Information and Contract Design for Repeated Interactions Between Agents with Misaligned Incentives

Nanda Kishore Sreenivas , Kate Larson
University of Waterloo
{nksreenivas, kate.larson}@uwaterloo.ca

Abstract

We study the consequences of information asymmetries and misaligned incentives in settings with multiple independent agents. We model an interaction between a Sender, who holds vital private information but cannot act, and a Receiver, who must make decisions but is dependent on the Sender's information. We find that the Sender learns an optimal communication strategy that the Receiver reliably acts on. Importantly, this strategy is highly sensitive to the degree of conflict in the agents' rewards and the amount of environmental information the Receiver can already observe. We introduce a mechanism allowing the agents to form linear contracts, where a price is established for the information. We demonstrate that the Sender learns to use these payment structures to improve its rewards, though this comes at a cost of "fairness" between agents as the Sender is able to extract much of the Receiver's surplus. This raises questions about fairness, contract design, and learning in the context of multi-agent systems.

1 Introduction

Agents are often faced with the challenge of making decisions under incomplete information. In many real-world scenarios, better-informed agents exist who can provide valuable information without direct intervention. If interests are fully aligned, the informed party can disclose everything it knows and the decision-maker's action will deliver the best outcome for both. However, in practice, this alignment rarely holds. This misalignment introduces a strategic dimension to information sharing, where the information-rich agent may selectively disclose information to guide the decision-maker's choices and steer outcomes toward its own objectives.

Online marketplaces exemplify this dynamic [Liu *et al.*, 2021; Zhang and Nault, 2024]. While the platform possesses a global view of demand through aggregated data, the sellers control product selection and pricing. This creates a complex incentive structure: the platform seeks to steer seller behavior via strategic information disclosure to maximize commissions, yet its broader goals, like market-wide volume and variety, often clash with individual seller interests. This friction allows the platform to monetize its information advantage, offering paid recommendations that sellers purchase to refine their competitive strategy. Similar dynamics are also seen in ride-sharing and navigation apps. Recent work argues that multi-agent LLM systems are best understood through the lens of principal-agent theory [Rauba *et al.*, 2026], precisely because information asymmetry and misaligned incentives emerge naturally in these settings too.

In this paper we explore the tensions that arise in such settings. In particular, we propose a model where there is an information-rich *Sender* agent and an action-taking, but less informed *Receiver* agent. While both agents are cumulative-reward maximizers, their interests may be misaligned. While this type of interaction has been well studied in the one-shot setting through the Bayesian persuasion literature [Kamenica and Gentzkow, 2011; Kamenica, 2019], our focus is on sequential decision-making, where the Receiver must take a series of actions over time without full access to the underlying state. The Sender cannot act directly but can influence the Receiver by sending signals. Before any observations are made, the Sender commits to a signalling policy—a probabilistic mapping from the true state of the world (captured by all payoff-relevant information) to the set of signals. The Receiver then uses this signal and its own partial observation to select actions. The rewards for both agents depend on the true state and the Receiver's chosen action.

We further augment the model through the use of linear contracts. The Sender proposes a contract that stipulates the signalling policy along with the return share it seeks from the Receiver. If the contract is accepted, the agreed-upon fraction of the Receiver's reward is transferred to the Sender. This allows the Sender to monetize its information advantage.

We conduct experiments to explore the behavior of the learning agents by varying (i) the degree of alignment between the agents' incentives, and (ii) the quality of the Receiver's own observations. Our key findings are as follows:

- The Sender learns effective signalling policies, and the Receiver learns to leverage these signals to improve its decision-making.
- When incentives are misaligned, the influence of the signalling policy shifts depending on the quality of the Receiver's private observations: it favors the Sender's

interests when the Receiver's observations are poor, and vice versa.

- When allowed to price its signalling policy through contracts, the Sender achieves higher cumulative rewards by learning to extract value from the Receiver. The pricing is also sensitive to the Receiver's observation quality, with higher prices in low-information settings.

A full version of the paper, containing additional experimental results and complete derivations, is available online [Sreenivas and Larson, 2026].

1.1 Related Work

Our research incorporates three distinct, yet related fields, Bayesian persuasion, contract theory, and multi-agent reinforcement learning, to examine and understand how optimal communication and incentive alignment occur between agents whose rewards are potentially misaligned. In this section, we review key related research in these areas.

