 DOI: https://doi.org/10.12345/bdai.v6i3.28235

Algorithm deterrence and man-machine game in AI era from the perspective of prospect theory

Lei Huang Zhe Cui Hao He Rui Song Lindong Yu

Shandong Qingzhou High-tech Research Institute, Qingzho, Shandong, 262500, China

Abstract

In view of the failure of the "rational actor hypothesis" of traditional deterrence theory in AI governance, this paper constructs a human-machine dynamic game model based on prospect theory and proposes a cognitive adaptive deterrence framework. Core contributions include: 1) Designing a two-modal deterrence mechanism to solve irrational decision bias through reference point anchoring and probability weight correction; 2) The human prospect utility function (loss aversion coefficient λ=2.18) and AI depth strategy gradient model (DDPG) were established to realize probabilistic correction and adaptive loss amplification; 3) The mixed experiment showed that the adaptive signal increased the violation suppression rate by 32% (OR=0.68), the dynamic strategy increased the compliance rate by 13% (p<0.01), and the strategy variance decreased by 42% (F=9.37). The results show that the algorithm deterrence needs to integrate the behavior parameter (probability distortion δ=0.76) rather than rely on the technical advantage. This study provides a theoretical framework for AI governance based on "psychological adaptation"

Keywords

algorithmic deterrence; Prospect theory; Man-machine game; Cognitive adaptation; Loss aversion

前景理论视域下 AI 时代算法威慑与人机博弈

黄雷 崔喆 贺浩 宋睿 余林栋

山东青州高新技术研究所，中国·山东 青州 262500

摘 要

针对传统威慑理论的“理性行为体假设”在AI治理中的失效问题，本文基于前景理论构建人机动态博弈模型，提出认知适配威慑框架。核心贡献包括：（1）设计双模态威慑机制，通过参考点锚定与概率权重修正解决非理性决策偏差；（2）建立人类前景效用函数（损失厌恶系数λ=2.18）与AI深度策略梯度模型（DDPG），实现概率纠偏与损失自适应放大；（3）混合实验显示：适配信号使违规抑制率提升32%（OR=0.68），动态策略使合规率提高13%（p<0.01），策略方差缩减42%（F=9.37）。结果表明，算法威慑需融合行为参数（概率扭曲δ=0.76）而非依赖技术优势。研究为AI治理提供基于“心理适配性”的理论框架。

关键词

算法威慑；前景理论；人机博弈；认知适配；损失厌恶

1 引言

随着人工智能技术的颠覆性发展，其在军事冲突、社会治理、商业竞争等场景中的战略性应用正在重新塑造传统威慑理论的实践范式。算法驱动的自动化决策系统不仅能够快速执行风险分析、行为预测与应对策略生成，更通过实时反馈机制形成了动态威慑能力。例如，社交平台利用自然语言处理算法监控虚假信息传播，并以即时封禁或限流机制威慑违规用户。这些案例表明，算法已突破传统威慑中“人类 - 人类”的交互框架，演化成为“人类 - 机器 - 环境”的复杂博弈格局。然而，现有威慑理论研究仍严重依赖“理性行为体假设”，即假定行为主体能够准确评估威慑成本与收益，并通过效用最大化原则进行决策。这一假设在人机博弈的异质性场景中面临着决策主体的非理性特征和算法威慑的动态矛盾性的双重挑战。

在此背景下，如何通过行为经济学的前景理论框架，重新构建人工智能时代的算法威慑逻辑，破解传统理论对人机博弈中非理性决策行为的解释困境？这一问题的解答可以推动威慑理论在数字时代的适应性重构，也为解决 AI 治理中的现实困境提供新视角。

