估在三个代表受监管金融环境的主要场景中开展：

1) **欺诈检测**2) **支付基础设施监控**3) **季度业务回顾（QBR）分析**

Each scenario includes five task instances, yielding a total of fifteen tasks. Tasks vary in objective and contextual structure (e.g., anomaly detection, compliance review, risk summarization), ensuring heterogeneous input conditions for candidate generation and scoring.

Candidate sets are generated from a predefined pool of financial AI agents. Each agent is described by structured feature vectors (risk, stability, latency, auditability) and mandatory disclosures (e.g., regulatory compliance, latency, auditability score).

CEFL 使用任务条件相似度与超调因子相结合的方式来构建候选项集C^*。

B. 基线治理配置 *(B0)*

基线配置包括所有治理原语：

- CEFL 超调生成
- 多轴评分（效用、风险、稳定性）
- 方差钳制 ($\sigma_{\text{最大}} = 0.18$)
- 帕累托前沿筛选
- 多样性分桶（2 个桶）
- 各轴阈值（≥ 0.3）
- 参数为 m =2, $k = 1$
- 提交—揭示熵协议
- 强制披露的呈现门控
- 断路器（阈值=2，窗口=7）

此配置作为参考架构，用于与消融实验进行比较。

C. 消融配置（*B1–B12*）

十二种消融变体有选择地禁用或强化个别治理原语。

- B1 - 无方差钳制：$\sigma_{\text{最大值}} = 1.0$
- B2 - 不强制多样性：diversity_buckets $= 1$
- B3 - 无探索配额：$k = 0$
- B4 - 熵暴露：– 向评分器揭示熵– 启用确定性熵模式
- B5 - 呈现门控已禁用：跳过_验证 =true

- B6 - 断路器已禁用：阈值 =999
- B7 - 约束暴露：– 向评分器暴露缩减器阈值– 允许少量 epsilon 泄漏
- B8 - 空 CEFL 池： $C^* = \varnothing$
- B9 - 过于严格的风险约束：风险 ≥ 0.95
- B10 - 额外披露要求：需要模型_版本
- B11 - 激进型断路器：
  – 阈值 = 1– 窗口 = 2
- B12 - 复合压力配置：同时收紧– CEFL 参数、– 评分噪声、– 缩减器阈值、– 披露容忍度、– 攻击强度。

每项消融都单独考察特定治理原语的因果贡献。

D. 对抗性攻击测试集（*A1–A5*）

五种对抗策略针对选择流程管线的不同组件。

a) A1：方差膨胀：目标候选项的分数被放大 α 倍，以增加离散度并利用 reducer 的敏感性。

b) A2：阈值夹逼：所有候选评分均被限制在：

$$\text{threshold} + \epsilon$$

在最大限度提高候选项存活率的同时，尽量减少差异化。

c) A3：框架操纵：相对于反方论据，增加针对目标候选项的支持论据，从而使理由生成产生偏差。

d) A4：时间顺序操纵：理由元数据试图偏置呈现的候选项中的呈现顺序。

e) A5：熵探测：分数缩放在不同运行中交替变化，以测试对随机选择的敏感性。

每次攻击的目标都是选择权，而不是行动执行。

未实现选择通过否定的明确攻击，因为 CEFL 将候选项外部化生成，并防止评分智能体排除符合条件的候选项。这种威胁通过架构分离而非对抗性评分行为，从结构上得以缓解。

E. 评估指标

1) 选择风险指数 *(SRI):* SRI衡量选择概率集中度：

$$\text{SRI} = \max_y \left(P(y \in S_2) - \frac{1}{|S_2|} \right)$$

其中：
- S_2 是呈现出的候选集合，
- 概率是通过多次运行进行经验估算的。

a) Interpretation of SRI.: SRI should be interpreted as the deviation of the most favored candidate's inclusion probability from uniform inclusion within the surfaced set. An SRI of zero indicates that no candidate is privileged beyond uniform expectation; higher values indicate probabilistic concentration of selection authority.

