American Journal of Artificial Intelligence
2026; 10(1): 101-113
http://www.sciencepublishinggroup.com/j/ajai
doi: 10.11648/j.ajai.20261001.20

Incentive-Aware Learning in Stateful Environments: Internalizing Externalities in Principal-Agent MDPs and Welfare-Maximizing Diffusion Mechanisms

Nassim Helou*

Department of Business, University of Harrisburg, Harrisburg PA, United States of America

Email address:
nhelou@my.harrisburgu.edu (Nassim Helou)
*Corresponding author

To cite this article:
Nassim Helou. (2026). Incentive-Aware Learning in Stateful Environments: Internalizing Externalities in Principal-Agent MDPs and Welfare-Maximizing Diffusion Mechanisms. *American Journal of Artificial Intelligence, 10*(1), 101-113.
https://doi.org/10.11648/j.ajai.20261001.20

Received: 8 January 2026 ; **Accepted:** 6 February 2026 ; **Published:** 26 February 2026

Abstract: Modern AI systems increasingly operate in economic environments such as markets and insurance, where data, behavior, and incentives are endogenous and mutually reinforcing. This paper develops a microeconomic foundation for multi-agent learning by studying a repeated principal-agent interaction in a finite-horizon Markov decision process with strategic externalities, where both the principal and the agent learn over time and the agent's actions affect payoffs and the environment's dynamics. We design incentive schemes that align decentralized learning with social welfare via a two-phase mechanism: in Phase 1, the principal estimates the minimal transfers required to implement targeted actions by identifying how incentives shift the agent's effective preferences; in Phase 2, the principal uses these estimates to steer long-run state-action visitation toward welfare-optimal behavior. We evaluate performance using social-welfare regret relative to the best achievable welfare benchmark, and we show that the mechanism achieves sublinear social-welfare regret under mild conditions (sublinear agent regret and sufficient exploration/coverage), implying asymptotically optimal welfare despite endogenous externalities and simultaneous learning. Simulations in a simple pollution-control environment illustrate that even coarse incentives can correct inefficient learning outcomes and substantially improve welfare. These results underscore that incentive-aware design, grounded in contract theory and mechanism design, is essential for safe, welfare-aligned AI deployed in strategic economic systems.

Keywords: Artificial Intelligence, Machine Learning, Multi-Agent Systems, Computational Game Theory

1. Introduction

Artificial intelligence is no longer a technology acting in isolation, but an economic force embedded inside markets, institutions, and large–scale systems. Modern AI systems, large foundation models, multi-agent simulators, autonomous decision-makers, data markets, and algorithmic insurers – operate in environments filled with strategic actors whose objectives shape the data and information flows on which AI relies. With the increasing deployment of such systems, one may wander how the interactions of such systems can be made oriented towards greater social welfare. Thus, a central challenge emerges: are there ideas and concepts from economic theory that could be employed in order to improve such systems? Typical and important questions are: how should AI reason about incentives, how should economic mechanisms shape the behavior of learning agents, are there insights from game theory that may explain some learning and decision–making behaviors? Understanding this interface is essential for ensuring that AI systems behave safely and align their target *policy* with social welfare. All this will also help to understand if deployed algorithms are robust to strategic behaviors and collusion.

Insurance and automated markets are particularly affected because they sit at the intersection of prediction, incentives, and strategic behavior—all of which are fundamentally altered once learning algorithms become central decision-makers.

Premiums, deductibles, coverage limits, and exclusions are not merely pricing parameters: they are incentive instruments that shape reporting, effort and prevention, risk-taking, and ultimately the underlying risk distribution itself. In such environments, data are endogenous: observed claims and behavioral signals reflect how participants respond to the mechanism, rather than an exogenous distribution.

Traditional insurance theory combines statistical estimation, risk pooling, and contract design under relatively stable behavioral assumptions. By contrast, as AI systems increasingly mediate risk prediction, dynamic pricing, underwriting, and fraud detection, these tasks become multi-agent learning problems with strategic participants. Customers learn how to respond to pricing and coverage signals, platform models learn risk distributions from data that are themselves behavior-dependent, and insurers must design incentive structures that sustain truthful reporting and socially efficient risk choices over time. Similar feedback loops arise in AI-driven marketplaces more broadly—for instance in automated trading and pricing platforms, autonomous logistics networks, and online advertising auctions—where externalities, strategic information revelation, and misaligned exploration incentives are central. Markets in which AI systems interact with humans and with each other are therefore, fundamentally, *games of learning agents*.

This transformation exposes a profound theoretical gap. Classical machine learning assumes that data is exogenous and agent behavior is myopic. Classical economics assumes known environments and fully rational optimization. But in the settings above, neither assumption holds: agents learn, adapt, explore, and manipulate each other in an unknown environment. Leveraging the vocabulary from mechanism design, we can now call the platform (insurer, regulator, etc) the *principal* and the other players in interaction the *agents*. Since utility functions or the environment are unknown, the principal must learn and make decisions simultaneously. At the same time, the environment evolves as a consequence of these learning processes. The result is a new regime of uncertain and strategic environment [1, 2] made of interacting learners. Tools from reinforcement learning [3, 4], contract theory, game theory, and mechanism design must be fused at a core level to provide valuable insights.

Motivated by the scarcity of work at the intersection of industry actors (e.g., insurers, online platforms) and academia, which tends to focus either on classical game theory or on more pure ML-oriented research, we aim to formulate several core questions and provide initial algorithmic insights. This paper contributes to this agenda by developing a unified framework for studying incentive–compatible learning in multi–agent systems, grounded in the economic theory of contracts and externalities and in modern tools from online learning and statistics. We start from the observation that AI systems increasingly function as principals that must elicit information and effort from human or artificial agents whose internal objectives, learning dynamics, and types are unknown. As shown in the literature on contract design [5–7] and delegated learning [8], incentives see the very extensive book, [9] shape statistical performance and exploration behavior in essential ways. For instance, when data collection is delegated to learning agents, the principal must account for both hidden states and hidden actions, designing transfer schemes that remain robust despite noisy evaluation [10]. When agents face costly exploration, standard RL algorithms violate incentive compatibility, requiring information–design mechanisms to ensure proper exploration [11]. Likewise, externalities, moral hazard, and strategic manipulation appear in repeated bandit and MDP settings [12], emphasizing how classical economic forces reemerge in learning environments.

Our work builds on these insights and pushes them into a genuinely dynamic and stateful setting. As formalized in this paper, we consider a Markov decision process (MDP) [13, 14] in which both the principal and the agents are learning over time, and where the agent's actions influence not only their own returns but also the principal's reward and the transition dynamics of the system. This environment captures essential features of AI in the context of data–powered markets and insurance systems: feedback loops between predictions and behavior, exploration that may impose externalities, and incomplete information about agent preferences. In the context of theoretical works in the field of statistics, feedback loops where a predictor influences the system from which it learns is increasingly studied as *performative prediction*, as in [15, 16] or [17]. In such a system, classical efficiency theorems break down unless the principal can infer the agent's learning dynamics and design transfers that internalize externalities. We show that, despite these challenges, a carefully constructed two–phase mechanism yields asymptotically optimal social welfare: the principal can first learn how to influence the agent and then use this influence to steer long–run favorable outcomes.

