arXiv:2601.03451v1 [stat.ML] 6 Jan 2026

多智能体学习的微观经济学基础

Nassim Helou

哈里斯堡大学，美国哈里斯堡；
Thatch，美国纽约；2026年1月

摘要

现代人工智能系统日益在市场和制度中运行，在这些环境中，数据、行为和激励都是内生的。本文通过研究马尔可夫决策过程中的委托人—代理人互动，为多智能体学习构建经济学基础。该过程存在战略外部性，委托人和代理人都会随时间不断学习。我们提出一种两阶段激励机制：首先估算可实施的转移支付，然后利用这些支付引导长期动态；在温和的基于遗憾的理性和探索条件下，该机制能够实现次线性社会福利遗憾，从而达到渐近最优福利。模拟结果表明，即使粗略的激励也能纠正存在状态依赖外部性时低效的学习，凸显了在市场和保险领域进行激励感知设计、实现安全且符合福利目标的人工智能的必要性。

1 引言

人工智能不再是孤立运行的技术，而是一股嵌入市场、制度和大规模系统中的经济力量。现代人工智能系统——大型基础模型、多智能体模拟器、自主决策者、数据市场和算法保险机构——运行于充满战略行为者的环境中，他们的目标塑造着人工智能所依赖的数据和信息流。随着这类系统日益广泛地部署，人们不禁会想，这些系统之间的互动如何才能更有利于提升社会福利。

因此，一个核心挑战随之浮现：能否运用经济理论中的理念和概念来改进这类系统？典型且重要的问题包括：人工智能应如何推理激励？经济机制应如何塑造学习代理的行为？博弈论是否能为某些学习和决策行为提供解释？理解这一交汇领域，对于确保人工智能系统安全运行，并使其目标策略策略 与社会福利保持一致至关重要。所有这些也有助于了解已部署的算法能否抵御战略行为和串谋。

保险自动化市场受到的影响尤为显著，因为它们处于预测、激励和策略性行为的交汇点，而一旦学习算法成为核心决策者，这些方面都会发生根本性变化。保费、免赔额、承保限额和除外责任都发挥着激励作用，会影响报告行为、风险承担，甚至风险本身。因此，市场和保险都深受上述各方面影响。然而，机器学习 领域的相关文献尚未充分沿着这一方向发展，这正是本文的切入点。传统保险建立在统计估计、风险分摊和契约设计之上。但随着人工智能成为调和风险预测、动态定价和欺诈检测的关键代理人，这些任务逐渐演变为多智能体学习

问题，其中存在策略性参与者：客户会学习如何响应定价信号，大型模型会从受行为影响的数据中学习风险分布，而保险公司则必须设计激励结构，以维持如实报告或适当的风险水平。新兴的人工智能驱动型市场（例如自动化市场、自主物流网络、在线广告拍卖）也面临同样的结构性挑战：经济外部性、策略性信息揭示，以及未能协调一致的探索激励。人工智能系统与人类及彼此互动的市场，从根本上说，都是学习代理人的博弈。

这一转变揭示了一个深刻的理论鸿沟。经典机器学习假设数据是外生的，代理人的行为是短视的。经典经济学假设环境是已知的，并且优化是完全理性的。但在上述情境中，这两个假设都不成立：代理人在未知环境中学习、适应、探索，并相互操纵。借用机制设计中的术语，我们现在可以将平台（保险人、监管者等）称为委托人，将互动中的其他参与者称为代理人。由于效用函数或环境是未知的，委托人必须同时进行学习和决策。与此同时，环境也会随着这些学习过程而演化。其结果是形成了一种由相互作用的学习者构成的新型不确定且具有策略性的环境 [Rothschild 等人，2025年, Immorlica 等人， 2024]。来自强化学习 [Kaelbling 等人，1996年，Sutton 等人， 1999], 契约理论、博弈论和机制设计的工具必须在核心层面加以融合，才能提供有价值的洞见。

由于行业参与者（例如保险公司、在线平台）与学术界交叉领域的研究较为稀缺——相关研究往往侧重于经典博弈论或更纯粹的以机器学习为导向的研究——我们旨在提出若干核心问题，并提供初步的算法见解。本文通过构建一个统一框架，为研究多代理人系统中的激励相容学习作出贡献。该框架以契约经济学理论和外部性为基础，并采用在线学习与统计学领域的现代工具。我们从这样一个观察出发：人工智能系统日益扮演委托人的角色，需要从人类或人工代理人那里获取信息和努力，而这些代理人的内部目标、学习动态和类型均未知。正如契约设计文献 [Guesnerie 和 Laffont，1984年, Bolton 和 Dewatripont，2004年, Kőszegi， 2014] 以及委托学习文献 [Saig 等人， 2023], 所示，激励 [参见 Laffont 和 Martimort 的大型专著，2002] 会以关键方式影响统计性能和探索行为。例如，当数据收集委托给学习代理时，委托人必须兼顾隐蔽状态和隐蔽行动，设计出即使评估含噪也依然稳健的转移支付方案 [Ananthakrishnan 等人， 2024]。当代理人面临有成本的探索时，标准强化学习算法会违反激励相容性，因此需要信息设计机制来确保适当的探索 [Simchowitz 和 Slivkins, 2024]。同样，在重复博弈、老虎机问题和马尔可夫决策过程环境中，也会出现外部性、道德风险和策略性操纵 [Scheid 等人, 2024b]，这凸显了经典经济力量如何在学习环境中重新显现。

我们的研究建立在这些洞见之上，并将其拓展到一个真正动态且具有状态依赖性的环境中。正如本文所形式化的，我们考虑一个马尔可夫决策过程（MDP）[Bellman, 1957, Puterman，1990] ，其中委托人和代理人都在随时间学习，而且代理人的行动不仅会影响自身的回报，也会影响委托人的奖励以及系统的状态转移动态。这一环境体现了人工智能在数据驱动型市场和保险系统中的若干关键特征：预测与行为之间的反馈回路、可能造成外部性的探索，以及关于代理偏好的不完全

信息。在统计学领域的理论研究中，预测器影响其学习所依托系统的反馈回路正日益受到关注，这种现象被称为表征性预测，例如 Perdomo 等人 [2020], 、Mendler-Dünner 等人 [2020] 或 Brown 等人 [2022]的研究。在这类系统中，除非委托人能够推断代理人的学习动态，并设计出能够内化外部性的转移支付，否则经典效率定理便不再成立。我们表明，尽管面临这些挑战，精心构建的两阶段机制仍能实现渐近最优的社会福利：委托人可以先学习如何影响代理人，再利用这种影响引导长期结果朝有利方向发展。