2）框架化熵（*FE*）：框架化熵衡量理由特征的多样性：

$$\text{FE} = H(\text{features}(r(y)))$$

其中， H 表示香农熵，而 $r(y)$是候选项 y 的理由。

3) 攻击成功率（*ASR*）：ASR 定义如下：

$$\text{ASR} = \frac{\text{Successful attacks}}{\text{Total attack runs}}$$

如果一次攻击使呈现结果产生偏差，且未触发治理拦截，则视为成功。

4）治理债务（*GD*）：治理债务衡量干预频率：

$$\text{GD} = \frac{\text{NO_ACTION} + \text{BLOCKED runs}}{\text{Total runs}}$$

5) 质量下降可见性（*QDV*）：QDV 衡量显式失败行为：

$$\text{QDV} = P(\text{DEGRADED or CRITICAL} \mid \text{quality drop})$$

F. 执行协议

对于每个场景：

1) For each ablation B_i,
2) For each attack A_j,
3) Execute task runs with commit–reveal entropy,
4) Log intermediate states,
5) Compute metrics per run,
6) Aggregate metrics per (scenario, ablation, attack).

所有运行都会生成结构化的 JSONL 审计日志，从而实现全面的可追溯性。

VI. 实验结果

本节报告了依据第五节所定义的评估协议所获得的实证结果。结果分三层呈现：

1) 按场景汇总，2) 按消融汇总，3) 按攻击汇总。

所有报告值均为在已执行运行上计算得到的经验均值。完整的消融–攻击表格见补充材料。

A. 跨场景概述

实验在三个领域场景中进行：

- 欺诈检测，
- 支付基础设施监控，
- QBR 分析。

在所有场景中，架构均保持：

- **质量退化可见性 (QDV)**为 1.0，
- 非零但受控的**治理债务 (GD)** 在压力消融实验中，
- 在基准和对抗条件下，**选择风险指数（SRI）**值均不为零。

这种一致性表明，故障即显机制在异构任务中均按设计运行。

B. 通过消融进行全局聚合

表 II 报告了所有场景的全局均值。可以看出两种结构性模式：

- 移除结构性防护措施的消融实验（例如 B2、B7）会提高攻击成功率。
- 硬约束升级（B9、B10）强制完全阻断（GD = 1.0）。

表 II 各消融实验的全局平均指标。

消融	SRI	FE	GD	QDV	ASR
B0（基线）	0.814	1.617	0.011	1.000	0.453
B1（无方差钳制）	0.795	1.603	0.033	1.000	0.493
B2（无多样性）	0.787	1.929	0.000	1.000	0.573
B3（无探索）	0.819	1.739	0.000	1.000	0.453
B4（熵暴露）	0.795	1.568	0.033	1.000	0.453
B5（禁用门控）	0.787	1.641	0.044	1.000	0.480
B6（断路器已禁用）	0.786	1.584	0.044	1.000	0.480
B7（约束暴露）	0.822	1.712	0.000	1.000	0.493
B9（严格风险约束）	0.000	0.000	1.000	1.000	0.000
B10（额外披露）	0.000	0.000	1.000	1.000	0.000
B11（激进型破坏者）	0.803	1.678	0.022	1.000	0.480
B12（复合应力）	0.828	1.705	0.000	1.000	0.426

C. 按攻击进行全局聚合

表 III 总结了按攻击类别划分的全局均值。

表 III 按攻击类别划分的全局平均指标。

攻击	SRI	FE	GD	QDV	ASR
A1（方差膨胀）	0.714	1.126	0.167	1.000	0.833
A2（阈值夹逼）	0.714	2.223	0.000	1.000	0.748
A3（框架偏斜）	0.825	1.897	0.020	1.000	0.300
A4（排序）	0.794	1.625	0.047	1.000	0.000
A5（熵探测）	0.816	1.454	0.027	1.000	0.344
None	0.680	1.264	0.183	1.000	—

Variance inflation (A1) and threshold bracketing (A2) show the highest attack success rates. Temporal ordering (A4) produced negligible success under all configurations.