Crucially, we extend these ideas beyond contract–based systems. Recent works reveals that the generative modeling techniques from diffusion models can be interpreted as economic aggregation mechanisms, implementing welfare – maximizing estimators and equilibrium decision rules. We formalize this connection and show that the denoising step of diffusion models corresponds to the unique solution of a social planner problem, and can be implemented as an equilibrium in a large–agent economy. This offers a surprising connection between economic theory and state of the art generative AI: diffusion models perform a form of efficient market aggregation. When integrated with principal–agent learning, this provides whole new ideas for designing collaborative AI systems in economic terms and mapping them to generative models.

Putting these elements together, this paper argues for a future in which AI systems behave as economic institutions, mediators of incentives, coordinators of decentralized learners. Questions then arise about how such systems can be oriented towards welfare optimization in environments shaped by strategic feedback. Nowhere is this more relevant than in insurance, where AI is used to evaluate risks, generate contracts or detect fraud. Consumer behavior could even be framed as a large multi–agent learning problem. In

such environments, insurance cannot be merely actuarial; it must be algorithmic, incentive–aware, and grounded in learning dynamics. Our framework provides both theoretical foundations and practical insights from the industry toward this vision.

In summary, this work provides (i) a rigorous model of principal-agent learning in MDPs with endogenous externalities, (ii) welfare and regret guarantees for incentive-compatible mechanisms in dynamic systems, and (iii) a conceptual and mathematical bridge between diffusion models and economic aggregation. Together, these contributions shed light about how AI and economics must be tightly linked to build safe and and welfare–aligned systems for the markets and insurance infrastructures of the future.

Existing work on incentives in learning environments typically falls into one of two regimes: (i) static principal-agent models in which the outcome distribution is known and only the contract is optimized, or (ii) bandit-style incentive-learning models where actions do not induce stateful feedback and externalities are not propagated through dynamics. These assumptions are increasingly misaligned with modern deployments in markets and insurance, where incentives change behavior, behavior changes data, and data changes subsequent decisions in a closed loop. In such settings, incentives are not a one-shot design choice: they are part of the learning process itself, and they must be designed under uncertainty while the environment evolves endogenously.

This paper addresses that gap by modeling a repeated principal-agent interaction inside a finite-horizon Markov decision process with externalities, where both principal and agent learn over time and agent actions jointly affect rewards and transitions. The key technical challenge is that the principal cannot directly control actions, must learn how transfers map into implementable behavior, and must do so in a stateful environment where exploration and incentives interact. We resolve this by introducing a two-phase mechanism that (i) estimates minimal implementable transfers needed to induce target actions and (ii) leverages those estimates to steer long-run dynamics toward welfare-optimal behavior while maintaining incentive compatibility under no-regret learning.

Our contributions are threefold. First, we formalize a principal-agent learning model in an MDP with endogenous externalities and simultaneous adaptation on both sides, making precise the welfare benchmark and the notion of social-welfare regret. Second, we provide a regret guarantee showing that, under mild conditions (sublinear agent regret and sufficient state visitation under exploration), the proposed two-phase mechanism achieves sublinear social-welfare regret, hence asymptotically optimal welfare. Third, we connect welfare maximization to diffusion-model denoising through a planner/aggregation interpretation, offering a micro-founded perspective on generative inference as a welfare-optimal decision rule. Together, these results supply both a theoretical framework and practical guidance for incentive-aware learning in strategic economic systems.

We also summarize abbreviations here: AI - Artificial Intelligence; MDP - Markov Decision Process; RL - Reinforcement Learning; SPNE - Subgame-Perfect Nash Equilibrium; MFG - Mean-Field Game; RLHF - Reinforcement Learning from Human Feedback.

2. Related Work

Before diving into the model, we review some important works linked with our setting. The study of learning and incentives in multi-agent systems lies at the intersection of contract theory and modern reinforcement-learning approaches. Classical principal-agent theory provides the foundation: beginning with the seminal formulations of hidden-action and hidden-information problems [18], the economic literature characterizes how a principal induces an agent to take costly, unobservable actions by offering outcome-contingent transfers. These models traditionally assume static or small dynamic environments with full knowledge of outcome distributions. Their central contribution is the articulation of incentive-compatibility constraints, participation constraints, and the structure of optimal contracts when types, costs, or actions are not directly observed. Our work inherits this conceptual logic but extends it to environments where both the principal and the agents are learning players, repeatedly interacting together in large state spaces under uncertainty about transition dynamics and reward structures.

A first major strand of work extends contract theory into algorithmic and high-dimensional domains. Combinatorial contracts [19] and multi-agent contract design [20] study settings where the principal's reward depends on combinatorial interactions between multiple agents or many possible effort dimensions. These works develop approximation schemes, impossibility results, and structural characterizations of optimal linear or bounded contracts in complex environments. More recent results show how contract classes can be understood through their pseudo-dimension, yielding sample-complexity guarantees for offline learning of near-optimal contracts from agent-type datasets [21]. In the same direction, some works explore the trade-offs between expressiveness and learnability of menus and piecewise-linear contracts. Together, this literature establishes the algorithmic foundations of large-scale contract design.

A growing line of research integrates contract theory with online learning. Several works investigate repeated principal-agent interactions under bandit feedback, in which the principal observes only the stochastic realization of outcomes but not the agent's type or reward function. Online contract-learning frameworks [22, 23] study how a principal can ensure incentive compatibility while simultaneously learning unknown rewards. Similar ideas appear in more complex multi-agent structures in tree-like graphs: [24, 25] demonstrate that local, one-step transfers suffice to globally steer all players toward the optimal joint action, effectively achieving welfare maximization in fully decentralized systems. These works show that without structural assumptions on agent response, principal regret is necessarily linear, while mild

restrictions, such as empirically-greedy or elimination-based behavior, recover sublinear regret. More sophisticated models incorporate strategic learning on the agent side: [26] study agents who maintain their own empirical estimates and may explore arbitrarily, proving nearly optimal regret bounds for robust incentivization [26]. Complementing these works, [27] introduce a general *learning to lead* model where the agent may strategically manipulate the principal's learning by misreporting or inducing misleading observations. These results collectively highlight the delicate interplay between incentive compatibility and statistical learning, a theme central to our paper.

Recent literature highlights the strategic delegation of learning tasks, specifically focusing on how incentive structures govern data quality and exploration. A significant subset of this work investigates delegated data acquisition within decentralized or federated architectures, where the principal must navigate the inherent trade-offs between agent effort and statistical precision. [10] show that when the principal relies on strategic agents to collect data that will later be used for training, both hidden actions and hidden states arise naturally, and performance-based contracts can achieve near-optimal delegation despite uncertainty in rewards and data quality [10]. Closely related are incentive-compatibility constraints in exploration: in reinforcement-learning settings where exploration is costly for the agent, [11] demonstrate that standard RL algorithms violate classic incentive compatibility, and that exploration must be orchestrated using controlled information disclosure rather than monetary transfers [11] while [28] study how a principal can orchestrate data collection by agents in the purpose of collaborative learning when such collection is costly to the agents. Earlier principal-agent bandit models take the opposite view: the principal directly pays agents to explore, allowing the principal to learn unknown reward functions [22]. Our work follows this line of thought but embeds the interaction inside a Markovian system and allows both sides to learn. Another major direction concerns incentive problems arising from externalities and coordination failures. For the fixed and fully rational scenario, results from the Coase theorem have existed for decades [29–32]. Previous works have been developed to extend the setup to a game in an unknown environment with learning. In two-agent bandit settings, [12] show that without property rights, welfare-maximizing outcomes may be impossible because agents fail to internalize externalities; surprisingly, appropriate transfer schemes restore an online analogue of the Coase theorem. [33] extends these ideas to dynamic environments with learning and uncertainty, emphasizing the importance of online bargaining and stability notions [33]. Fairness considerations have also entered the literature: [34] show that linear contracts can be adapted to satisfy fairness constraints across heterogeneous agents while preserving sublinear regret and high welfare in repeated interactions. Such works illustrate how classical concepts, externality internalization, bargaining, and fairness, must be reinterpreted when agents are learners rather than fully rational optimizers.