至关重要的是，我们将这些思路拓展到对抗式系统之外。近期研究表明，扩散模型中的生成建模技术可被解释为经济聚合机制，从而实现福利最大化估计量和均衡决策规则。我们对这种联系进行了形式化，并证明扩散模型的去噪步骤对应于社会规划者问题的唯一解，且可在大型代理人经济中作为均衡实现。这揭示了经济理论与最先进生成式人工智能之间出人意料的联系：扩散模型实现了某种形式的高效市场聚合。将其与委托人—代理人学习相结合，为从经济学角度设计协作式人工智能系统，并将其映射到生成模型，提供了全新的思路。

综合这些要素，本文主张构想这样一个未来：人工智能系统作为经济制度发挥作用——成为激励的调解者、去中心化学习者的协调者。由此产生的问题是，如何使这类系统在受策略性反馈影响的环境中，以福利优化为导向。没有哪一领域比保险更能体现这一点：人工智能被用于评估风险、生成合同或侦测欺诈——消费者行为甚至可以被视为一个大型多代理人学习问题。在这样的环境中，保险不能仅仅依赖精算；它还必须是算法驱动的、关注激励的，并以学习动态为基础。我们的框架为这一愿景提供了理论基础，以及来自行业的实践洞见。

总而言之，本研究提供了：(i) 一个关于具有内生外部性的马尔可夫决策过程中的委托人—代理人学习的严谨模型；(ii) 动态系统中激励相容机制的福利与遗憾保证；以及 (iii) 扩散模型与经济聚合之间的概念和数学桥梁。这些贡献共同揭示了人工智能与经济学如何

cal必须紧密结合起来，为未来的市场和保险基础设施构建安全且符合福利目标的系统。

2 相关工作

在深入探讨模型之前，我们先回顾一些与本文设定相关的重要研究。多智能体系统中的学习与激励研究，处于契约理论与现代强化学习方法的交汇处。经典委托—代理人理论奠定了基础：从隐藏行动和隐藏信息问题的开创性表述开始 [Mir- rlees, 1999], 经济学文献刻画了委托人如何通过提供结果依存型转移支付，促使代理人采取成本高昂且不可观测的行动。这些模型传统上假设环境是静态的或规模较小的动态环境，且结果分布完全已知。其核心贡献在于阐明了类型、成本或行动无法直接观测时的激励相容约束、参与约束，以及最优合同的结构。我们的研究继承了这一概念逻辑，但将其扩展至这样的环境：委托人和代理人都是学习参与者，在对状态转移

动态和奖励结构。

第一条重要研究脉络将契约理论拓展到算法领域和高维领域。组合契约 [Dütting 等人，2022]以及多智能体契约设计[Dütting 等人，2023]研究了这样一类情境：委托人的奖励取决于多个代理人之间的组合交互，或许多可能的努力维度。这些研究针对复杂环境中的最优线性合同或有界合同，提出了近似方案、不可能性结果和结构刻画。近期研究表明，可以通过伪维度来理解各类合同，从而为利用代理人类型数据集离线学习近最优合同提供样本复杂度保证 Dütting 等人 [2025]。沿着这一方向，一些研究探讨了菜单和分段线性合同在表达能力与可学习性之间的权衡。总体而言，这些文献奠定了大规模契约设计的算法基础。

越来越多的研究将契约理论与在线学习相结合。多项研究考察了多臂老虎机反馈下重复的委托—代理交互，在这种情况下，委托人只能观测结果的随机实现，而无法观测代理人类型或奖励函数。在线契约学习框架 [Scheid 等人，2024年c、Liu 等人， 2025] 研究委托人如何在学习未知奖励的同时确保激励相容性。类似思想也出现在树状图中更复杂的多智能体结构里：Scheid 等人 [2025b、a]证明，局部的单步转移支付足以在全局上引导所有参与者趋向最优联合行动，从而在完全去中心化系统中有效实现福利最大化。这些研究表明，在缺乏关于代理人响应的结构性假设时，委托人遗憾必然呈线性增长；而适度的限制——例如经验贪婪行为或基于淘汰的行为——则可以恢复次线性遗憾。更复杂的模型

在代理人一侧纳入策略性学习：Liu 和 Ratliff [2024] 研究了会维护自身经验估计、并可能进行任意探索的代理人，为稳健激励机制证明了近乎最优的遗憾界 [Liu 和 Ratliff， 2024]。作为对这些研究的补充，Wu 等人 [2025] 提出了一般性的学习以引导模型，其中代理人可能通过虚报或诱导误导性观测，在策略上操纵委托人的学习过程。这些结果共同凸显了激励相容性与统计学习之间微妙的相互作用，这也是本文的核心主题。

近期研究强调学习任务的委托，以及针对数据质量或探索激励设计激励机制。一个值得关注的方向是在去中心化或联邦环境中委托数据收集。Ananthakrishnan 等人 [2024] 表明，当委托人依赖策略性智能体收集数据，并将这些数据用于后续训练时，隐蔽行动和隐蔽状态都会自然出现；尽管奖励和数据质量存在不确定性，基于绩效的合同仍能实现近乎最优的委托 [Ananthakrishnan 等人，2024]。与之密切相关的是探索中的激励相容约束：在对代理人而言探索成本高昂的强化学习场景中，Simchowitz 和 Slivkins [2024]证明，标准强化学习算法违反了经典的激励相容性；探索必须通过受控的信息披露来组织，而非依靠货币转移支付 [Simchowitz 和 Slivkins，2024]; Capitaine 等人 [2024]则研究了当数据收集会给代理人带来成本时，委托人如何组织代理人收集数据以开展协作学习。早期的委托人—代理人多臂老虎机模型则采取相反的思路：委托人直接付费让代理人进行探索，从而使委托人能够学习未知的奖励函数 Scheid 等人 [2024c]。我们的研究沿袭了这一思路，但将交互置于马尔可夫系统之中，并允许双方都进行学习。另一个主要研究方向