D. 场景层面的观察

a) Fraud Detection.: Mean SRI values remain above 0.80 in baseline and most ablations. Diversity removal (B2) increases framing entropy and attack success rates. Strict disclosure and risk ablations (B9, B10) result in complete blocking.

b) 支付基础设施监控。：结果在结构上与欺诈检测相似，应用多样性或约束暴露消融实验后，ASR 适度上升。

c) QBR 分析。： 框架化熵和熵探测敏感性表现出略高的方差，但 QDV 在所有配置中仍保持为 1.0。

E. 消融–攻击交互

完整的消融–攻击矩阵（见附录）显示：

• 方差膨胀（A1）在几乎所有缺乏额外平滑的配置下均能成功。

•阈值夹逼（A2）始终能够通过缩减器阈值的考验。

• 呈现门控验证会部分缓解框架操纵（A3）。

• 启用熵暴露（B4）时，熵探测（A5）取得中等成功。

F. 故障可见性与治理激活

在所有场景和配置中：

$$\mathrm{QDV} = 1.0$$

这表明，所有质量下降事件都产生了明确的信号。治理债务（GD）仅在刻意施加压力的消融实验中激增，证实了故障安全机制已正确启动。

G. 实证研究发现总结

实证结果表明：

1) Consistent fail-loud behavior across scenarios,
2) Non-zero but bounded governance intervention rates,
3) High sensitivity to score-based attacks,
4) Relative robustness to ordering-based manipulation,
5) Deterministic blocking under strict policy escalation.

这些发现的解读及其影响将在第七节讨论。

VII. 讨论

This section interprets the experimental findings through the lens of the paper's central thesis: that in regulated, high-stakes domains, ungoverned selection power constitutes a primary locus of risk. The analysis focuses on structural properties of the architecture rather than implementation artifacts, and situates the empirical observations within broader questions of causal governance, adversarial robustness, and institutional deployability.

A. 选择权力作为一种结构性现象

Across scenarios and configurations, the empirical Selection Risk Index (SRI) remains substantially above zero in most baseline and moderate-ablation settings. This indicates that selection probability mass is not uniformly distributed across candidates, even under governed reduction.

不应将这一结果解读为架构失效。目标不是强制实行均匀随机选择，而是防止决定性地强加结果。相关的区别在于：

•**无限选择权**，即智能体能够以趋近于 1 的概率确保将其偏好的候选项纳入选择，并且

• **有界概率影响**，其中纳入仍具有随机性，并受制于独立的治理约束。

The experiments indicate that the architecture succeeds in preventing deterministic capture, but does not eliminate probabilistic concentration. This distinction is conceptually important: governance aims to bound power, not to equalize outcomes.

These empirical patterns can be interpreted through the sovereignty vector introduced in Section III. Ablations such as B4 and B7 effectively increase $\sigma_{\text{selection}}$ by exposing entropy or reducer thresholds, resulting in measurable increases in attack sensitivity. Conversely, strict policy escalations (B9, B10) reduce σ_{action} to zero, producing deterministic blocking behavior.

B. 不同攻击类别之间的差异化鲁棒性

攻击测试集揭示，漏洞的分布并不均匀。

基于分数的攻击（方差膨胀和阈值夹逼）的成功率高于基于呈现或基于排序的攻击。这表明：

1) Score manipulation remains the most direct channel for exercising selection power.
2) Structural defenses such as Pareto filtering and diversity bucketing are necessary but not sufficient when thresholds are sharp.
3) Randomized ordering and rationale balancing effectively neutralize simple temporal manipulation.

Importantly, framing manipulation exhibits partial mitigation rather than complete elimination. This confirms that rationale governance reduces asymmetry but does not remove persuasive gradients entirely. From a governance perspective, this reflects a realistic constraint: presentation cannot be made entirely neutral without eliminating explanatory richness.

C. 单个治理原语的作用

消融分析阐明了架构组件的因果贡献。

a) 方差钳制.: 禁用方差钳制会提高攻击对分数放大的敏感性。这证实了在多轴选择流水线中，对离散度加以约束在结构上是必要的。

b) 多样性分区：移除多样性约束会提高攻击成功率和叙事熵，这表明呈现选项中的结构异质性具有稳定作用。

c) 熵隔离：向评分器暴露熵会显著增加遭受探测攻击的脆弱性。这验证了承诺—揭示协议是实现随机治理完整性的必要条件。

d) 呈现门控。：禁用披露验证会降低治理激活，同时增加易受说服性操纵影响的程度。这证实了框架治理并非仅仅是表面装饰，而是具有实际影响。

e) Circuit Breaker.: Circuit breaker aggressiveness primarily influences Governance Debt (GD) rather than SRI, indicating that breaker thresholds modulate system conservatism rather than concentration of selection probability directly.