The foundational framework for our information design component is Bayesian Persuasion [Kamenica and Gentzkow, 2011], which models an informed Sender strategically influencing a Receiver's action based on information held by the Sender concerning the state of the world. Building upon this, recent reinforcement learning literature has explored dynamic Bayesian Persuasion, where interactions unfold in a sequential, Markovian manner. Key findings indicate that while optimal signalling is tractable for myopic Receivers, the problem becomes significantly more complex for far-sighted agents [Gan *et al.*, 2022; Wu *et al.*, 2022]. Furthermore, research has also investigated signalling games where the Sender lacks commitment, resulting in coupled learning dynamics between the agents [Lin *et al.*, 2023]. Our work builds upon these dynamic settings by explicitly introducing reward misalignment, which fundamentally alters the Sender's optimal strategy and the informational efficiency of the system.

To address the challenge of reward misalignment, we integrate concepts from algorithmic contract theory, specifically by incorporating simple linear, payment-based contracts with the commitment-based information design of BP [Duetting *et al.*, 2019]. This allows the Sender to influence the Receiver's utility not only through reliable information transfer but also through direct financial incentives. This mechanism draws a parallel to the use of contracts in mixed-motive multiagent reinforcement learning (MARL) settings, where reward transfers are employed to resolve social dilemmas and enhance cooperation among otherwise self-interested agents [Christoffersen *et al.*, 2023]. Notably, the successful application of contracts in these MARL environments often relies on imposing structure on the large contract space, a methodological parallel we adopt to manage the complexity of the joint information and payment strategy space.

Our work distinguishes itself from the complementary research line concerning information markets [Bergemann *et al.*, 2018]. In information markets, an informed seller offers a menu of signalling policies and prices to a buyer, aiming to maximize revenue; the seller's utility is primarily revenue-driven. In contrast, our model maintains the core characteristic of Bayesian Persuasion: the Sender's utility is directly dependent on the Receiver's final action, not just the price paid for the information. This critical difference ensures that the design of the signalling policy and the contract is fundamentally driven by the operational outcome desired by the Sender, rather than solely by a pure profit motive.

Finally, our research provides insight for the wider field of MARL, particularly by contrasting with the focus on cooperative settings within action-advising and communication literature [Da Silva and Costa, 2019]. While action-advising often involves an expert outside the environment providing advice to an agent [da Silva *et al.*, 2017], the research often, though not always (see [Subramanian *et al.*, 2022]), assumes the advisor is well-intentioned or purely cooperative. By modelling agents with imperfectly aligned rewards, our work provides a necessary bridge to explore the complex dynamics of communication, learning, and incentive design in more realistic, mixed-motive MARL environments [Zhu *et al.*, 2024]. Our findings offer new insights into how structured information and payments can enable constructive coordination even when agents harbour conflicting interests.

2 Model

In this section, we describe the components of our model and introduce the learning objectives.

We study settings where there are two different agents, a *Sender (S)* and a *Receiver (R)*, interacting in some environment. There are some key asymmetries in the problem. Namely,

1. The Sender (S) has an informational advantage over the Receiver (R) in that it can observe more of the environment. S can decide whether or not to share this information with R.
2. The Receiver (R) has agency, in that it can act in the environment. The Sender (S) must rely on the actions of the Receiver.
3. Both R and S obtain rewards from the environment. However, the rewards may be different and may not be fully aligned.

We formalize the model by introducing two interrelated decision problems, one for the Receiver and one for the Sender, which can be viewed as a Markov Game. For ease of understanding, we present the model from the perspective of the two agents separately, coupling them when necessary.

2.1 The Receiver

Define

$$\mathcal{M}_R = (S_R, A_R, P_R, \Omega_R, Re_R, \gamma_R)$$

to be the MDP for R, capturing the state space, the action space, the probability transition function, the observation space, the reward function and discount factor respectively.

The Receiver does not observe the full environment state $s \in S_R$, but instead perceives a projection of it through an observation function $O : S_R \mapsto O_R$, where O_R is the set of possible local observations available to the Receiver. The Receiver's effective observation space is $\Omega_R = O_R \times \Sigma$, where $\Sigma \supseteq A_R$ is a signal or message sent by the Sender, to the Receiver, specifying an action suggestion.