关注外部性和协调失灵引发的激励问题。对于固定且完全理性的情形，科斯定理的相关结论早已存在数十年[Coase，2013，Medema，2020，Farrell，1987，Deryugina 等人， 2021]。先前研究已将这一设定扩展为未知环境中的博弈，并纳入学习因素。在双智能体多臂老虎机设置中，Scheid 等人 [2024b] 表明，如果没有产权，福利最大化结果可能无法实现，因为代理人不会将外部性内化；令人意外的是，恰当的转移支付方案可以恢复科斯定理的在线类比。Zuo [2024]将这些观点扩展到包含学习和不确定性的动态环境，并强调在线议价和稳定性概念的重要性 [Zuo， 2024]。公平性考量也已进入相关研究：Tłuczek 等人 [2025] 表明，线性合约可以调整，以满足异质智能体之间的公平性约束，同时在重复交互中保持次线性遗憾和高福利。这些研究表明，当代理人是学习者而非完全理性的优化者时，必须重新诠释外部性内化、议价和公平性等经典概念。

此外，近期研究将委托人与代理人的推理与马尔可夫决策过程中的强化学习联系起来以及马尔可夫博弈。Ivanov 等人 [2024]提出了委托—代理强化学习，引入一种元算法，通过对策略进行交替优化，使委托—代理马尔可夫决策过程收敛到子博弈完美纳什均衡（SPNE）。他们证明，基于合约的支付可以解释为一种具有明确经济学含义的奖励塑形形式，而深度强化学习能够将此类机制扩展到大型马尔可夫决策过程。将其扩展到多智能体马尔可夫博弈 [的研究（例如，Littman，1994；Nowé 等人，2012；Zhang 等人，2021；Yang 和 Wang，2020；有关总体概述，参见]）表明，基于合约的干预能够缓解硬币游戏等环境中的序贯社会困境。除此之外，此类设置向平均场博弈的扩展也已得到研究 [Lasry 和 Lions，2007；Bensoussan 等人，2013；Guo 等人，2019],在这些研究中，调解者通过激励大量无后悔智能体，使其趋向预期的均衡，即便存在模型不确定性[Widmer 等人， 2025]。这些结果凸显了基于学习的激励设计在序贯环境和大规模群体环境中的重要性，也预示着未来人工智能生态系统的复杂性。

最后，这些研究方向与实验设计领域的文献密切相关 [Kirk，2009；Berger 等人，2018；Federer， 1956], 该领域研究应如何选择数据，以最大化统计效率和下游决策质量。在经典统计学中，实验设计将信息获取与资源约束之间的权衡形式化。在现代机器学习中，这些问题在自适应、序贯和交互式环境中再次出现，此时数据会受到学习代理的行为以及系统中嵌入的激励影响。这一视角将激励设计、探索和数据收集与最优设计原则联系起来，而后者将学习视为在信息结构上的优化问题

en而非单一的估计任务。最优设计 [s Atkinson，2014年，Goos 等人，2016] 已成为一种基础工具，可用于选择哪些数据有助于训练和推断。从基于人类反馈的强化学习中的奖励模型训练这一视角来看 [Wang 等人，2024a,b，Fu 等人， 2025], 人工标注的偏好对的选择可以被表述为一个纯探索多臂老虎机问题[Zhao 等人，2024年，Scheid 等人，2024a]。通过刻画简单遗憾并构造匹配的上界和下界，可以说明激励（在此情形下，即对高成本人工标注工作量的分配）如何影响奖励推断的统计效率。委托—代理学习问题可以被理解为内生实验设计的实例，其中机制和转移支付不仅决定代理人的行为

也决定学习本身的统计效率——这一主题直接推动并补充了本文所提出的框架。

总体而言，这些研究方向汇聚为一个核心洞见：随着人工智能系统日益由激励和信息分散的交互代理组成，必须将经典契约理论与在线学习相融合，以开发现代且公平的系统。本文通过研究马尔可夫环境中双方均进行学习的委托人—代理人交互，为这一综合研究作出了贡献，并展示了激励设计、探索策略和多智能体协调如何在动态且不确定的环境中相互作用。

3 具有外部性的委托人—代理人马尔可夫决策过程中的激励设计

随着人工智能系统日益介入经济活动、社会协调和大规模决策过程，理解学习型代理人如何进行策略性交互，对于确保这些系统安全、高效且公平地运行变得至关重要。本文建立的理论框架为应对这些挑战提供了坚实的基础，展示了如何将经济机制融入学习系统，使个体代理人为全球期望的结果作出贡献。通过证明即使人工智能系统无法直接控制环境、必须推断其他代理人的偏好和学习动态，也能够在理论上恢复社会福利，我们希望这项研究能够开启设计人工智能平台的大门，使其无需强制手段或关于代理人理性的不切实际假设，便能引导去中心化生态系统。

随着人工智能不断从实验室走向开放、复杂的市场，这类结果至关重要：网约车平台、生成模型市场、多智能体仿真环境和协作机器人，都围绕激励而非直接控制构建。了解如何设计转移支付、讨价还价方案 [Muthoo, 1999, Powell, 2002, Ståhl, 1973]，以及激励相容协议，有助于预测和规范代理人的行为，降低遭受剥削或福利崩溃的风险。同样重要的是，这些洞见有助于开发能够推理激励、与人类协商，并承诺采用公平透明机制以协调不同利益相关方行为的人工智能。随着在线科斯式结果从简单的多臂老虎机推广到复杂的马尔可夫决策过程环境，我们获得的不仅是新的理论保证，也获得了一份概念路线图，指导我们构建融合学习、合同与战略推理的人工智能系统。本研究凸显了从根本层面将经济学与在线学习结合起来的必要性，并有助于确保新一代人工智能技术能够在其必然要运行的多智能体、激励驱动的世界中蓬勃发展。

我们将多臂老虎机环境下建立的外部性与讨价还价框架扩展到马尔可夫决策过程，其中代理人控制环境，而委托人通过转移支付影响代理人的行为。代理人的行动同时决定委托人的奖励和转移概率，且两位参与者都会随时间学习。与多臂老虎机相关的在线议价一样，委托人试图通过动态转移支付策略将外部性内部化，但马尔可夫决策过程的结构带来了更长的探索阶段和更复杂的学习动态。

3.1 设置

环境是一个有限时域马尔可夫决策过程，其状态空间为 $\mathcal{S}$，$|\mathcal{S}| = S$，动作空间为 $\mathcal{A}$，$|\mathcal{A}| = K$。对于任意状态 $s,s' \in \mathcal{S}$ 和行动 $a \in \mathcal{A}$，我们有一个转移核 $P(s' \mid s,a)$；智能体奖励 $r_a(s,a) \in [0,1]$