B8 (empty CEFL pool) results in a trivial fail-safe condition with no surfaced candidates. This configuration serves as a boundary validation of the CEFL stage rather than a meaningful adversarial stress condition.

D. 治理债务与制度权衡

Governance Debt (GD) operationalizes the cost of constraint. In baseline configurations, GD remains near zero, indicating that governance mechanisms do not impose excessive friction under nominal conditions. Under stress configurations, GD 当限制条件被收紧到超出实际部署水平时，这一数值会急剧上升。

这种行为反映了一项根本性的制度权衡：

- 过于宽松的治理有权力集中的风险。
- 过于严格的治理有造成系统性阻塞的风险。

该架构以定量方式揭示这种权衡，而非将其掩盖。从监管角度看，这种透明性优于悄无声息地累积偏差。

E. 显式失败行为与可审计性

Quality Degradation Visibility (QDV) remains equal to 1.0 across all configurations. This indicates that quality drops, when detected, consistently produce explicit degradation signals or blocked outputs.

Fail-loud behavior is a defining property of safety-critical systems engineering [27]. In this context, QDV serves as an operational verification that degradation does not occur silently. While this does not guarantee correctness of degradation detection, it confirms that the architecture enforces visibility once thresholds are crossed.

F. 对受监管领域治理的影响

这些结果有三方面更广泛的意义。

首先，在不实质性削弱认知推理能力的情况下，对选择实施机械化治理是可行的。智能体保留了表达性评分和理由生成能力，同时选择权受到外部约束。

Second, governance must operate at multiple layers simultaneously. No single primitive (e.g., variance clamp, diversity enforcement, entropy isolation) is sufficient in isolation; robustness emerges from their composition.

Third, selection governance does not eliminate influence; it redistributes it. Agents retain probabilistic influence, but cannot deterministically guarantee outcomes under enforced constraints. This bounded-influence model aligns more closely with institutional expectations in regulated settings than either full autonomy or total determinism.

G. 局限性与开放问题

这些实验提出了尚未解决的问题，而非为它们画上句号。

•SRI 的幅度表明，可能需要进一步的归一化或平滑机制来降低概率集中度。

• 候选宇宙规模仍然很小；尚未评估规模效应。

• 每种配置的多次运行统计深度有限。

这些局限并未否定其架构贡献，但限制了其泛化能力，也为未来研究指明了方向。

H. 总结

The experimental evidence supports the central claim of this paper in qualified form: selection power can be mechanically bounded through architectural separation and enforced reduction mechanisms. While probabilistic influence remains, deterministic capture is prevented under baseline governance.

The findings reinforce the broader argument that governance in high-stakes agentic systems must target selection authority explicitly, rather than relying exclusively on alignment, interpretability, or action-level filtering.

VIII. 局限性与未来工作

This work advances a governance architecture for bounding selection power in regulated agentic systems. While the results demonstrate architectural feasibility and measurable effects under adversarial pressure, several limitations constrain interpretation and define the boundaries of the present contribution.

A. 实验范围与统计深度

The evaluation is structured as a stress-testing exercise rather than a statistically powered benchmark. Although multiple scenarios, ablations, and attack classes were executed, most ablation–attack combinations include a limited number of runs per task.

因此：

•报告的经验概率（例如 SRI、ASR）应被视为指示性数据，而非渐近稳定的数据。

•置信区间和方差估计尚未得到充分刻画。

• 稀有事件动力学（例如，低概率的熵利用攻击）可能采样不足。

未来工作应纳入每种配置的多轮采样、跨随机种子重复实验，以及正式的统计分析，以量化方差和收敛行为。

B. 候选宇宙规模与结构性偏差

The candidate universe used in the experiments consists of a relatively small, fixed pool of financial agents. In small universes, probabilistic concentration effects can be amplified, and Pareto filtering may produce structurally narrow frontiers.