A policy for the Receiver, π_R, is a mapping from its observation space to (a distribution over) actions. That is

$$\pi_R : (O_R, \Sigma) \mapsto \Delta(A_R).$$

The Receiver aims to maximize its expected discounted sum of rewards by following an optimal policy

$$\pi_R^* = \arg\max_{\pi_R} \mathbb{E}\left[\sum_t \gamma_R^t Re_R(s_R^t, \pi_R(o^t, \sigma^t))\right]$$

where $\sigma^t \in \Sigma$ is the signal sent by the Sender, and $o^t = O(s_R^t)$ is the Receiver's local environmental observation at time t.

2.2 The Sender

Define

$$\mathcal{M}_S = (S_S, A_S, P_S, Re_S, \gamma_S)$$

to be the decision problem for Sender S, capturing the state space, action space, probability transition function, reward structure, and discount factor, respectively.

We assume that $S_S = S_R \cup S_\emptyset$. That is, the state space of the Sender and Receiver is the same, except for one special state, $S_\emptyset$, which indicates that S takes an action before R considers any actions. Furthermore, the environmental dynamics are the same as the Receiver, that is, $P_S = P_R$, for all $s \in S_S \cap S_R$. Recall that the Sender is not able to act directly in the same environment as the Receiver. Its actions, instead, take the form of action advice to the Receiver, $\sigma \in \Sigma$.

The Sender's reward function, Re_S depends on the actions taken by the Receiver. In particular,

$$Re_S : S_R \times A_R \mapsto \mathbb{R}$$

where $Re_S(s, a_R)$ is the reward earned by the Sender when the Receiver takes action a_R in state s. Critically, even though the Sender cannot take direct action in the environment of the Receiver, its reward is determined by how it influences the action choice of the Receiver, through the signals it sends.

A policy of the Sender is thus its signalling policy, which takes the form of a *commitment probability*, p with which the Sender commits to informing the Receiver of the best action to take, *from the Receiver's perspective*. With probability $1 - p$ the Sender will recommend an action that is optimal from its perspective. We further assume that the commitment probability is set at the start of an episode, before the Receiver agent takes any actions in the environment, *i.e.*, the Sender's choice of p is its action at the initial state $S_\emptyset$. We can formalize the policy of the Sender as follows:

$$\pi_S(s) = \begin{cases} p \in [0,1] & \text{if } s = S_\emptyset \\ \begin{cases} \pi'_R(s) & \text{with probability } p \\ \pi'_S(s) & \text{with probability } 1-p \end{cases} & \text{if } s \in S_R \end{cases}$$

Note that we assume the Sender has full observability of the environment in which the Receiver acts and, furthermore, is able to compute optimal policies for both the Receiver and itself.[1] We define $\pi_S^{'}$ as the Sender's hypothetical optimal policy *if the Sender could act directly in the environment*. Further, we define $\pi_R^{'}$ as the hypothetical optimal policy from the Receiver's perspective, *i.e.*, using the Receiver's reward structure.

$$\pi_S^{'} = \arg\max_{\pi} \mathbb{E}\left[\sum_t \gamma_S^t Re_S(s_R^t, \pi(s_R^t))\right]$$

$$\pi_R^{'} = \arg\max_{\pi} \mathbb{E}\left[\sum_t \gamma_R^t Re_R(s_R^t, \pi(s_R^t))\right]$$

Considering the Sender's policy, it is clear that the Sender's influence over the Receiver is through the choice of value of p, its commitment probability. The optimal policy of the Sender, π_S^*, is defined as

$$\pi_S^* = \arg\max_{p\in[0,1]} \mathbb{E}[\sum_t \gamma_S^t (pRe_S(s^t, \pi_R^*(o^t, \pi_R^{'}(s^t))) + (1-p)Re_S(s^t, \pi_R^*(o^t, \pi_S^{'}(s^t))))].$$

Critically, the optimal polices of both agents depend on each other, leading best-response dynamics. Determining whether the Receiver and Sender can learn to best respond to each other is a key focus of this work.

This model allows us to answer the following questions:

RQ1: Can the Sender learn optimal signalling policies (p) under different levels of misalignment? Can the Receiver learn to better navigate the environment with the provided information?

RQ2: How does the quality of the Receiver's private observations (O) affect the agents' learned strategies and outcomes?