【基金项目】山东省自然科学基金青年项目（项目编号：ZR2023QD087）。

【作者简介】黄雷（1984–），男，中国四川成都人，硕士，副教授，从事人工智能、算法博弈研究。

2 算法威慑的机制与场景

2.1 算法威慑的运作模式

人工智能驱动的威慑机制可划分为主动威慑与被动威慑两类[1]，二者共同构成了“预测 - 响应 - 反馈”的动态闭环系统。

2.1.1 主动威慑：AI 系统主导的威胁实施

主动威慑技术特征是基于机器学习与实时数据分析，AI 通过预测性威胁与预设规则执行主动施加威慑。该类型的行为影响机制主要是：①参考点操控：算法将用户的“合规状态”设为默认参考点，任何偏离此状态的行为（如发布违规内容）均被标记为“损失域事件”，触发用户损失厌恶反应。②确定性效应强化：通过即时反馈（如“违规后 5 分钟限权”）放大威慑信号的可信度，避免传统威慑中“承诺可信度不足”的缺陷。

2.1.2 被动威慑：人类预判驱动的行为矫正

被动威慑的技术特征是即使 AI 未直接发出威胁信号，人类也会基于对算法潜在反制能力的认知主动调整行为，形成前瞻性自我审查。其行为影响机制：

①心理模拟与损失预判：人类在决策前会模拟与 AI 互动的可能结果，对高阶概率事件（如“被算法标记为可疑交易”）赋予过度权重，从而选择保守策略。

②隐性威胁的认知启动：算法通过历史数据透明度（如公开部分处罚案例）激活用户的“可得性启发式”，使其高估当前行为的风险概率。

2.2 典型案例分析

2.2.1 案例 1：自动驾驶系统的风险预控威慑——以特斯拉的安全策略为例

威慑机制：

概率权重干预：特斯拉的 AI 会以“每千公里分心次数”为单位向驾驶员发送安全报告，将低概率风险（如 0.1% 的事故率）转换为高频事件（“每 10 万公里可能发生 100 次危险”），利用概率权重偏差加剧用户风险感知。

梯度惩罚设计：系统根据分心程度实施三级响应：

Level 1：仪表盘闪烁警告（轻度损失框架）；

Level 2：自动降低巡航速度（中度损失框架）；

Level 3：强制靠边停车并锁定人工驾驶权限（重度损失框架）。梯次威慑使 94% 的驾驶员在 Level 1 阶段即修正行为，避免触发更高惩罚。

威慑效果：NHTSA 数据显示，自动驾驶用户因分心引发的事故率比传统车辆低 53%[2]。

2.2.2 案例 2：金融风控系统的反欺诈博弈——以蚂蚁金服的动态评估模型为例

威慑机制：

参考点动态锚定：支付宝的 AI 根据用户历史交易模式生成个性化“安全基线”，若检测到异常（如深夜大额转账），则动态下调信用评分并冻结部分功能。研究发现，用户对模糊威胁的风险感知均值为 67%，显著高于实际风险率（12%）[3]，印证概率权重偏差的放大效应。

威慑效果：虚假交易拦截率提升至 99.8%，但部分用户因频繁验证产生“警报疲劳”，导致 3.2% 的客户关闭风控提醒，反映威慑强度的边际递减规律。

3 前景理论下的人机博弈模型

3.1 模型构建

基于前景理论的核心假设，本文提出一种非对称人机动态博弈框架，旨在刻画算法威慑场景中人类决策者的非理性行为特征与智能体策略的自适应优化过程。模型由以下核心组件构成。

3.1.1 定义 1：人类前景效用函数

人类决策主体$\mathcal{H}$的感知效用函数U_h可分解为价值函数和概率权重函数$w(\cdot)$的耦合形式：

$$U_h(a_h|s_t) = \sum_{i=1}^{n} v\left(\Delta x_i\right) \cdot w(p_i)$$

其中：

价值函数$v(\Delta x)$表征结果相对于参考点$r^{(t)}$的偏离感知，满足分段特性：

$$v(\Delta x) = \begin{cases} (\Delta x)^{\alpha} & \text{if } \Delta x \geq r^{(t)} \\ -\lambda \cdot \left(r^{(t)} - \Delta x\right)^{\beta} & \text{if } \Delta x < r^{(t)} \end{cases}$$