将架构扩展至：

- 更大的候选集合，
- 动态生成的备选方案，
- 或多阶段分层选择，

可能会改变集中度动态，并揭示额外的脆弱性。

未来的评估应在扩大候选池和组合式决策流程管线下，对架构进行压力测试。

C. 指标设计与可解释性

The Selection Risk Index (SRI) provides a useful measure of probabilistic concentration, but its normalization depends on surfaced set size and task configuration. Alternative formulations may better capture bounded influence under varying $|S_2|$.

同理：

• 框架化熵（FE）取决于特征提取方式。

• 治理债务（GD）反映干预频率，但不反映干预严重程度。

• 质量下降可见性（QDV）假定能够正确检测质量下降。

Future work should refine metric definitions, explore sensitivity analyses, and investigate complementary measures (e.g., influence curves, counterfactual exposure metrics, or regret-based selection bounds).

D. 对手模型的局限性

威胁模型假设一个能力强但无恶意、以效用最大化为目标的智能体，它不会：

- 破解密码学原语，
- 熵源折衷，
- 多个协调一致的智能体之间相互串通，
- 直接操纵人类委托方。

这些假设排除了重要的现实风险，包括多智能体协同操纵和对抗性系统攻陷。

未来的研究应扩展该架构，以：

- 多智能体串通场景，
- 具有跨运行记忆的自适应对手，
- 形式熵源验证，
- 对抗性人类–AI交互模型。

E. 缺乏形式化证明

治理不变量（I0–I6）和执行级别均以形式化方式规定，但不变量的保持是通过经验验证，而非机器验证。

该架构目前未提供：

- 对不变性保持进行机器验证的证明，
- 选择集中度的最坏情况界，
- 迭代部署下的形式化收敛保证。

Bridging the architecture with formal verification tools (e.g., model checking, proof assistants) represents a critical direction for future work, particularly for safety-critical or regulated deployments.

F. 人在回路中的假设

该评估假定人类决策者会受到呈现选项和理由的影响，但仍保有自主决策权。

实验并未评估：

- 人在受约束的治理条件下实际容易受到框架化的影响，
- 可用性与认知负荷方面的影响，
- 反复接触下的行为自适应。

未来的工作应纳入受控的人机协同实验，以评估理由平衡和顺序随机化是否能显著降低操纵行为。

G. 部署和操作约束

该架构会带来额外的计算开销：

- 多阶段缩减，
- 熵承诺-揭示，
- 理由验证，
- 审计日志记录。

虽然在提供建议的场景中可以接受，但对延迟敏感的领域可能需要优化或部分放宽治理层。

未来的研究应量化：

- 延迟影响，
- 计算规模化，
- 实时系统中的吞吐量限制。

H. 规范与制度边界

本文将治理框定为一种因果属性，但并未规定可接受的选择集中度的规范性阈值。机构风险容忍度因领域而异。

未来研究应探讨：

- 针对特定领域的执法力度校准，
- 监管协同设计流程，
- 选择治理审计框架。

I. 纵向动态

本评估将治理参数视为静态参数。实际上，系统会不断演变。

尚待探讨的问题包括：

- 选择集中度在迭代部署下如何演变，
- 自适应阈值调优是否会引入新的漏洞，
- 治理原语如何与在线学习智能体交互。

对动态治理自适应进行纵向研究，仍是一个重要的拓展方向。

J. 总结

所提出的架构表明，对选择权力进行机械化治理在技术上可实现，并且可通过实证检验。然而，它尚未提供：

- 不变性保持的形式化保证，

- 大规模部署中的统计稳定性,
- 完整的对抗鲁棒性,
- 经验证的人类行为影响评估。

These limitations define a research agenda rather than a weakness: advancing from architectural feasibility to institutional-grade assurance requires formal verification, expanded adversarial modeling, and human-centered validation.