以及委托方奖励 $r_p(s,a) \in [0,1]$。

回合的时域为 H。在回合 k 的步骤 h，状态为 s_h^k。首先，委托人选择转移支付向量 $\tau_h^k(\cdot) \in \mathbb{R}_+^K$，然后代理人采取行动 $a_h^k \in A$，代理人获得奖励:

$$r_a(s_h^k, a_h^k) + \tau_h^k(a_h^k) \,,$$

而委托人的奖励为

$$r_p(s_h^k, a_h^k) - \tau_h^k(a_h^k) \,,$$

其中，转移支付在每一轮加总后构成参与者各自的效用。如果设想保险人（委托人）与客户（代理人）的情境，激励措施通常是为客户提供折扣或促销，以换取某些有利的合同。最后，状态转移为$s_{h+1}^k \sim P(\cdot \mid s_h^k, a_h^k)$。

理性的代理人会最大化预期累积回报

$$\mathbb{E}\left[\sum_{k=1}^{T}\sum_{h=1}^{H}(r_a(s_h^k, a_h^k) + \tau_h^k(a_h^k))\right],$$

而委托人会最大化

$$\mathbb{E}\left[\sum_{k=1}^{T}\sum_{h=1}^{H}(r_p(s_h^k, a_h^k) - \tau_h^k(a_h^k))\right].$$

将参与者获得的效用相加，我们将社会福利定义为

$$W = \mathbb{E}\left[\sum_{k=1}^{T}\sum_{h=1}^{H}(r_a(s_h^k, a_h^k) + r_p(s_h^k, a_h^k))\right],$$

因为转移支付在此处相互抵消。因此，转移支付用于塑造参与者的行为，但不计入全局福利。其唯一用途是使参与者的效用保持一致。

策略与社会福利。一个平稳代理人策略是一个映射 $\pi_a : \mathcal{S} \to \Delta(\mathcal{A})$，而平稳转移支付策略是一个映射 $\pi_\tau : \mathcal{S} \to \mathbb{R}_+^K$。

令 $V^{\pi_a}{}_{,a}{}^{\pi_\tau}$和 $V^{\pi_a}{}_{,p}{}^{\pi_\tau}$分别表示代理人和委托人的价值函数。在 (π_a, π_τ) 下，社会福利为

$$W(\pi_a, \pi_\tau) = V_a^{\pi_a,\pi_\tau} + V_p^{\pi_a,\pi_\tau} \,.$$

因此，我们将最优全局福利定义为

$$W^\star := \max_{\pi_a} W(\pi_a, \pi_\tau) \,,$$

注意到转移支付不会影响福利。

3.2 参与者行为

智能体学习与理性。令$\pi_{a,\,k}$表示代理人在回合 k 中的策略，该策略由根据过去回合进行更新的强化学习算法生成。我们提出一个回合遗憾假设，该假设类似于一种事后理性条件。

定义 1（智能体理性）。如果存在 $C > 0$ 和 $\zeta > 0$，使得对于任意转移支付序列（$\pi_{\tau,\,1}$，……，$\pi_{\tau,\,T}$），以至少 $1 - T^{-\zeta}$ 的概率，代理人满足回合遗憾指数 $\kappa \in [0,1)$

$$\sum_{k=1}^{T}(V_a^{\pi_a^\star,\pi_{\tau,k}} - V_a^{\pi_{a,k},\pi_{\tau,k}}) \le CT^\kappa,$$

其中 $\pi_a^\star$ 是代理人（在 r_a下）的最优平稳策略。

委托人的目标与福利遗憾。现在，大部分设定已经定义完毕，我们将注意力转向参与者的目标。形式上，我们将全局福利定义为

$$W_k = V_a^{\pi_{a,k},\pi_{\tau,k}} + V_p^{\pi_{a,k},\pi_{\tau,k}} \,,$$

社会福利遗憾为

$$R_{\mathrm{sw}}(T) = TW^\star - \sum_{k=1}^{T} W_k \,.$$

目标是设计 $\pi_{\tau,\,k}$ 使得 $R_{\mathrm{sw}}(T) = o(T)$。再以一家采用人工智能算法的保险公司为例，福利就是客户（乐于投保并愿意为此支付费用）和保险人（其目标是获取收入）双方获得的效用。

3.3 委托人的两阶段算法

第一阶段：转移支付估计。 对于每个（s, a），委托人寻求最小的转移支

$$\tau_s^\star(a) = \max_{a'} \big(Q_a(s,a') - Q_a(s,a)\big)_+,$$

其中，Q_a 是代理人的最优状态-动作价值函数。

通过分批二分搜索，委托人会在多个回合批次中提供固定的转移支付，并观察代理人在状态 s下选择 a 的频率，以估计 $\tau_s^\star(a)$。

第二阶段：福利优化。 一旦估计值 $\hat{}\tau_s(a)$ 满足

$$|\hat{\tau}_s(a) - \tau_s^\star(a)| \le T^{-\beta}\,,$$

委托人就能通过提供$\hat{}\tau_s(a)$，在 s 有效地实施任意期望的行动。随后，她会在平移后的马尔可夫决策过程中，使用有效奖励运行无遗憾强化学习算法（例如 UCB-VI）

$$\tilde{r}_p(s,a) = r_p(s,a) - \hat{\tau}_s(a),$$

从而保持福利不变。请注意，我们在此给出了一个简单的理论结果，但也可以在使用上下文强化学习算法时纳入特征。进一步而言，我们认为深度强化学习算法或许能为我们的设定带来益处。

3.4 主要结果

现在，方法和算法已介绍完毕，我们给出主要定理，证明见附录。

定理1（委托人–代理人马尔可夫决策过程中的社会效率）。 假设代理人满足指数为 $\kappa < 1$ 的事后理性，并且在探索策略下，该马尔可夫决策过程是一致遍历的，从而每个批次中每个状态都会被访问 $\Theta(T^\alpha)$ 次，其中 $\alpha > 0$。假设委托人选择的指数 $\alpha, \beta \in (0,1)$ 满足

$$\kappa < \alpha < 1, \qquad \frac{\beta}{\alpha} < 1 - \kappa.$$

那么，存在一种两阶段的委托人—代 $gorith^{m\ 成功}\ h\ that\ w^{ith\ hi}\ gh\ probabilit\ y:$

$$R_{\mathrm{sw}}(T) = O(T^\alpha \,\mathrm{polylog}(T) + T^\gamma \,\mathrm{polylog}(T) + T^\kappa \,\mathrm{polylog}(T))\,,$$