参数$\alpha = 0.88$ 与 $\beta = 0.92$分别表征收益和损失域的风险敏感程度，$\lambda =$ 2.25 反映损失厌恶系数，验证数据源自实证行为实验[4]。

概率权重函数$w(p)$描述人类对客观概率的非线性扭曲：

$$w(p) = \frac{p^{\gamma}}{(p^{\gamma} + (1-p)^{\gamma})^{1/\gamma}}$$

超参数γ随概率区间动态调整：当$p < 0.5$时$\gamma = 0.61$（低估大概率事件），反之$\gamma = 0.69$（高估小概率事件），从而捕捉“可能性效应”与“确定性效应”[5]。

3.1.2 定义 2：算法威慑的强化学习模型

算法智能体$\mathcal{A}$基于深度确定性策略梯度（DDPG）框架优化威慑策略，目标为最小化长期违规频率。其关键组件包括：

状态空间：$s_t = \left(\delta p_t,\ \Delta r^{(t)},\ f_c\right)^{\mathrm{T}} \in \mathbb{R}^8$，其中$\delta p_t$表示当前风险偏离度，$\Delta r^{(t)} = r^{(t)} - r^{(t-1)}$为参考点动态调整量，$f_c$为历史合规率构成的时序特征。

Q 函数更新规则：

$$Q^{\pi_a}\left(s_t,\ a_a\right) \leftarrow \mathbb{E}st + 1\left[R_t + \gamma \cdot \max_{a_a} Q^{\pi_a}\left(s_{t+1},\ a_a^{'}\right)\right]$$

其中即时奖励 $R_t = \omega_1 \mathbb{I}$ compliant $- \omega_2 D$ KL $(\pi_a^t \| \pi_a^{t-1})$ $\gamma = 0.95$融合合规激励与策略振荡惩罚， 为未来折扣因子。

3.2 同步控制逻辑优化

人机交互过程可分解为三阶段闭环动态博弈，具体机制如下：

阶段 I：认知适配威慑信号生成算法根据当前状态s_t生成原生威慑信号$D_t=(p_t,\ l_t)$，其中p_t为预测违规概率，l_t为预期损失强度。为避免人类对低概率 - 高损失事件的感知偏差，算法执行认知适配转换：

概率纠偏：构建逆向概率权重函数$w^{-1}(p)$，修正原生概率：

$$\tilde{p}_t=\frac{p_t^{\gamma/(1-\gamma)}}{\left(p_t^{\gamma/(1-\gamma)}+(1-p_t)^{\gamma/(1-\gamma)}\right)}$$

损失放大：针对损失厌恶特性放大信号强度：

$$L_t=\lambda\cdot l_t^{1/\beta}\ (\lambda=2.25,\ \beta=0.92)$$

最终输出适配信号$D_t^{'}=(\tilde{p}_t,\ L_t)$。

阶段Ⅱ：人类风险感知与决策人类接收$D_t^{'}$后，通过前景效用评估各行动$a_h\in\mathcal{A}_h$：

效用计算：基于参考点$r^{(t)}$计算相对效用值：

$$\Delta x_i=x\left(a_h^{(i)}\right)-r^{(t)},\quad U_h\left(a_h^{(i)}\right)=v(\Delta x_i)\cdot w(\tilde{p}_t)$$

随机选择：采用 Logit 响应规则生成行为策略：

$$\pi_h\left(a_h^{(i)}|D_t^{'}\right)=\frac{\exp\left(\kappa U_h\left(a_h^{(i)}\right)\right)}{\sum_j\ \exp\left(\kappa U_h\left(a_h^{(j)}\right)\right)}$$

其中系数$\kappa\in[0,\ +\infty)$控制决策理性度：$\kappa\to0$时完全随机化，而$\kappa\to+\infty$逼近理性最优解。

阶段Ⅲ：策略动态调整算法通过观察人类反馈$(a_h,\ s_{t+1})$更新威慑策略：

经验回放：存储交互元组$\langle s_t,\ a_a,\ R_t,\ s_{t+1}\rangle$至缓存池 D，采用优先采样策略提升 TD-error 样本的学习效率。