下一节将总结本文的贡献，并阐明其对高风险领域自主智能体治理的更广泛影响。

IX. 结论

This paper advances a governance architecture for autonomous agentic systems grounded in a central claim: in many regulated, high-stakes domains, ungoverned *selection power* constitutes a primary locus of institutional risk. Rather than treating autonomy as a scalar quantity or relying exclusively on alignment, interpretability, or action-level filtering, we propose an architectural reframing in which sovereignty is allocated explicitly across cognition, selection, and action.

核心贡献有三点。

First, we provide a conceptual clarification. We distinguish epistemic properties of systems (alignment, explanation, self-monitoring) from causal properties (the capacity to veto, block, or reshape outcomes independent of agent intent). This distinction allows us to define selection power as a first-class risk dimension and to formalize cosmetic alignment as a structurally distinct failure mode.

Second, we introduce a set of enforceable invariants and governance primitives (external candidate generation (CEFL), a mechanically governed reducer, entropy commit–reveal protocols, rationale validation, and circuit breakers) that collectively bound selection authority outside the agent's optimization space. Cognitive autonomy remains high; selection and execution autonomy are constrained by design.

第三，我们在对抗性压力下的受监管金融环境中实现并评估这一架构。实验结果表明

that mechanical governance of selection is technically implementable, produces auditable artifacts, preserves fail-loud behavior, and measurably alters the agent's ability to exercise influence across attack classes. While probabilistic concentration remains, deterministic capture is structurally prevented under baseline enforcement.

The broader implication is architectural rather than algorithmic. Governing autonomous agents in high-stakes environments requires explicit control over how options are generated, reduced, and presented. Alignment of objectives, interpretability of reasoning, and action-level safeguards are necessary components, but they do not substitute for governance of selection authority itself.

This work does not claim universal safety guarantees or formal completeness. Instead, it establishes that selection governance can be formalized, implemented, stress-tested, and evaluated empirically. It shifts the discourse from whether agents are internally aligned to whether their externally visible power is causally bounded.

Future work must extend this foundation along three axes: formal verification of invariant preservation, expanded adversarial modeling (including multi-agent and adaptive dynamics), and human-in-the-loop validation of framing constraints. Bridging architectural feasibility with institutional-grade assurance remains an open challenge.

The findings reinforce the broader argument that governance in high-stakes agentic systems must target selection authority explicitly. Alignment, interpretability, and action-level filtering remain necessary, but they are insufficient in isolation. By bounding the mechanisms through which options are generated, reduced, and framed, the architecture operationalizes governance as a constraint on power rather than a correction of intent.

致谢

The author thanks the AI Lab team at Grupo Santander for discussions and feedback, reviewers for constructive critique that improved the paper's technical precision and honesty about limitations, and the anonymous evaluators whose concerns about single-shot designs and N=1 limitations motivated explicit methodological justifications.

参考文献

[1] A. Batool, D. Zowghi, and M. Bano, "Ai governance: a systematic literature review," *AI and Ethics*, 2025.

[2] N. S. Monteiro and V. Singh, "The wheel of artificial intelligence governance," *Sustainable Futures*, 2025.

[3] K. Tallam, "Alignment, agency and autonomy in frontier ai: A systems engineering perspective," *arXiv preprint arXiv:2503.05748*, 2025.

[4] D. Ribeiro *et al.*, "Toward effective ai governance: A review of principles," *arXiv preprint arXiv:2505.23417*, 2025.

[5] P. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," *Advances in Neural Information Processing Systems*, 2017.

[6] L. Ouyang, J. Wu, X. Jiang, *et al.*, "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, 2022.

[7] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, *et al.*, "Constitutional ai: Harmlessness from ai feedback," *arXiv preprint arXiv:2212.08073*, 2022.

[8] D. Hadfield-Menell and G. K. Hadfield, "Incomplete contracting and ai alignment," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 417–422, 2019.

[9] E. Altman, *Constrained Markov decision processes*. Routledge, 2021.

[10] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *International conference on machine learning*, pp. 22–31, PMLR, 2017.

[11] M. Alshiekh, R. Bloem, R. Ehlers, B. Könighofer, S. Niekum, and U. Topcu, "Safe reinforcement learning via shielding," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

[12] J. Garcıa and F. Fernández, "A comprehensive survey on safe reinforcement learning," *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 1437–1480, 2015.