其中，$\gamma < 1$ 是委托人的强化学习算法在第二阶段的遗憾指数。特别地，

$$R_{\mathrm{sw}}(T) = o(T)\,,$$

因此，委托人实现了渐近最优的社会福利。

4 作为福利最大化经济机制的扩散模型

最后，我们正式建立扩散模型与经典经济聚合原则之间的联系，从而完成这项处于人工智能与经济理论交叉领域的工作。我们证明，扩散模型所学习的去噪器是二次福利下社会规划者问题的唯一解，并且同一对象还作为微观基础委托人—代理人模型中的均衡聚合器出现。因此，扩散模型可以被解释为聚合关于潜在状态的噪声信息的经济机制。

4.1 扩散模型预备知识

设 $x_0 \in \mathbb{R}^d$ 取自未知分布 $p_{\mathrm{data}}(x_0)$。扩散模型（保持方差）定义了一个前向加噪过程

$$x_t = \alpha_t x_0 + \sigma_t \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I_d)\,, \quad (1)$$

其中 $t \in [0,1]$、$\alpha_0 = 1$、$\sigma_0 = 0$ 和 $\alpha_1 \approx 0$、$\sigma_1 \approx 1$。

令 $\varepsilon_\theta(x_t,\ t)$ 为通过以下损失训练得到的去噪网络

$$\mathcal{L}(\theta) = \mathbb{E}_{t,\ x_0,\ \varepsilon}\big\|\varepsilon - \varepsilon_\theta(\alpha_t x_0 + \sigma_t \varepsilon,\ \ t)\big\|^2\,. \quad (2)$$

众所周知（例如在去噪分数匹配中），唯一的极小化解为

$$\varepsilon^\star(x_t, t) = \mathbb{E}[\varepsilon \mid x_t] \,. \tag{3}$$

线性高斯模型下的贝叶斯法则给出

$$\mathbb{E}[x_0 \mid x_t] = \frac{1}{\alpha_t}(x_t - \sigma_t\, \varepsilon^\star(x_t, t)) \,, \tag{4}$$

这具有一项显著优势，即能够提供一个简洁的闭式表达式。

4.2 一个社会规划者问题

考虑一位规划者，他只能观察到 x_t，并选择一个重构$\hat{x}(x_t) \in \mathbb{R}^d$。将福利定义为负的平方误差:

$$W(\hat{x}) \coloneqq -\mathbb{E}\big[\|x_0 - \hat{x}(x_t)\|^2\big] \,. \tag{5}$$

规划者要解决

$$\max_{\hat{x}} W(\hat{x}) \iff \min_{\hat{x}} \mathbb{E}\|x_0 - \hat{x}(x_t)\|^2 \,. \tag{6}$$

命题1 (贝叶斯去噪器最大化福利)。 规划者问题(6) 的唯一解为

$$\hat{x}^\star(x_t) = \mathbb{E}[x_0 \mid x_t]. \tag{7}$$

证明。 任取一个可测的 $\hat{x}(x_t)$。记

$$\begin{aligned} x_0 - \hat{x}(x_t) = \big(x_0 - \mathbb{E}[x_0 \mid x_t]\big) \\ + \big(\mathbb{E}[x_0 \mid x_t] - \hat{x}(x_t)\big). \end{aligned}$$

取平方范数、展开并取期望:

$$\begin{aligned} \mathbb{E}\|x_0 - \hat{x}(x_t)\|^2 = \mathbb{E}\|x_0 - \mathbb{E}[x_0 \mid x_t]\|^2 \\ + \mathbb{E}\|\mathbb{E}[x_0 \mid x_t] - \hat{x}(x_t)\|^2 \,, \end{aligned}$$

由于根据条件期望的定义，交叉项为零。该表达式当且仅当 $\hat{x}(x_t) = \mathbb{E}[x_0 \mid x_t]$ 几乎处处时取最小值。 □

利用（4），我们得到以下推论。

推论1 （扩散训练 =福利最大化）。 扩散损失 L 的最小化者 $\varepsilon^\star$ 通过线性关系实现福利最大化决策规则 $\hat{x}^\star(x_t)$

$$\hat{x}^\star(x_t) = \frac{1}{\alpha_t}\big(x_t - \sigma_t \varepsilon^\star(x_t, t)\big).$$

因此，求解扩散训练问题等价于求解 (6)。

4.3 微观基础的经济学诠释

我们建立一个由 $i \in [0,1]$ 索引的连续统代理人模型。存在一个隐藏状态 $x_0 \in \mathbb{R}^d$。每个代理人都观察到一个信号

$$y_i = x_0 + \eta_i, \qquad \eta_i \sim \mathcal{N}(0, \sigma^2 I), \tag{8}$$

在各个 i 上相互独立。委托人选择一项公共决策 $d \in \mathbb{R}^d$。每个代理人的效用为

$$u_i(d, x_0) = -\|d - x_0\|^2 \,,$$

而社会福利为

$$W(d, x_0) = \int_0^1 u_i(d, x_0)\, di = -\|d - x_0\|^2 \,.$$

考虑一种直接机制，其中每个代理人报告 m_i，委托人使用一项结果规则

$$d = g(m_{[0,1]}) = \phi\Big(\int_0^1 m_i \, di\Big).$$

引理1 (大型经济中的如实报告)。如果委托人采用线性规则 $d = A\int_0^1 m_i\, di$，那么在代理人连续统的情况下，如实报告$m_i = y_i$ 是贝叶斯纳什均衡。

证明。 在存在连续统代理人的情况下，任何单个代理人对$\int_0^1 m_i\, di$的影响都微乎其微。因此，每个代理人都会将结果视为固定不变。报告不会改变d，所以代理人在各种报告策略之间并无差异；因此，如实报告构成一种均衡（前提是作出有利的平局裁决，而这在此类情形中是非常常见的假设）。 □