Optimal Signalling

We derive the optimal signalling policy for the Sender. We denote the episodic return of an agent following policy π as $J(\pi)$, given by:

$$J(\pi) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{T-1} \gamma^t r_{t+1}\right],$$

The expected utility of the Receiver over an episode is:

$$U_R(p) = p\, V_{RR} + (1-p)\, V_{RS}, \quad (1)$$

$$U_S(p) = p\, V_{SR} + (1-p)\, V_{SS} \quad (2)$$

where $V_{ij} = J_i(\pi'_j)$ denotes agent i's episodic return when the Receiver follows policy π'_j, for $i, j \in \{R, S\}$.

Let U_R^0 denote the Receiver's independent episodic reward, which primarily depends on its quality of observations. Then, the Receiver incentive compatibility constraint dictates that:

$$U_R(p) \geq U_R^0 \implies p\, V_{RR} + (1-p)\, V_{RS} \geq U_R^0. \quad (3)$$

$$p_{\min} = \frac{U_R^0 - V_{RS}}{V_{RR} - V_{RS}}. \quad (4)$$

$U_S(p)$ is linear and decreasing in p (because $V_{SS} \geq V_{SR}$ by definition), so the optimal signalling policy is:

$$p^* = \min(\max(p_{\min}, 0), 1) \quad (5)$$

Note that V_{RS} and V_{RR} depend on the degree of misalignment between S and R while U_R^0 varies with the level of Receiver's observability.

[1] We acknowledge that this is a strong assumption. We justify the assumption since it allows us to abstract away some of the learning problems of the Sender and focus on the central problem of information asymmetry and agency.

2.3 Contracts and Information Pricing

We introduce the possibility of the Sender charging for information through the use of "take it or leave it" linear contracts [Duetting *et al.*, 2019]. A contract is an agreement between the Sender and the Receiver where the Receiver agrees to share a portion of their collected reward in exchange for information provided by the Sender. While contracts can take many forms, linear contracts, captured by a single parameter, $c \in [0, 1]$, have been shown to be very powerful and robust in settings with uncertainty [Duetting *et al.*, 2019; Carroll, 2015]. This gives rise to several questions:

RQ3: Can the Sender learn how to use contracts effectively?

RQ4: How does the use of contracts impact the Receiver's accumulated reward?

To explore these questions we expand the action space of the Sender and Receiver. The Sender announces $\langle p, c\rangle$ to the Receiver, specifying its signalling policy (p) and the reward share $c \in [0, 1]$ it will collect from the Receiver's collected rewards. The Receiver can decide to accept or reject the proposal. If the proposal is rejected, the Receiver must act in the environment with no further interaction from the Sender. If the proposal is accepted, the process is the same as described earlier, except that the reward structure changes. The effective reward structures become

$$Re_S^{'} = Re_S + cRe_R \tag{6}$$

$$Re_R^{'} = (1 - c)Re_R \tag{7}$$

The optimal contract is characterized by:

$$p^* = \begin{cases} p_{\min}, & c < \hat{c}, \\ 1, & c > \hat{c}. \end{cases} \tag{8}$$

where $\hat{c}$ is a threshold that depends on the degree of incentive alignment (for full derivation, see [Sreenivas and Larson, 2026]). For $c < \hat{c}$, transfers are weak and the agents' incentives remain misaligned. So, the interaction reduces to the signalling setting ($p^* = p_{min}$), and there is little surplus to be extracted through contracts. For $c > \hat{c}$, transfers are large enough to align incentives, and full revelation $p^* = 1$ becomes optimal.

3 Experiments

In this section, we present our experimental findings. We ground our work in two settings. The first is a classic recommendation letter scenario from the Bayesian persuasion literature [Dughmi, 2017], while the second is a grid-world environment which allows us to explore the impact that reward alignment has on agents' learned policies.

3.1 Recommendation Letter

In the recommendation letter problem, there are two agents, a professor (Sender) and a recruiter (Receiver) [Dughmi, 2017]. The professor is writing a recommendation letter for their student who is being recruited by the recruiter. The student is either a strong candidate or a weak candidate, and the student quality is known to the professor but not to the recruiter. The recommendation letter serves as a binary signal (recommend/don't recommend) from the professor to the recruiter. The recruiter receives a reward of +1 for hiring a strong student and -1 for hiring a weak student. If no student is hired, both agents receive a reward of 0. The professor receives a reward of +1 whenever their student is hired, regardless of quality. This problem captures the challenges of asymmetric information and misaligned incentives. If the professor (Sender) truthfully reported student quality, the recruiter (Receiver) would only hire strong students. By recommending all strong students and randomly recommending weak students, the professor can increase their expected utility.