Separately, recent works connect principal-agent reasoning with reinforcement learning in MDPs and Markov games. [35] propose principal-agent reinforcement learning, introducing a meta-algorithm that converges to subgame-perfect Nash equilibrium (SPNE) in principal-agent MDPs through alternating optimization over policies. They show that contract-based payments can be interpreted as a form of reward shaping with principled economic meaning, and that deep RL can scale such mechanisms to large MDPs. Extensions to multi-agent Markov games [see, e.g. 36–39, for a general overview] demonstrate how contract-based interventions can mitigate sequential social dilemmas in environments such as the Coin Game. Complementary, extensions of such setups to mean-field games have been studied [40–42], where a mediator incentivizes a large population of no-regret agents towards desired equilibria despite model uncertainty [43]. These results highlight the importance of learning-based incentive design in sequential and population-scale environments, foreshadowing the complexity of future AI ecosystems.

Finally, these lines of work are closely connected to the experimental design literature [44–46], which studies how data should be selected in order to maximize statistical efficiency and downstream decision quality. In classical statistics, experimental design formalizes the trade-off between information acquisition and resource constraints. In modern machine learning and AI research, these concerns reemerge in adaptive, sequential, and interactive settings, where data is shaped by the behavior of learning agents and by the incentives embedded in the system. This perspective links incentive design, exploration, and data collection to optimal design principles, which frame learning as an optimization problem over information structures rather than a single estimation task. Optimal design [47, 48] has appeared to be fundamental as a tool to select which data can be useful for training and inference. In the perspective on reward-model training in RLHF [49–51], the selection of human-labeled preference pairs can be framed as a pure-exploration bandit problem [52, 53]. By characterizing simple regret and constructing matching upper and lower bounds, it can be shown how incentives (in this case, allocation of costly human annotation effort) shape the statistical efficiency of reward inference. Principal-agent learning problems can be interpreted as instances of endogenous experimental design, in which mechanisms and transfers determine not only agent behavior but also the statistical efficiency of learning itself, a theme that directly motivates and complements the framework developed in this paper.

Despite the richness of these individual threads, a significant theoretical and structural gap remains at the intersection of dynamic incentives [54, 55] and modern generative architectures [56, 57]. Existing literature on principal-agent learning [63] largely confines itself to "static" or "bandit-style" interactions where the environment does not evolve as a consequence of the players' actions. While Markov games [67] have begun to address sequential dependencies, they frequently overlook the endogenous externalities that arise when a principal's learning process directly competes with or relies upon the agent's own exploration dynamics. This

lack of a unified framework for stateful, incentive-compatible learning means that current AI systems are ill-equipped for environments like insurance or automated markets [66], where every prediction shapes the future state of the risk pool and every incentive alters the quality of incoming data.

Furthermore, there is a distinct conceptual disconnect between the algorithmic design of AI models [64] and the economic objectives they are increasingly meant to serve. While generative models, specifically diffusion-based architectures [65], have achieved state-of-the-art performance in data synthesis, they are typically treated as pure statistical estimators rather than mechanisms for economic coordination [58]. There is no established theory that explains how these high-dimensional models [59] act as aggregators of decentralized information or how they relate to social welfare optimization in a multi-agent economy. Our work addresses this void by not only providing a regret-optimal framework for principal-agent MDPs [35] with externalities but also by demonstrating that the mathematical backbone of generative AI can be rigorously interpreted as a welfare-maximizing social planner's tool.

Overall, these lines of research converge towards a central insight: *as AI systems increasingly consist of interacting agents whose incentives and information are distributed, classical contract theory must be fused with online learning* to develop modern and fair systems. This paper contributes to this synthesis by studying principal-agent interactions in Markovian environments with learning on both sides, demonstrating how incentive design, exploration strategies, and multi-agent coordination interact in dynamic and uncertain settings.

3. Incentive Design in a Principal Agent MDPs with Externalities

As AI systems increasingly mediate economic activity, social coordination, and large-scale decision processes, understanding how learning agents interact strategically becomes essential for ensuring that these systems behave safely, efficiently, and fairly. The theoretical framework developed in this work provides a principled foundation for addressing these challenges, showing how economic mechanisms can be integrated into learning systems so that individual agents contribute to globally desirable outcomes. By demonstrating that social welfare can provably be recovered, even when the AI system does not control the environment directly and must infer the preferences and learning dynamics of other agents, we hope that this research opens the door to designing AI platforms that can steer decentralized ecosystems without coercion or unrealistic assumptions about agent rationality.

These results are especially important as AI transitions from laboratory settings to open, complex markets (e.g., ride-sharing platforms, generative-model marketplaces, multi-agent simulation environments, and collaborative robotics), in which behavior is shaped primarily by incentives rather than direct control. Understanding how to design transfers, bargaining schemes [60–62], and incentive-compatible protocols allows to predict and regulate how agents behave, reducing risks of exploitation or welfare collapse. Equally importantly, these insights support the development of AI that can reason about incentives, negotiate with humans, and commit to fair and transparent mechanisms that align behavior across diverse stakeholders. As online Coasean results generalize from simple bandits to rich MDP environments, we gain not only new theoretical guarantees but also a conceptual roadmap for building AI systems that combine learning, contracts, and strategic reasoning. This work highlights the necessity of merging economics and online learning at a fundamental level and helps ensure that the next generation of AI technologies can thrive within the multi-agent, incentive-driven world in which they will inevitably operate.

We extend the externality and bargaining framework developed in the bandit setting to a MDP in which the agent controls the environment while the principal influences the agent's behavior through transfers. The agent's actions determine both the principal's reward and the transition probabilities, and both players learn over time. As in the online bargaining linked with bandits, the principal seeks to internalize externalities through dynamic transfer policies, but the MDP structure introduces a longer exploration phase and more complex learning dynamics.

3.1. Setting

3.2. Players' Behaviors

The environment is a finite-horizon Markov decision process (MDP) with state space $\mathcal{S}$, $|\mathcal{S}| = S$, and action space $\mathcal{A}$, $|\mathcal{A}| = K$. For any $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$, the dynamics are governed by a transition kernel $P(s' \mid s, a)$. The agent and the principal receive bounded stage rewards $r_a(s, a) \in [0, 1]$ and $r_p(s, a) \in [0, 1]$, respectively.