现在假设委托人观察到一个由以下方式生成的带噪声宏观信号

$$x_t = \alpha_t x_0 + \sigma_t \varepsilon \ , \tag{9}$$

如扩散前向过程中所述。委托人的福利问题再次是 $\max_d -\mathbb{E}\|x_0 - d\|^2$。

将命题1与引理1相结合，最终得到以下结果。

命题2 （扩散漂移作为福利最大化均衡）。在二次效用、对称先验和如实报告（引理*1*）的条件下，给定带噪宏观信号 x_t，使预期福利最大化的机制实现了

$$d^\star(x_t) = \mathbb{E}[x_0 \mid x_t],$$

这与扩散模型的最优去噪器完全对应，即

$$d^\star(x_t) = \frac{1}{\alpha_t}\big(x_t - \sigma_t \varepsilon^\star(x_t, t)\big).$$

因此，反向扩散更新方向是大规模经济中对带噪私有信息进行福利最大化聚合的聚合器。

本节确立了扩散模型与经济聚合之间的等价性：扩散模型学习到的得分函数或去噪器，同时由以下两者刻画：(i) 二次福利下的社会规划者最优解 ，以及(ii) 大规模代理人经济中的均衡机制 。这为将扩散模型理解为聚合关于潜在状态的分散且带噪信息的工具提供了有原则的经济学解释，并将其自然地与多智能体学习系统中的激励设计联系起来。尽管我们关于扩散模型与大规模委托—代理经济学之间联系的结果尚属初步，但我们希望这些结果能为这一交叉领域的未来研究提供启示

5 模拟

最后，我们通过一个简单的、具有状态依赖外部性的委托—代理马尔可夫决策过程，说明激励的作用。该环境是一个有限时域的线性世界，代理人可以从三种行动中进行选择：快速行动能让代理人迅速前进，但会造成严重污染；缓慢行动会产生中等排放；绕行行动虽然会给代理人带来成本，却能减少累积的污染。污染是一个显式的状态变量，会随时间演变，并在每个步骤和终止状态下对委托人的奖励产生负面影响。相比之下，代理人只在意尽快到达目标，不会直接将污染成本纳入考量。社会福利定义为代理人奖励与委托人奖励之和，货币转移支付则相互抵消。这一情境经典地展现了委托人与代理人之间效用不一致、角色各异的情况。我们旨在说明，激励能够让参与者以有利的方式协调效用，从而恢复全局福利。

我们比较两种设定。第一种不提供转移支付，代理人仅根据自身奖励，通过 Q学习进行学习。第二种中，委托人提供一项简单的、与状态无关的补贴，鼓励代理人采取绕行行动。该补贴经过校准，用于抵消代理人承担污染减排的私人成本，但不取决于状态或博弈历史。在这两种情况下，代理人都采用 ε-贪心表格型 Q学习算法，并根据长期平均社会福利和终端污染水平评估表现。

图1和图2所示的结果表明，这两种机制之间存在明显的定性差异。在没有转移支付的情况下，代理人几乎总是偏好快速行动，导致污染持续累积，社会福利较低。引入简单的补贴后，代理人学到的行为发生了显著变化：代理人在回合初期越来越多地选择绕行行动，从而减少污染随时间的累积。因此，平均社会福利显著提高，而平均回合结束时的污染水平则降低。

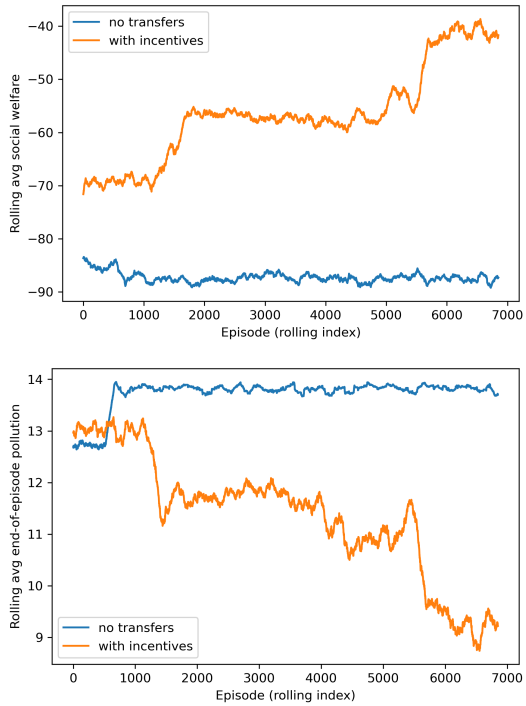

图1：具有有状态外部性的委托人—代理人马尔可夫决策过程中的激励效果。上图：滚动平均社会福利。下图：滚动平均终端污染。通过引入简单的补贴来促使污染减排，可显著提升福利。

重要的是，尽管补贴的设计十分简单，这些改善仍然得以实现。

这项实验展示了本文的核心观点：在存在状态依赖外部性的多智能体学习环境中，自利强化学习可能收敛到系统性低效的结果，因此必须采用激励方案——即使是非常简单的方案——才能使个体学习行为与社会福利保持一致。本文采用的补贴机制刻意设计得十分简约，但观察到的收益表明，有必要进一步研究前文从理论上探讨的、更具结构性的激励相容机制。

6 结论

我们认为，下一代人工智能系统应当作为经济机制来研究，并最终以经济机制的方式加以设计。当人工智能在市场和保险基础设施中介入预测、签约、定价和执行时，“从数据中学习”与“设计激励”之间的经典区分便不复存在。数据变得内生，行为会对策略作出响应，而优化则在由相互作用的学习者构成的耦合系统中展开。我们的首项贡献是在马尔可夫决策过程中建立了一个正式的委托人—代理人框架，其中双方都会随时间学习，且代理人的行动共同影响奖励和状态转移。在这一模型中，我们证明，转移支付虽然事后对福利没有影响，却是实现事前福利对齐的有力工具：它们重塑代理人的学习问题，使私人激励能够内化外部性。由此形成的两阶段机制提供了清晰的概念模板。在第一阶段，委托人确定实施期望行动所需的最低转移支付；在第二阶段，委托人利用这些估计有效引导系统的长期动态。在温和的正则条件下，这种方法能够实现次线性社会福利遗憾。我们的第二项贡献是在经济聚合与现代生成式建模之间搭建概念和数学桥梁。

最后，我们的模拟在一个透明环境中阐明了这一核心信息。总体而言，本文推动了人工智能部署领域一种新兴观点的发展：要在战略环境中设计安全且符合福利目标的系统，就需要协同设计学习动态与经济机制。