信任域约束：为规避策略突变导致威慑失效，引入 KL 散度约束策略更新幅度：

$$\pi_a^{t+1}=\underset{\pi}{\operatorname{argmin}}D_{\mathrm{KL}}\left(\pi^{'}\parallel\pi_a^t+\eta\nabla J(\pi_a)\right)\qquad \text{s.t. } D_{\mathrm{KL}}\le\epsilon$$

其中学习率η和信任域半径ϵ通过自适应优化算法动态调整[6]。

4 实验验证与结果分析

验证基于前景理论的人机博弈模型在动态交互环境下的有效性，具体检验以下假设：

假设 H1：人类决策者的行为模式显著受认知适配威慑信号的修改参数（$\tilde{p}_t$，L_t）影响，且符合价值函数与概率权重函数的预测；

假设 H2：算法动态调整策略后，系统的长期合规率f_c显著高于静态威慑策略；

假设 H3：信任域约束策略能有效降低策略振荡方差σ_π^2（相对于无约束策略降低至少 30%）。

4.1 实验场景

采用混合主体实验框架，将真人被试与算法智能体嵌入标准化的风险决策博弈环境。

构建基于 Web 的“合规 - 威慑”博弈平台，人类被试在交互中面临以下任务序列：

4.1.1 任务描述

被试作为企业管理者需连续完成 100 轮生产决策，每轮选择是否以降低安全成本（节省$\mathbb{S}x$）为代价进行违规操作（对应真实实验报酬）。

4.1.2 威慑机制

每轮开始前系统显示两类信息：

原生威慑信号：客观违规概率$p_t\in[0.05,\ 0.8]$与损失$l_t\in[10,\ 50]$；

适配威慑信号：修正概率$\tilde{p}_t$与放大损失L_t（通过模型参数计算）；

决策反馈：若被试选择违规：以概率p_t被检测，损失l_t（反映为扣除实验积分），实际扣除值按适配信号L_t执行（被试仅观察适配后的损失结果）。

4.1.3 样本选择

实验采用分层随机抽样获取 N=200 有效样本（每组 50 人）。

标准排除：剔除实验经验超过 1000 小时或决策类实验参与次数 ≥5 次的被试。

4.1.4 分组设计

组 1（控制组）：接收原生威慑信号，算法策略无约束；

组 2（修正组）：接收适配信号$D_t^{'}$，算法策略无约束；

组 3（约束组）：接收适配信号$D_t^{'}$，算法策略 KL 约束。

4.1.5 数据收集

通过时间序列面板数据记录全实验周期动态。

4.2 分析步骤

4.2.1 假设 H1 检验（行为模式验证）

价值函数参数估计：构建极大似然估计模型：

$$\max_{\alpha,\ \beta,\ \lambda}\prod_{t=1}^{100}\pi_h\left(a_h^{(t)}|\widehat{U}_h^{(t)}\right)$$

其中$\widehat{U}_h^{(t)}$依据适配后的$D_t^{'}$计算。通过对比预设值$\lambda=2.25$与估计值$\hat{\lambda}$的偏差分析模型有效性。

概率权重验证：通过二元 Logit 回归检验适配信号的修正效应：

$$\ln\left(\frac{\mathbb{P}(a_h=1)}{\mathbb{P}(a_h=0)}\right)=\theta_0+\theta_1\tilde{p}_t+\theta_2L_t$$

若θ_1显著为负且θ_2显著为负，则 H1 成立。

4.2.2 假设 H2 检验（动态策略效能）

采用混合效应模型评估长期合规率的组间差异：

$$f_{c,\ it} = \beta_0 + \beta_1 \cdot \text{修正组}_i + \beta_2 \cdot \text{时间}\ t + \epsilon it + u_i$$

其中个体随机效应 $u_i \sim \mathcal{N}(0,\ \sigma_u^2)$，若 $\beta_1 > 0$ 且 p<0.05，则 H2 成立。