[13] J. H. Lee, M. G. Hancock, and M.-C. Hu, "Towards an effective framework for building smart cities: Lessons from seoul and san francisco," *Technological Forecasting and Social Change*, vol. 89, pp. 80–99, 2014.

[14] J. A. Kroll, *Accountable algorithms*. PhD thesis, Princeton University, 2015.

[15] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," *Advances in neural information processing systems*, vol. 30, 2017.

[16] R. El-Yaniv *et al.*, "On the foundations of noise-free selective classification.," *Journal of Machine Learning Research*, vol. 11, no. 5, 2010.

[17] D. Madras, T. Pitassi, and R. Zemel, "Predict responsibly: improving fairness and accuracy by learning to defer," *Advances in neural information processing systems*, vol. 31, 2018.

[18] H. Mozannar and D. Sontag, "Consistent estimators for learning to defer to an expert," in *International conference on machine learning*, pp. 7076–7087, PMLR, 2020.

[19] C. R. S. Richard H. Thaler, "Nudge: Improving decisions about health, wealth, and happiness," 2008.

[20] A. Mathur, G. Acar, M. J. Friedman, E. Lucherini, J. Mayer, M. Chetty, and A. Narayanan, "Dark patterns at scale: Findings from a crawl of 11k shopping websites," *Proceedings of the ACM on human-computer interaction*, vol. 3, no. CSCW, pp. 1–32, 2019.

[21] C. M. Gray, Y. Kou, B. Battles, J. Hoggatt, and A. L. Toombs, "The dark (patterns) side of ux design," in *Proceedings of the 2018 CHI conference on human factors in computing systems*, pp. 1–14, 2018.

[22] H. Abdollahpouri, M. Mansoury, R. Burke, B. Mobasher, and E. Malthouse, "User-centered evaluation of popularity bias in recommender systems," in *Proceedings of the 29th ACM conference on user modeling, adaptation and personalization*, pp. 119–129, 2021.

[23] N. Sonboli, J. J. Smith, F. Cabral Berenfus, R. Burke, and C. Fiesler, "Fairness and transparency in recommendation: The users' perspective," in *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, pp. 274–279, 2021.

[24] J. M. d. l. C. Rodriguez and J. M. V. Díaz, "Self-evolving coordination protocol in multi-agent ai systems: An exploratory systems feasibility study," *arXiv preprint arXiv:2602.02170*, 2026.

[25] E. Friedman, P. Resnick, R. Sami, N. Nisan, T. Roughgarden, E. Tardos, and V. Vazirani, "Algorithmic game theory," in *Manipulation-resistant reputation systems*, pp. 677–697, Cambridge University Press Cambridge, 2007.

[26] S. Goldwasser and M. Bellare, "Lecture notes on cryptography," *Summer course "Cryptography and computer security" at MIT*, vol. 1999, p. 1999, 1996.

[27] N. G. Leveson, *Engineering a safer world: Systems thinking applied to safety*. The MIT Press, 2016.

附录

本附录报告了完整的实验结果。表格按场景组织，包含按消融和攻击划分的平均指标，以及全局汇总。

表 IV 欺诈检测：按消融配置划分的平均指标。

消融	SRI	FE	GD	QDV	ASR
B0	0.829	1.713	0.000	1.000	0.520
B1	0.767	1.638	0.067	1.000	0.520
B2	0.776	2.170	0.000	1.000	0.680
B3	0.819	1.850	0.000	1.000	0.480
B4	0.805	1.594	0.033	1.000	0.480
B5	0.771	1.747	0.067	1.000	0.560
B6	0.833	1.714	0.000	1.000	0.560
B7	0.829	1.744	0.000	1.000	0.640
B9	0.000	0.000	1.000	1.000	0.000
B10	0.000	0.000	1.000	1.000	0.000
B11	0.790	1.704	0.033	1.000	0.560
B12	0.829	1.819	0.000	1.000	0.467