参考文献

Nivasini Ananthakrishnan、Stephen Bates、Michael Jordan 和 Nika Haghtalab。去中心化机器学习中的数据收集委托。载于人工智能与统计国际会议，第 478–486 页。PMLR，2024年。Anthony C Atkinson。最优设计。*Wiley StatsRef: Statistics Reference Online*，第 1–17 页，2014年。Richard Bellman。马尔可夫决策过程。数学与力学杂志，第 679–684 页，1957年。Alain Bensoussan、Jens Frehse、Phillip Yam 等。平均场博弈与平均场型控制理论，第 101 卷。施普林格，2013年。Paul D Berger、Robert E Maurer 和 Giovana B Celli。实验设计。施普林格，2018年。Patrick Bolton 和 Mathias Dewatripont。契约理论。麻省理工学院出版社，2004年。Gavin Brown、Shlomi Hod 和 Iden Kalemaj。有状态世界中的施为预测。载于人工智能与统计国际会议，第 6045–6061 页。PMLR，2022年。Aymeric Capitaine、Etienne Boursier、Antoine Scheid、Eric Moulines、Michael I Jordan、El-Mahdi El-Mhamdi 和 Alain Durmus。协作学习中的解缠。*arXiv* 预印本 *arXiv:2407.14332*，2024年。Ronald Harry Coase。社会成本问题。法律与经济学杂志，56（4）：837–877，2013年。Tatyana Deryugina、Frances Moore 和 Richard SJ Tol。科斯定理的环境应用。《环境科学与政策》，120：81–88，2021年。Paul Dütting、Tomer Ezra、Michal Feldman 和 Thomas Kesselheim。组合契约。载于*2021 IEEE* 第 *62* 届计算机科学基础年度研讨会（*FOCS*），第 815–826 页。IEEE，2022年。Paul Dütting、Tomer Ezra、Michal Feldman 和 Thomas Kesselheim。多智能体契约。载于第 *55* 届 *ACM* 理论计算年度研讨会论文集，第 1311–1324 页，2023年。Paul Dütting、Michal Feldman、Tomasz Ponitka 和 Ermis Soumalias。契约的伪维数。载于第 *26* 届 *ACM* 经济学与计算会议论文集，第 514–539 页，2025年。Joseph Farrell。信息与科斯定理。《经济展望杂志》，1（2）：113–129，1987年。Walter F Federer。实验设计，第 81 卷。LWW，1956年。Jiayi Fu、Xuandong Zhao、Chengyuan Yao、Heng Wang、Qi Han 和 Yanghua Xiao。通过奖励塑形缓解基于人类反馈的强化学习中的奖励投机。*arXiv* 预印本 *arXiv:2502.18770*，2025年。

Peter Goos、Bradley Jones 和 Utami Syafitri。《混合试验的 I-最优设计》。美国统计学会期刊，111(514):899–911，2016年。Roger Guesnerie 和 Jean-Jacques Laffont。一个委托代理问题类别的完整解及其在自我管理企业控制中的应用。《公共经济学杂志》，25(3):329–369，1984年。Xin Guo、Anran Hu、Renyuan Xu 和 Junzi Zhang。学习平均场博弈。神经信息处理系统进展，32，2019年。Nicole Immorlica、Brendan Lucier 和 Aleksandrs Slivkins。生成式人工智能作为经济代理。*ACM SIGecom Exchanges*，22(1):93–109，2024年。Dima Ivanov、Paul Dütting、Inbal Talgam-Cohen、Tonghan Wang 和 David C Park。委托代理强化学习：通过合同协调人工智能代理。*arXiv* 预印本 *arXiv:2407.18074*，2024年。Leslie Pack Kaelbling、Michael L Littman 和 Andrew W Moore。强化学习：综述。《人工智能研究杂志》，4:237–285，1996年。Roger E Kirk。实验设计。《*Sage*心理学定量方法手册》，第23–45页，2009年。Botond Kőszegi。行为契约理论。《经济文献杂志》，52(4): 1075–1118，2014年。Jean-Jacques Laffont 和 David Martimort。《激励理论：委托代理模型》。普林斯顿大学出版社，2002年。Jean-Michel Lasry 和 Pierre-Louis Lions。平均场博弈。《日本数学杂志》，2(1):229–260，2007年。Michael L Littman。作为多智能体强化学习框架的马尔可夫博弈。载于《*1994*年机器学习论文集》，第157–163页。爱思唯尔，1994年。Junyan Liu 和 Lillian J Ratliff。面向自利型与探索型学习代理的委托人—代理人老虎机博弈。*arXiv* 预印本 *arXiv:2412.16318*，2024年。Junyan Liu、Arnab Maiti、Artin Tajdini、Kevin Jamieson 和 Lillian J Ratliff。在对抗性代理到达的重复委托代理问题中学习如何提供激励。*arXiv* 预印本 *arXiv:2505.23124*，2025年。Steven G Medema。科斯定理六十年。《经济文献杂志》，58(4): 1045–1128，2020年。Celestine Mendler-Dünner、Juan Perdomo、Tijana Zrnic 和 Moritz Hardt。施为预测的随机优化。神经信息处理系统进展，33:4929–4939，2020年。James A Mirrlees。道德风险与不可观测行为理论：第一部分。《经济研究评论》，66(1):3–21，1999年。Abhinay Muthoo。《谈判理论及其应用》。剑桥大学出版社，1999年。