4.2.3 假设 H3 检验（策略稳定性）

计算各组策略振荡方差：

$$\sigma_\pi^2 = \frac{1}{T-1}\sum_{t=1}^{T-1}(\pi_a^{t+1} - \pi_a^t)^2$$

若约束组 σ_π^2 显著小于无约束组（ANOVA 检验，效果量 η^2>0.15），则 H3 成立。

4.2.4 结果解释

假设 H1 验证：修正组被试的损失厌恶系数估计均值为 $\hat{\lambda} = 2.18 \pm 0.23$（p=0.62 vs λ =2.25），理论值与实际行为无显著偏离（t=1.23，p=0.22）。

适配后的 $\tilde{p}_t$ 对违规行为的抑制效应比原生 p_t 强 32%（OR=0.68，95%Cl [0.57，0.82]）。

假设 H2 验证：长期合规率组间差异显著（约束组 $f_c = 0.74 \pm 0.06$ vs 控制组 $f_c = 0.61 \pm 0.09$，$\beta 1$ =0.13，p <0.01）。

假设 H3 验证：约束组的策略振荡方差降低 42%（σ_π^2 从 0.019 降至 0.011，F=9.37，p<0.001），且未导致收敛速度下降（迭代次数差异 Δ=3.2%，p=0.54）。

4.2.5 模型有效性结论

实验结果支持人机博弈模型的关键假设，表明：

①认知适配机制能有效校正人类决策者的概率感知偏差。

②动态策略更新机制在保障稳定性的同时显著提升威慑效能。模型的拟合优度（AIC=1123 vs 静态模型 AIC=1452）证实其解释力优势，为算法威慑系统的工程实现提供实证依据。

5 结论

算法威慑的有效性并非单纯取决于技术能力，而在于对人类行为规律的深刻解读。研究发现，用户对威胁的感知具有显著的非理性特征：轻微的风险提示可能在心理上被放大（如“损失规避效应”），而高强度的威慑反而因引发麻木或抵触情绪而失效。例如，社交平台通过隐性标签而非直接删除内容来抑制虚假信息传播，本质上是对用户“选择性关注”心理的适配；自动驾驶系统动态调整安全介入阈值，则是对用户习惯逐渐适应的周期性修正。这种“心理适配性”要求算法设计以前景理论为框架构建预测模型，从硬性规则控制转向柔性行为引导，进而平衡威慑力度与用户接受度。

展望未来研究，还需着重解决两大维度的挑战。一方面是人机博弈的长期动态均衡亟待探索：当人类逐渐识破或适应算法威慑策略时，如何设计自进化机制以避免威慑失效？另一方面是复杂场景的协同控制能否实际落地。例如，在金融欺诈检测等涉及多方博弈的场景中，需协调多个智能体间的策略冲突，防止过度威慑引发的系统性误判。研究的最终目标是要将人机博弈从“威胁 - 服从”的简单逻辑，演进为“引导 - 协作”的共生范式，使算法威慑既具备伦理稳健性，又能融入实际社会治理与商业场景。

参考文献

[1] Crandall, J. W., Oudah, M., Ishowo-Oloko, F., Abdallah, S., Bonnefon, J. F. (2018). Cooperating with machines. Nature Communications, 9(1), 233.

[2] Tesla Safety Team. (2023). Autopilot’s Collision Avoidance Mechanism: Technical Whitepaper. https://www.tesla.com/autopilot.

[3] Chen，X.，Zhang，Y.，Wang，H.，& Liang，Y. (2022). A dynamic adversarial risk analysis model for financial fraud detection. Expert Systems with Applications，187，115895.

[4] Tversky A, Kahneman D. Advances in prospect theory：Cumulative representation of uncertainty[J]. Journal of Risk and Uncertainty, 1992.

[5] Barberis N C. Thirty years of prospect theory in economics：A review and assessment[J]. Journal of Economic Perspectives, 2013.

[6] Schulman J, et al. Trust region policy optimization[C]. ICML, 2015.