Episodes have horizon H. In episode $k \in [T]$ and step $h \in [H]$, the state is s_h^k. The principal first posts a nonnegative transfer vector $\tau_h^k(\cdot) \in \mathbb{R}_+^K$, after which the agent selects an action $a_h^k \in \mathcal{A}$. The agent's realized reward is

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

while the principal's realized reward is

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

In the economic interpretation, $\tau_h^k(a)$ represents an incentive such as a discount, rebate, or promotion associated with choosing contract/action a. Finally, the next state is drawn according to

$$s_{h+1}^k \sim P(\cdot \mid s_h^k, a_h^k).$$

Given a sequence of transfers $\{\tau_h^k\}_{k,h}$ and the induced

sequence of agent actions, the interaction generates a trajectory distribution over $\{(s_h^k, a_h^k)\}_{k\le T,\, h\le H}$. The agent is assumed to act rationally so as to maximize the expected cumulative utility

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

while the principal seeks to maximize

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

Summing both objectives yields the social welfare for the players

$$\begin{aligned} W(T) &:= J_a(T) + J_p(T) && (1)\\ &= \mathbb{E}\left[\sum_{k=1}^{T}\sum_{h=1}^{H}\big(r_a(s_h^k, a_h^k) + r_p(s_h^k, a_h^k)\big)\right], && (2)\end{aligned}$$

since the transfer terms cancel exactly. Thus transfers do not directly enter welfare ex post; rather, they act as a mechanism-design instrument that reshapes the agent's incentives and thereby the induced state–action distribution, allowing the principal to internalize externalities and steer learning toward high-welfare behavior.

Policies and Social Welfare. A stationary agent policy is a mapping $\pi_a : \mathcal{S} \to \Delta(\mathcal{A})$, and a stationary transfer policy is a mapping $\pi_\tau : \mathcal{S} \to \mathbb{R}_+^K$.

Let $V_a^{\pi_a,\pi_\tau}$ and $V_p^{\pi_a,\pi_\tau}$ denote the value functions for the agent and principal. Social welfare under (π_a, π_τ) is

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

We thus define the optimal global welfare as

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

noting that transfers do not affect welfare.

3.3. Players' Behaviors

(i) Agent Learning and Rationality. Let $\pi_{a,k}$ be the agent's policy in episode k, produced by a reinforcement learning algorithm that updates from past episodes. We impose an episodic regret assumption analogous to a form of hindsight-rationality condition.

Definition 3.1 (Agent Rationality). The agent satisfies episodic regret exponent $\kappa \in [0, 1)$ if there exists $C > 0$ and $\zeta > 0$ such that for any transfer sequence $(\pi_{\tau,1}, \ldots, \pi_{\tau,T})$, with probability at least $1 - T^{-\zeta}$,

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

where $\pi_a^\star$ is an optimal stationary policy for the agent (under r_a).

(ii) Principal's Objective and Welfare Regret. Now that we most of the setting is defined, we turn our attention to the objectives that the players have. Formally, we define the global welfare as

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

and the social welfare regret is

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

The goal is to design $\pi_{\tau,k}$ such that $R_{\mathrm{sw}}(T) = o(T)$. Again thinking to an insurance company powered with AI algorithms, the welfare would be the utility obtained by both the client (happy to be insured, and ready to pay a fare for that) and the insurer (whose aim is to collect revenues).

3.4. Principal's Two-Phase Algorithm

(i) Phase 1: Transfer Estimation. For each (s, a), the principal seeks the minimal transfer

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

where Q_a is the agent's optimal state-action value function.

Using batched binary search, the principal estimates $\tau_s^\star(a)$ by offering fixed transfers during batches of episodes and observing the fraction of times the agent selects a when in state s.

(ii) Phase 2: Welfare Optimization. Once the estimates $\hat\tau_s(a)$ satisfy

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

the principal can effectively implement any desired action at s by offering $\hat\tau_s(a)$. She then runs a no-regret RL algorithm (e.g., UCB-VI) on the shifted MDP with effective rewards

$$\tilde r_p(s, a) = r_p(s, a) - \hat\tau_s(a),$$

which preserves welfare. Note that we formulate a simple and theoretical result here, but features can be incorporated while using contextual reinforcement learning algorithms. Pushing things further, we believe that our setting could benefit from Deep RL algorithms.

3.5. Main Result

Now that the method and algorithms are exposed, we provide our main theorem, with the proof given in the Appendix.

Theorem 3.1 (Social Efficiency in Principal–Agent MDPs). Assume the agent satisfies hindsight rationality with exponent $\kappa < 1$ and that the MDP is uniformly ergodic under exploratory policies, ensuring that each state is visited $\Theta(T^\alpha)$ times per batch for some $\alpha > 0$. Suppose the principal chooses

exponents $\alpha, \beta \in (0,1)$ satisfying

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

Then there exists a two-phase principal's algorithm such that with high probability:

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

where $\gamma < 1$ is the regret exponent of the principal's RL algorithm in Phase 2. In particular,

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

so the principal achieves asymptotically optimal social welfare.

4. Diffusion Models as Welfare Maximizing Economic Mechanisms

Finally, we establish a formal connection between diffusion models and economic aggregation, completing the paper's broader theme that modern AI systems can be studied as welfare-oriented mechanisms operating in strategic environments. In our principal–agent MDP framework, the central object is a *welfare criterion* and the design of procedures that implement welfare-improving outcomes despite endogenous behavior and imperfect information. We show that an analogous welfare principle underlies diffusion models: the denoiser learned by standard diffusion training is the *unique* solution to a social planner problem under quadratic welfare (negative squared error), i.e., it implements the Bayesian decision rule that maximizes expected welfare given a noisy signal. Moreover, we provide a micro-founded interpretation in which the same denoising map emerges as an equilibrium aggregation rule in a large-agent economy, where dispersed private information is reported and combined through a simple mechanism. This perspective places diffusion models within the same conceptual toolkit as the rest of the paper: they can be interpreted not only as statistical estimators, but as *economic mechanisms* that aggregate noisy information into a welfare-maximizing decision, thereby connecting generative modeling to incentive-aware design in multi-agent learning systems.

4.1. Diffusion Preliminaries

Let $x_0 \in \mathbb{R}^d$ be drawn from an unknown distribution $p_{\mathrm{data}}(x_0)$. A (variance-preserving) diffusion model defines a forward noising process

$$x_t = \alpha_t x_0 + \sigma_t \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I_d), \tag{3}$$

where $t \in [0,1]$, $\alpha_0 = 1$, $\sigma_0 = 0$, and $\alpha_1 \approx 0$, $\sigma_1 \approx 1$.

Let $\epsilon_\theta(x_t, t)$ be the denoising network trained via the loss

$$\mathcal{L}(\theta) = \mathbb{E}_{t,x_0,\varepsilon}\left\|\varepsilon - \epsilon_\theta(\alpha_t x_0 + \sigma_t \varepsilon, t)\right\|^2. \tag{4}$$

It is classical (for instance in denoising score matching) that the unique minimizer is

$$\epsilon^\star(x_t, t) = \mathbb{E}[\varepsilon \mid x_t]. \tag{5}$$

Bayes' rule under the linear-Gaussian model yields

$$\mathbb{E}[x_0 \mid x_t] = \frac{1}{\alpha_t}\left(x_t - \sigma_t\, \epsilon^\star(x_t, t)\right), \tag{6}$$

which has the great advantage of offering a close form expression.