Ann Nowé、Peter Vrancx 和 Yann-Michaël De Hauwere。博弈论与多智能体强化学习。载于《强化学习：最新进展》，第441–470页。施普林格，2012年。Juan Perdomo、Tijana Zrnic、Celestine Mendler-Dünner 和 Moritz Hardt。施为预测。载于国际机器学习会议，第7599–7609页。PMLR，2020年。Robert Powell。议价理论与国际冲突。《政治学年度评论》，5(1):1–30，2002年。Martin L Puterman。马尔可夫决策过程。《运筹学与管理科学手册》，2:331–434，1990年。David M Rothschild、Markus Mobius、Jake M Hofman、Eleanor W Dillon、Daniel G Goldstein、Nicole Immorlica、Sonia Jaffe、Brendan Lucier、Aleksandrs Slivkins 和 Matthew Vogel。智能体经济。*arXiv* 预印本 *arXiv:2505.15799*，2025年。Eden Saig、Inbal Talgam-Cohen 和 Nir Rosenfeld。委托分类。神经信息处理系统进展，36:13200–13236，2023年。Antoine Scheid、Etienne Boursier、Alain Durmus、Michael I Jordan、Pierre Ménard、Eric Moulines 和 Michal Valko。基于人类反馈的强化学习中奖励建模的最优设计。*arXiv* 预印本 *arXiv:2410.17055*，2024a。Antoine Scheid、Aymeric Capitaine、Etienne Boursier、Eric Moulines、Michael Jordan 和 Alain Durmus。学习缓解外部性：具有事后理性的科斯定理。神经信息处理系统进展，37:15149–15183，2024b。Antoine Scheid、Daniil Tiapkin、Etienne Boursier、Aymeric Capitaine、Eric Moulines、Michael Jordan、El-Mahdi El-Mhamdi 和 Alain Oliviero Durmus。委托人—代理人老虎机博弈中的激励学习。载于国际机器学习会议，第43608–43631页。PMLR，2024c。Antoine Scheid、Etienne Boursier、Alain Durmus、Eric Moulines 和 Michael I Jordan。分层多智能体系统中的合同学习。2025a。Antoine Scheid、Etienne Boursier、Alain Durmus、Eric Moulines 和 Michael I Jordan。具有转移支付的树状多智能体博弈中的在线决策。*arXiv* 预印本 *arXiv:2501.19388*，2025b。Max Simchowitz 和 Aleksandrs Slivkins。强化学习中的探索与激励。《运筹学》，72(3):983–998，2024年。Ingolf Ståhl。谈判理论。博士论文，斯德哥尔摩经济学院，1973年。Richard S Sutton、Andrew G Barto 等。强化学习。《认知神经科学杂志》，11(1):126–134，1999年。Jakub Tłuczek、Victor Villin 和 Christos Dimitrakakis。异质类型委托代理博弈中的公平合约。*arXiv* 预印本 *arXiv:2506.15887*，2025年。Binghai Wang、Rui Zheng、Lu Chen、Yan Liu、Shihan Dou、Caishuang Huang、Wei Shen、Senjie Jin、Enyu Zhou、Chenyu Shi 等。大语言模型中基于人类反馈的强化学习的奥秘（第二部分）：奖励建模。*arXiv* 预印本 *arXiv:2401.06080*，2024a。

Zhichao Wang、Bin Bi、Shiva Kumar Pentyala、Kiran Ramnath、Sougata Chaudhuri、Shubham Mehrotra、Xiang-Bo Mao、Sitaram Asur 等。《大语言模型对齐技术综述：基于人类反馈的强化学习、基于人工智能反馈的强化学习、近端策略优化、直接偏好优化等》。*arXiv* 预印本 *arXiv:2407.16216*，2024b。

Leo Widmer、Jiawei Huang 和 Niao He。《模型不确定性下引导无后悔智能体参与均场博弈》。*arXiv* 预印本 *arXiv:2503.09309*，2025年。

Yuchen Wu、Xinyi Zhong 和 Zhuoran Yang。《学习以引导：激励暗处的策略性智能体》。*arXiv* 预印本 *arXiv:2506.08438*，2025年。

Yaodong Yang 和 Jun Wang。《从博弈论视角概述多智能体强化学习》。*arXiv* 预印本 *arXiv:2011.00583*，2020年。

Kaiqing Zhang、Zhuoran Yang 和 Tamer Bacsar。多智能体强化学习：理论与算法精选综述。强化学习与控制手册，第 321–384 页，2021年。

Heyang Zhao、Chenlu Ye、Quanquan Gu 和 Tong Zhang。KL正则化上下文老虎机与基于人类反馈的强化学习的精细分析。*arXiv* 预印本 *arXiv:2411.04625*，2024年。

Shiliang Zuo。在线契约设计的新视角：异质、同质、非短视代理人及团队生产。*arXiv* 预印本 *arXiv:2403.07143*，2024年。

附录

6.1 定理1的证明

步骤1：通过批次实现动作可识别性。固定 (s,a)。在长度为 $L=T^{\alpha}$ 的一个批次中，假设过程至少访问状态 s $\Omega(T^{\alpha})$ 次（一致遍历性）。如果所提供的转移支付 τ满足 $\tau>\tau_s^{\star}(a)+T^{-\beta}$，那么根据代理人的事后理性条件，在状态 s 选择任意 $a'\neq a$ 都会产生至少 $\Omega(T^{\alpha})$ 的后悔值；除非错误操作至多发生在 $O(T^{\kappa})$ 次访问中，否则这与遗憾界矛盾。因此，代理人以 $1-O(T^{\kappa-\alpha})$ 的频率采取 a。

反之，若 $\tau<\tau_s^{\star}(a)-T^{-\beta}$，则 a至少次优 $T^{-\beta}$，且代理人选择 a 的次数至多为 $O(T^{\kappa-\alpha})$。由于 $\alpha>\kappa$，这些情形在统计上可区分。

步骤 2：分批二分搜索。在 $O(\log T)$ 个批次中重复此测试，并在每次测试后将 $\tau_s^{\star}(a)$ 的区间缩小一半，即可得到估计值 $\hat{\tau}_s(a)$，其误差至多为 $T^{-\beta}$；前提是 $\beta/\alpha<1-\kappa$，以确保误分类概率为 $o(1)$。

对所有 s，a 应用并集界，可知所有估计值都满足

$$0\le\hat{\tau}_s(a)-\tau_s^{\star}(a)\le 2T^{-\beta},$$

且这一结论以高概率同时成立。

步骤 3：期望行动的可实施性。对于任意 a'，我们有

$$Q_a(s,a')\le Q_a(s,a)+\tau_s^{\star}(a)\le Q_a(s,a)+\hat{\tau}_s(a),$$

因此，当委托人提供 $\hat{\tau}_s(a)$ 时，a 对代理人而言是最优的。根据遗憾界，代理人在第二阶段总共仅偏离 a $o(T)$ 次。

步骤 4：委托方强化学习与福利遗憾。在第二阶段，委托人实际上控制马尔可夫决策过程，其后悔值为 $O(T^{\gamma}\mathrm{polylog}T)$。第一阶段带来的后悔值至多为 $O(T^{\alpha}\mathrm{polylog}T)$，代理人的偏离则带来 $O(T^{\kappa}\mathrm{polylog}T)$。因此

$$R_{\mathrm{sw}}(T) = O(T^{\alpha}\,\mathrm{polylog}\,T + T^{\gamma}\,\mathrm{polylog}\,T + T^{\kappa}\,\mathrm{polylog}\,T)\ ,$$

由于 $\alpha,\ \gamma,\ \kappa < 1$，因此为 $o(T)$。这就证明了该定理。 □