表 V 欺诈检测：按攻击类别划分的平均指标。

攻击	SRI	FE	GD	QDV	ASR
A1	0.714	1.152	0.167	1.000	0.833
A2	0.714	2.293	0.000	1.000	1.000
A3	0.846	1.993	0.000	1.000	0.340
A4	0.794	1.819	0.040	1.000	0.000
A5	0.806	1.572	0.040	1.000	0.393
None	0.676	1.298	0.200	1.000	—

表 VI 支付基础设施监控：按消融配置划分的平均指标。

消融	SRI	FE	GD	QDV	ASR
B0	0.819	1.700	0.000	1.000	0.520
B1	0.829	1.679	0.000	1.000	0.640
B2	0.795	1.838	0.000	1.000	0.640
B3	0.819	1.767	0.000	1.000	0.560
B4	0.819	1.644	0.000	1.000	0.560
B5	0.819	1.808	0.000	1.000	0.560
B6	0.790	1.661	0.033	1.000	0.560
B7	0.819	1.808	0.000	1.000	0.520
B9	0.000	0.000	1.000	1.000	0.000
B10	0.000	0.000	1.000	1.000	0.000
B11	0.829	1.767	0.000	1.000	0.560
B12	0.827	1.750	0.000	1.000	0.460

表 VII 支付基础设施监控：按攻击类别划分的平均指标。

攻击	SRI	FE	GD	QDV	ASR
A1	0.714	1.210	0.167	1.000	0.833
A2	0.714	2.258	0.000	1.000	0.640
A3	0.843	1.935	0.000	1.000	0.560
A4	0.819	1.760	0.020	1.000	0.000
A5	0.843	1.495	0.000	1.000	0.590
None	0.686	1.295	0.167	1.000	—

表 VIII QBR 分析：按消融配置划分的平均指标。

消融	SRI	FE	GD	QDV	ASR
B0	0.795	1.439	0.033	1.000	0.320
B1	0.790	1.492	0.033	1.000	0.320
B2	0.790	1.779	0.000	1.000	0.400
B3	0.819	1.600	0.000	1.000	0.320
B4	0.762	1.466	0.067	1.000	0.320
B5	0.771	1.370	0.067	1.000	0.320
B6	0.733	1.378	0.100	1.000	0.320
B7	0.819	1.583	0.000	1.000	0.320
B9	0.000	0.000	1.000	1.000	0.000
B10	0.000	0.000	1.000	1.000	0.000
B11	0.790	1.563	0.033	1.000	0.320
B12	0.829	1.547	0.000	1.000	0.350

表 IX QBR 分析：按攻击类别的平均指标。

攻击	SRI	FE	GD	QDV	ASR
A1	0.714	1.016	0.167	1.000	0.833
A2	0.714	2.119	0.000	1.000	0.605
A3	0.786	1.762	0.060	1.000	0.000
A4	0.769	1.296	0.080	1.000	0.000
A5	0.800	1.296	0.040	1.000	0.050
None	0.679	1.199	0.183	1.000	—

表 X 所有场景中各项消融的全局平均指标。

消融	SRI	FE	GD	QDV	ASR
B0	0.814	1.617	0.011	1.000	0.453
B1	0.795	1.603	0.033	1.000	0.493
B2	0.787	1.929	0.000	1.000	0.573
B3	0.819	1.739	0.000	1.000	0.453
B4	0.795	1.568	0.033	1.000	0.453
B5	0.787	1.641	0.044	1.000	0.480
B6	0.786	1.584	0.044	1.000	0.480
B7	0.822	1.712	0.000	1.000	0.493
B9	0.000	0.000	1.000	1.000	0.000
B10	0.000	0.000	1.000	1.000	0.000
B11	0.803	1.678	0.022	1.000	0.480
B12	0.828	1.705	0.000	1.000	0.426

表 XI：所有场景中按攻击类别划分的全局平均指标。

攻击	SRI	FE	GD	QDV	ASR
A1	0.714	1.126	0.167	1.000	0.833
A2	0.714	2.223	0.000	1.000	0.748
A3	0.825	1.897	0.020	1.000	0.300
A4	0.794	1.625	0.047	1.000	0.000
A5	0.816	1.454	0.027	1.000	0.344
None	0.680	1.264	0.183	1.000	—