4.2. A Social Planner Problem

Consider a planner who observes only x_t and chooses a reconstruction $\hat{x}(x_t) \in \mathbb{R}^d$. Define welfare as negative squared error:

$$W(\hat{x}) := -\mathbb{E}\left[\|x_0 - \hat{x}(x_t)\|^2\right]. \tag{7}$$

The planner solves

$$\max_{\hat{x}} W(\hat{x}) \iff \min_{\hat{x}} \mathbb{E}\|x_0 - \hat{x}(x_t)\|^2. \tag{8}$$

Proposition 4.1 (Bayesian Denoiser Maximizes Welfare). The unique solution to the planner's problem (8) is

$$\hat{x}^\star(x_t) = \mathbb{E}[x_0 \mid x_t]. \tag{9}$$

Proof: Fix any measurable $\hat{x}(x_t)$. Write

$$x_0 - \hat{x}(x_t) = \left(x_0 - \mathbb{E}[x_0 \mid x_t]\right) + \left(\mathbb{E}[x_0 \mid x_t] - \hat{x}(x_t)\right).$$

Taking squared norms, expanding, and taking expectations:

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

since the cross-term is zero by the definition of conditional expectation. The expression is minimized iff $\hat{x}(x_t) = \mathbb{E}[x_0 \mid x_t]$ almost surely.

Proposition 4.2 (Score representation of the welfare-optimal estimator). Let p_t denote the marginal density of x_t under the forward process $x_t = \alpha_t x_0 + \sigma_t \varepsilon$, and assume p_t is differentiable with suitable integrability so that $\nabla_{x_t} \log p_t(x_t)$ exists almost surely. Then the welfare-maximizing estimator admits the score form

$$\mathbb{E}[x_0 \mid x_t] = \frac{x_t}{\alpha_t} + \frac{\sigma_t^2}{\alpha_t}\,\nabla_{x_t} \log p_t(x_t). \tag{10}$$

Equivalently, the diffusion-optimal noise predictor satisfies

$$\epsilon^\star(x_t, t) \;=\; -\,\sigma_t\, \nabla_{x_t} \log p_t(x_t). \tag{11}$$

Using (6), we obtain the' folowing corolalry.

Theorem 4.1 (Diffusion Training = Welfare Maximization). The minimizer $\epsilon^\star$ of the diffusion loss $\mathcal{L}$ implements the welfare-maximizing decision rule $\hat{x}^\star(x_t)$ through the linear relationship

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

Thus solving the diffusion training problem is equivalent to solving (8).

4.3. A Micro-Founded Economic Interpretation

We model a continuum of agents indexed by $i \in [0,1]$. There is a hidden state $x_0 \in \mathbb{R}^d$. Each agent observes a signal

$$y_i = x_0 + \eta_i, \qquad \eta_i \sim \mathcal{N}(0, \sigma^2 I), \tag{12}$$

independently across i. A principal chooses a public decision $d \in \mathbb{R}^d$. Each agent has utility

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

and social welfare is

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

Consider a direct mechanism where each agent reports m_i and the principal uses an outcome rule

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

Lemma 4.1 (Truth-Telling in Large Economies). If the principal uses a linear rule $d = A\int_0^1 m_i\, di$, then with a continuum of agents, truth-telling $m_i = y_i$ is a Bayesian Nash equilibrium.

Proof: With a continuum of agents, any individual agent has negligible influence on $\int_0^1 m_i\, di$. Thus each agent treats the outcome as fixed. The report does not change d, so the agent is indifferent among reporting strategies; truthful reporting is therefore an equilibrium (whenever one assumes a favorable tie-breaking, which is a very common assumption in such settings).

Now assume that the principal observes a noisy macro signal generated as

$$x_t = \alpha_t x_0 + \sigma_t \varepsilon \ , \tag{13}$$

as in the diffusion forward process. The principal's welfare problem is again $\max_d -\mathbb{E}\|x_0 - d\|^2$.

Combining Proposition 4.1 with Lemma 4.1 finally leads to the following result.

Proposition 4.3 (Diffusion Drift as Welfare-Maximizing Equilibrium). Under quadratic utilities, symmetric priors, and truth-telling (Lemma 4.1), the mechanism that maximizes expected welfare given noisy macro signal x_t implements

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

which corresponds exactly to the diffusion model's optimal denoiser via

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

Thus the reverse-diffusion update direction is the welfare-maximizing aggregator of noisy private information in a large economy.

This section establishes an equivalence between diffusion models and economic aggregation: the score or denoiser learned by a diffusion model is characterized by both (i) a *social planner optimum* under quadratic welfare and (ii) an *equilibrium mechanism* in a large-agent economy. This provides a principled economic interpretation of diffusion models as devices for aggregating dispersed, noisy information about latent states, and connects them naturally to incentive design in multi-agent learning systems. Although our results linking diffusion models and large-scale principal–agent economics are preliminary, we hope that they offer insights for future research at this intersection

5. Simulations

Reproducible protocol. To ensure reproducibility, we specify the complete interaction protocol and estimation procedure. In each episode k and step h, the principal observes the current state s_h^k, posts a transfer vector $\tau_h^k(\cdot)$, the agent selects an action a_h^k, and both rewards (r_a, r_p) and the next state are realized. Phase 1 estimates $\tau_s^\star(a)$ for each (s, a) using batched tests with fixed transfers and explicit decision thresholds based on empirical action frequencies when s is visited. Phase 2 treats the estimated transfers $\hat{\tau}$ as an implementability oracle and runs a standard no-regret RL routine on the principal's shifted reward

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

while logging welfare using the original (unshifted) rewards since transfers cancel in social welfare. We provide pseudocode (Algorithm 1), all hyperparameters (batch length, confidence thresholds, exploration schedule), and the exact regret metric used in plots.

Fair comparison. All baselines are evaluated under the same interaction budget, the same MDP, the same random seeds, and the same information structure (what the principal observes and when). In particular, any baseline that uses privileged knowledge (e.g., access to the agent's Q-function or to $\tau^\star$) is reported as an oracle upper bound rather than a competing method. This distinction avoids unfair comparisons and makes performance claims interpretable.

Reproducibility statement. We will release a minimal reproducibility package including the environment specification, code to generate all figures, and fixed-seed scripts; all reported curves are averaged over N independent runs with confidence intervals.

We illustrate the role of incentives in a simple principal-

agent Markov decision process with a stateful externality. The environment is a finite-horizon line-world in which an agent chooses among three actions: a fast action that advances the agent quickly but generates significant pollution, a slow action with moderate emissions, and a detour action that is costly to the agent but reduces accumulated pollution. Pollution is an explicit state variable that evolves over time and negatively affects the principal's reward both per step and at the terminal state. The agent, by contrast, values only reaching the goal quickly and does not directly internalize pollution costs. Social welfare is defined as the sum of agent and principal rewards, with monetary transfers canceling out. This setting is a classic illustration of misaligned utilities and distinct roles between a principal and an agent. We show want to show that the incentives allow the players to align their utilities in a favorable way in order to recover global welfare.

We compare two settings. In the first, no transfers are offered and the agent learns via Q-learning using only its own rewards. In the second, the principal offers a simple, state-independent subsidy for taking the detour action. This subsidy is calibrated to offset the agent's private cost of pollution abatement but does not depend on the state or the history of play. In both cases, the agent follows an ϵ-greedy tabular Q-learning algorithm, and performance is evaluated over long-run averages of social welfare and terminal pollution levels.

The results, shown in Figure 1 and Figure 2, demonstrate a clear qualitative difference between the two regimes.

This experiment illustrates a core message of the paper: in multi-agent learning environments with stateful externalities, selfish reinforcement learning can converge to systematically inefficient outcomes, and incentive schemes, even very simple ones, are necessary to align individual learning behavior with social welfare. While the subsidy mechanism used here is intentionally minimal, the observed gains motivate the more structured incentive-compatible mechanisms studied theoretically in the preceding sections.

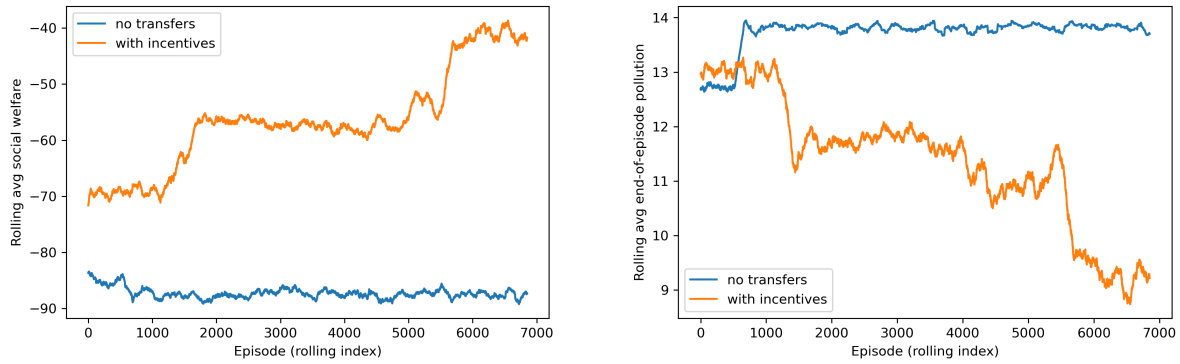

Figure 1. *Effect of incentives in a principal-agent MDP with a stateful externality. Above: rolling average social welfare. Below: rolling average terminal pollution. Introducing a simple subsidy significantly improves welfare by inducing pollution abatement.*

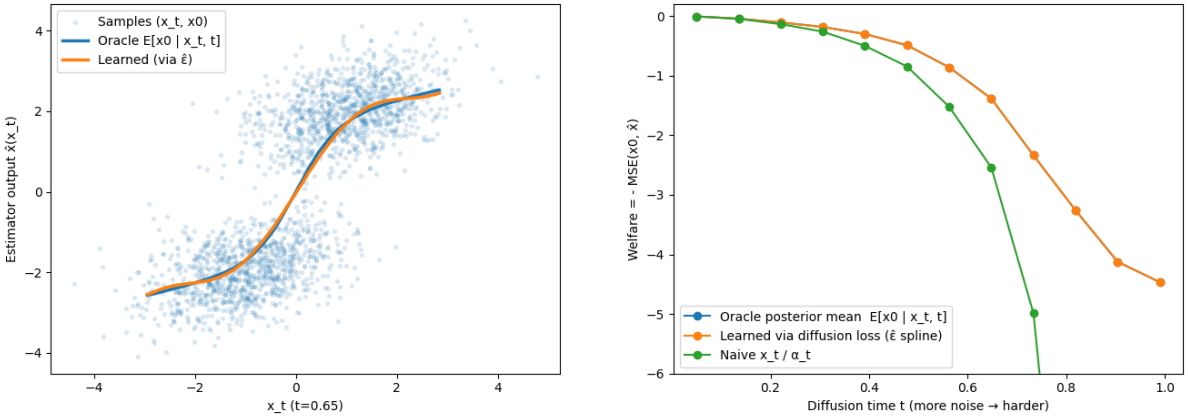

Figure 2. *Diffusion training as welfare maximization. Up: At fixed t, the learned decision rule $\hat{x}(x_t)$ closely matches the oracle posterior mean $\mathbb{E}[x_0 \mid x_t, t]$, illustrating the nonlinear welfare-maximizing aggregator predicted by Section 4. Down: Welfare (negative MSE) as a function of diffusion time t: the diffusion-trained estimator tracks the oracle posterior-mean welfare across noise levels, while the naive baseline collapses in the high-noise regime.*

The welfare-vs.-noise plot empirically validates the equivalence developed in Section 4: under quadratic welfare (negative squared error), the welfare-maximizing decision rule is the Bayesian posterior mean

$$\hat{x}^\star(x_t, t) \;=\; \mathbb{E}[x_0 \mid x_t, t]$$

(as formalized in Proposition 4.1). In the experiment, the estimator obtained via *diffusion training*—learning a denoiser $\hat{\varepsilon}(x_t, t)$ by minimizing the standard denoising (diffusion) loss—induces the corresponding estimate

$$\hat{x}(x_t, t) \;=\; \frac{x_t - \sigma_t\, \hat{\varepsilon}(x_t, t)}{\alpha_t},$$

and achieves essentially the same welfare curve as the oracle posterior mean across diffusion times. This matches the prediction of Theorem 4.1: optimizing the diffusion objective recovers the welfare-optimal aggregator. By contrast, the naive baseline $\hat{x}(x_t, t) = x_t/\alpha_t$ deteriorates sharply as $t \to 1$ (where $\alpha_t \approx 0$), illustrating that the denoiser is not an aesthetic modeling choice but the key ingredient that prevents welfare from collapsing in the high-noise regime.

The *decision-rule plot at fixed* t makes the mechanism interpretation explicit. Because the latent prior on x_0 is nontrivial (here, bimodal), the welfare-optimal mapping $x_t \mapsto \mathbb{E}[x_0 \mid x_t, t]$ is genuinely *nonlinear* (an "S-shaped" aggregation rule), and the learned diffusion estimator tracks this oracle curve closely throughout the relevant range of x_t. This is exactly the object emphasized by the micro-founded mechanism perspective in the paper: under quadratic objectives, the optimal mechanism/aggregator implements the posterior mean (Proposition 4.2), and the optimal diffusion denoiser is simply an alternative parameterization of that same welfare-maximizing rule through the relation above. In other words, the plot does not merely show accurate denoising; it provides a direct visualization of the paper's central message that diffusion updates can be read as an equilibrium, welfare-maximizing information aggregation procedure.

6. Conclusion

We argue that the next generation of AI systems should be studied, and ultimately engineered, as economic mechanisms. When AI mediates prediction, contracting, pricing, and enforcement inside markets and insurance infrastructures, the classical separation between "learning from data" and "designing incentives" collapses. Data become endogenous, behavior responds to policies, and optimization unfolds within a coupled system of interacting learners. Our first contribution is a formal principal-agent framework in a Markov decision process where both players learn over time and where agent actions jointly influence rewards and state transitions. Within this model, we show that transfers, while neutral to welfare ex post, are powerful instruments for welfare alignment ex ante: they reshape the agent's learning problem so that private incentives internalize externalities. The resulting two-phase mechanism provides a clean conceptual template. In Phase 1, the principal identifies the minimal transfers needed to implement desired actions; in Phase 2, the principal leverages these estimates to effectively steer the long-run dynamics of the system. Under mild regularity conditions, this approach achieves sublinear social-welfare regret. Our second contribution is a conceptual and mathematical bridge between economic aggregation and modern generative modeling.

Finally, our simulations illustrate the central message in a transparent environment. Overall, the paper contributes to an emerging view of AI deployment: designing safe and welfare-aligned systems in strategic environments requires co-designing learning dynamics and economic mechanisms.

ORCID

0009-0008-4620-4371 (Nassim Helou)

Abbreviations

AI	Artificial Intelligence
MDP	Markov Decision Process
RL	Reinforcement Learning
SPNE	Subgame-Perfect Nash Equilibrium
MFG	Mean-Field Game
RLHF	Reinforcement Learning from Human Feedback

Author Contributions

Nassim Helou: Conceptualization, Methodology, Formal analysis, Investigation, Writing - original draft, Writing - review & editing.

Conflicts of Interest

The authors declare no conflicts of interest.

Appendix: Proof of Theorem 3.1

Step 1: Action Identifiability via Batches. Fix (s, a). In a batch of length $L = T^\alpha$, assume the process visits state s at least $\Omega(T^\alpha)$ times (uniform ergodicity). If the offered transfer τ satisfies $\tau > \tau_s^\star(a) + T^{-\beta}$, then under the agent's hindsight rationality condition, choosing any $a' \neq a$ in state s incurs regret at least $\Omega(T^\alpha)$, contradicting the regret bound unless misplays occur on at most $O(T^\kappa)$ of the visits. Thus the agent plays a with frequency $1 - O(T^{\kappa-\alpha})$.

Conversely, if $\tau < \tau_s^\star(a) - T^{-\beta}$, then a is suboptimal by at least $T^{-\beta}$, and the agent will choose a at most $O(T^{\kappa-\alpha})$ times. Since $\alpha > \kappa$, these regimes are statistically distinguishable.

Step 2: Batched Binary Search. Repeating this test over $O(\log T)$ batches and shrinking the interval for $\tau_s^\star(a)$ by half each time yields an estimate $\hat{\tau}_s(a)$ with error at most $T^{-\beta}$, provided $\beta/\alpha < 1 - \kappa$ to ensure misclassification probability is $o(1)$.

A union bound across all s, a shows that all estimates satisfy

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

simultaneously with high probability.

Step 3: Implementability of Desired Actions. For any a', we have

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

so a is optimal for the agent whenever the principal offers $\hat{\tau}_s(a)$. By the regret bound, the agent deviates from a only $o(T)$ times in total during Phase 2.

Step 4: Principal's RL and Welfare Regret. In Phase 2, the principal effectively controls the MDP and faces regret $O(T^\gamma \mathrm{polylog} T)$. Phase 1 contributes at most $O(T^\alpha \mathrm{polylog} T)$ regret, and deviations by the agent contribute $O(T^\kappa \mathrm{polylog} T)$. Thus

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

which is $o(T)$ since $\alpha, \gamma, \kappa < 1$. This establishes the theorem.

Proof of Proposition 4.2: Define the rescaled observation $y := x_t/\alpha_t$, so that $y = x_0 + \tilde{\sigma}_t \varepsilon$ with $\tilde{\sigma}_t := \sigma_t/\alpha_t$. For Gaussian corruption, Tweedie's formula (a standard identity in empirical Bayes) gives

$$\mathbb{E}[x_0 \mid y] \ = \ y + \tilde{\sigma}_t^2 \nabla_y \log p_Y(y),$$

where p_Y is the marginal density of Y. Since $y = x_t/\alpha_t$, we have $\nabla_y = \alpha_t \nabla_{x_t}$ and $p_Y(y) = \alpha_t^d p_t(x_t)$, so $\nabla_y \log p_Y(y) = \alpha_t \nabla_{x_t} \log p_t(x_t)$. Substituting and simplifying yields (10). Finally, by definition $\epsilon^\star(x_t,t) = \mathbb{E}[\varepsilon \mid x_t] = (x_t - \alpha_t \mathbb{E}[x_0 \mid x_t])/\sigma_t$, and plugging (10) gives (11).

References

[1] David M Rothschild, Markus Mobius, Jake M Hofman, Eleanor W Dillon, Daniel G Goldstein, Nicole Immorlica, Sonia Jaffe, Brendan Lucier, Aleksandrs Slivkins, and Matthew Vogel. The agentic economy. *arXiv preprint arXiv: 2505.15799*, 2025.

[2] Nicole Immorlica, Brendan Lucier, and Aleksandrs Slivkins. Generative ai as economic agents. *ACM SIGecom Exchanges*, 22(1): 93–109, 2024.

[3] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4: 237–285, 1996.

[4] Richard S Sutton, Andrew G Barto, et al. Reinforcement learning. *Journal of Cognitive Neuroscience*, 11(1): 126–134, 1999.

[5] Roger Guesnerie and Jean-Jacques Laffont. A complete solution to a class of principal-agent probems with an application to the control of a self-managed firm. *Journal of public Economics*, 25(3): 329–369, 1984.

[6] Patrick Bolton and Mathias Dewatripont. *Contract theory*. MIT press, 2004.

[7] Botond Kőszegi. Behavioral contract theory. *Journal of Economic Literature*, 52(4): 1075–1118, 2014.

[8] Eden Saig, Inbal Talgam-Cohen, and Nir Rosenfeld. Delegated classification. *Advances in Neural Information Processing Systems*, 36: 13200–13236, 2023.

[9] Jean-Jacques Laffont and David Martimort. *The theory of incentives: the principal-agent model*. Princeton university press, 2002.

[10] Nivasini Ananthakrishnan, Stephen Bates, Michael Jordan, and Nika Haghtalab. Delegating data collection in decentralized machine learning. In *International Conference on Artificial Intelligence and Statistics*, pages 478–486. PMLR, 2024.

[11] Max Simchowitz and Aleksandrs Slivkins. Exploration and incentives in reinforcement learning. *Operations Research*, 72(3): 983–998, 2024.

[12] Antoine Scheid, Aymeric Capitaine, Etienne Boursier, Eric Moulines, Michael Jordan, and Alain Durmus. Learning to mitigate externalities: the coase theorem with hindsight rationality. *Advances in Neural Information Processing Systems*, 37: 15149–15183, 2024.

[13] Richard Bellman. A markovian decision process. *Journal of mathematics and mechanics*, pages 679–684, 1957.

[14] Martin L Puterman. Markov decision processes. *Handbooks in operations research and management science*, 2: 331–434, 1990.

[15] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR, 2020.

[16] Celestine Mendler-Dünner, Juan Perdomo, Tijana Zrnic, and Moritz Hardt. Stochastic optimization for performative prediction. *Advances in Neural Information Processing Systems*, 33: 4929–4939, 2020.

[17] Gavin Brown, Shlomi Hod, and Iden Kalemaj. Performative prediction in a stateful world. In *International conference on artificial intelligence and statistics*, pages 6045–6061. PMLR, 2022.

[18] James A Mirrlees. The theory of moral hazard and unobservable behaviour: Part i. *The Review of Economic Studies*, 66(1): 3–21, 1999.

[19] Paul Dütting, Tomer Ezra, Michal Feldman, and Thomas Kesselheim. Combinatorial contracts. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 815–826. IEEE, 2022.

[20] Paul Dütting, Tomer Ezra, Michal Feldman, and Thomas Kesselheim. Multi-agent contracts. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 1311–1324, 2023.

[21] Paul Dütting, Michal Feldman, Tomasz Ponitka, and Ermis Soumalias. The pseudo-dimension of contracts. In *Proceedings of the 26th ACM Conference on Economics and Computation*, pages 514–539, 2025.

[22] Antoine Scheid, Daniil Tiapkin, Etienne Boursier, Aymeric Capitaine, Eric Moulines, Michael Jordan, El-Mahdi El-Mhamdi, and Alain Oliviero Durmus. Incentivized learning in principal-agent bandit games. In *International Conference on Machine Learning*, pages 43608–43631. PMLR, 2024.

[23] Junyan Liu, Arnab Maiti, Artin Tajdini, Kevin Jamieson, and Lillian J Ratliff. Learning to incentivize in repeated principal-agent problems with adversarial agent arrivals. *arXiv preprint arXiv: 2505.23124*, 2025.

[24] Antoine Scheid, Etienne Boursier, Alain Durmus, Eric Moulines, and Michael I Jordan. Online decision-making in tree-like multi-agent games with transfers. *arXiv preprint arXiv: 2501.19388*, 2025.

[25] Antoine Scheid, Etienne Boursier, Alain Durmus, Eric Moulines, and Michael I Jordan. Learning contracts in hierarchical multi-agent systems. *arXiv preprint arXiv: 2501.19388v1*, 2025.

[26] Junyan Liu and Lillian J Ratliff. Principal-agent bandit games with self-interested and exploratory learning agents. *arXiv preprint arXiv: 2412.16318*, 2024.

[27] Yuchen Wu, Xinyi Zhong, and Zhuoran Yang. Learning to lead: Incentivizing strategic agents in the dark. *arXiv preprint arXiv: 2506.08438*, 2025.

[28] Aymeric Capitaine, Etienne Boursier, Antoine Scheid, Eric Moulines, Michael I Jordan, El-Mahdi El-Mhamdi, and Alain Durmus. Unravelling in collaborative learning. *arXiv preprint arXiv: 2407.14332*, 2024.

[29] Ronald Harry Coase. The problem of social cost. *The journal of Law and Economics*, 56(4): 837–877, 2013.

[30] Steven G Medema. The coase theorem at sixty. *Journal of Economic Literature*, 58(4): 1045–1128, 2020.

[31] Joseph Farrell. Information and the coase theorem. *Journal of Economic Perspectives*, 1(2): 113–129, 1987.

[32] Tatyana Deryugina, Frances Moore, and Richard SJ Tol. Environmental applications of the coase theorem. *Environmental Science & Policy*, 120: 81–88, 2021.

[33] Shiliang Zuo. New perspectives in online contract design: Heterogeneous, homogeneous, non-myopic agents and team production. *arXiv preprint arXiv: 2403.07143*, 2024.

[34] Jakub Tłuczek, Victor Villin, and Christos Dimitrakakis. Fair contracts in principal-agent games with heterogeneous types. *arXiv preprint arXiv: 2506.15887*, 2025.

[35] Dima Ivanov, Paul Dütting, Inbal Talgam-Cohen, Tonghan Wang, and David C Parkes. Principal-agent reinforcement learning: Orchestrating ai agents with contracts. *arXiv preprint arXiv: 2407.18074*, 2024.

[36] Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pages 157–163. Elsevier, 1994.

[37] Ann Nowé, Peter Vrancx, and Yann-Michaël De Hauwere. Game theory and multi-agent reinforcement learning. In *Reinforcement learning: State-of-the-art*, pages 441–470. Springer, 2012.

[38] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pages 321–384, 2021.

[39] Yaodong Yang and Jun Wang. An overview of multi-agent reinforcement learning from game theoretical perspective. *arXiv preprint arXiv: 2011.00583*, 2020.

[40] Jean-Michel Lasry and Pierre-Louis Lions. Mean field games. *Japanese journal of mathematics*, 2(1): 229–260, 2007.

[41] Alain Bensoussan, Jens Frehse, Phillip Yam, et al. *Mean field games and mean field type control theory*, volume 101. Springer, 2013.

[42] Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. Learning mean-field games. *Advances in neural information processing systems*, 32, 2019.

[43] Leo Widmer, Jiawei Huang, and Niao He. Steering no-regret agents in mfgs under model uncertainty. *arXiv preprint arXiv: 2503.09309*, 2025.

[44] Roger E Kirk. Experimental design. *Sage handbook of quantitative methods in psychology*, pages 23–45, 2009.

[45] Paul D Berger, Robert E Maurer, and Giovana B Celli. *Experimental design*. Springer, 2018.

[46] Walter F Federer. *Experimental design*, volume 81. LWW, 1956.

[47] Anthony C Atkinson. Optimal design. *Wiley StatsRef: Statistics Reference Online*, pages 1–17, 2014.

[48] Peter Goos, Bradley Jones, and Utami Syafitri. I-optimal design of mixture experiments. *Journal of the American Statistical Association*, 111(514): 899–911, 2016.

[49] Binghai Wang, Rui Zheng, Lu Chen, Yan Liu, Shihan Dou, Caishuang Huang, Wei Shen, Senjie Jin, Enyu Zhou, Chenyu Shi, et al. Secrets of rlhf in large language models part ii: Reward modeling. *arXiv preprint arXiv: 2401.06080*, 2024.

[50] Zhichao Wang, Bin Bi, Shiva Kumar Pentyala, Kiran Ramnath, Sougata Chaudhuri, Shubham Mehrotra, Xiang-Bo Mao, Sitaram Asur, et al. A comprehensive survey of llm alignment techniques: Rlhf, rlaif, ppo, dpo and more. *arXiv preprint arXiv: 2407.16216*, 2024.

[51] Jiayi Fu, Xuandong Zhao, Chengyuan Yao, Heng Wang, Qi Han, and Yanghua Xiao. Reward shaping to mitigate reward hacking in rlhf. *arXiv preprint arXiv: 2502.18770*, 2025.

[52] Heyang Zhao, Chenlu Ye, Quanquan Gu, and Tong Zhang. Sharp analysis for kl-regularized contextual bandits and rlhf. *arXiv preprint arXiv: 2411.04625*, 2024.

[53] Antoine Scheid, Etienne Boursier, Alain Durmus, Michael I Jordan, Pierre Ménard, Eric Moulines, and Michal Valko. Optimal design for reward modeling in rlhf. *arXiv preprint arXiv: 2410.17055*, 2024.

[54] Margaret A Meyer and John Vickers. Performance comparisons and dynamic incentives. *Journal of political economy*, 105(3): 547–581, 1997.

[55] Jorge Barrera and Alfredo Garcia. Dynamic incentives for congestion control. *IEEE Transactions on Automatic Control*, 60(2): 299–310, 2014.

[56] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553): 436–444, 2015.

[57] Numa Dhamani and Maggie Engler. *Introduction to generative AI*. Simon and Schuster, 2026.

[58] Richard N Cooper. Economic interdependence and coordination of economic policies. *Handbook of international economics*, 2: 1195–1234, 1985.

[59] Iain M Johnstone and D Michael Titterington. Statistical challenges of high-dimensional data, 2009.

[60] Abhinay Muthoo. *Bargaining theory with applications*. Cambridge University Press, 1999.

[61] Robert Powell. Bargaining theory and international conflict. *Annual Review of Political Science*, 5(1): 1–30, 2002.

[62] Ingolf Ståhl. *Bargaining theory*. PhD thesis, Stockholm School of Economics, 1973.

[63] Fidan Boylu, Haldun Aytug, and Gary J Koehler. Principal–agent learning. *Decision Support Systems*, 47 (2): 75–81, 2009.

[64] Takuya Ito, Murray Campbell, Lior Horesh, Tim Klinger, and Parikshit Ram. Quantifying artificial intelligence through algorithmic generalization. *Nature Machine Intelligence*, 7(8): 1195–1205, 2025.

[65] Chieh-Hsin Lai, Yang Song, Dongjun Kim, Yuki Mitsufuji, and Stefano Ermon. The principles of diffusion models. *arXiv preprint arXiv: 2510.21890*, 2025.

[66] Vijay Mohan. Automated market makers and decentralized exchanges: a defi primer. *Financial Innovation*, 8(1): 20, 2022.

[67] Lars Erik Zachrisson. *Markov games*, volume 52. Princeton University Press Princeton, NJ, 1964.