Formally, a student's type is drawn from $\{S, W\}$ with prior $Pr(S) = p_0 \leq \frac{1}{2}$. The professor chooses a signalling rule $\sigma : \{S, W\} \to \Delta(\{R, \bar{R}\})$, where R and $\bar{R}$ denote recommend and not recommend, parameterized by

$$p_1 = \Pr(R|S), \qquad p_2 = \Pr(R|W).$$

The optimal signalling policy for the professor is given by:

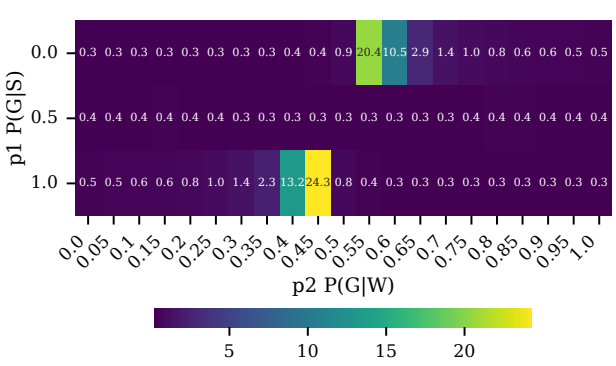
$$p_1^* = 1, \qquad p_2^* = \frac{p_0}{1 - p_0}$$

Consider a specific example, where $p_0 = 1/3$, i.e, a randomly drawn student is strong with probability $\frac{1}{3}$. Without any information about a particular student, the expected utility of the recruiter for hiring a student is negative ($-\frac{1}{3}$), and, therefore, the recruiter will not hire any student. Under perfect information, i.e., the professor recommends a strong student and does not recommend a weak student, the recruiter will hire only the strong students netting an expected utility of $\frac{1}{3}$ for both parties. However, if the professor uses the optimal signalling policy, i.e., recommend a weak student with one half probability, then the professor increases their expected utility to $\frac{2}{3}$. We arrive at the same signalling policy $p^* = \frac{1}{2}$ using our analysis from Section 2.2

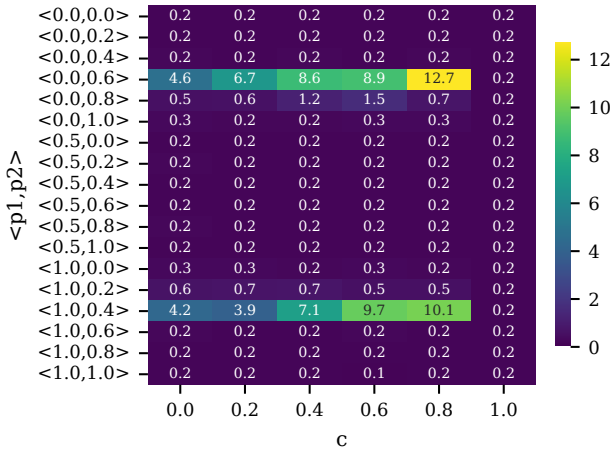
Experimental Setup

We model the learning processes of both agents using multi-armed bandits. The Sender's policy is a tuple $\langle p_1, p_2\rangle$ where p_1 is the probability that the Sender provides a good recommendation if the student is strong ($P(R|S)$), while p_2 is the probability that the Sender provides a good recommendation if the student is weak ($P(R|W)$). Thus, the arms for the Sender's bandit problem correspond to different signalling policies. The Receiver observes the signalling policy of the Sender and the recommendation. This forms the context for a contextual bandit problem with two arms, with one arm corresponding to the hire decision and the other arm corresponding to the not hire decision. Rewards for both the Receiver and Sender are observed after the hire/not hire decision and arm-values are updated.

We ensure there is a finite number of arms for the Sender's bandit problem by discretizing p_1 to the set $\{0, 0.5, 1\}$ and p_2 into 0.05 increments. Each trial consists of 200,000 interactions, and we define an episode to be 50 interactions. At the start of an episode, the Sender commits to a fixed strategy $\langle p_1, p_2\rangle$. The underlying learning algorithm is a discounted-



(a) Average selection percentages of signalling strategies $\langle p_1, p_2 \rangle$



(b) Average selection percentages of contract proposals

Figure 1: Results for Recommendation Letter

UCB algorithm [Garivier and Moulines, 2011]. [2] All of our results are averages computed over 100 trials.

Recommendation Letter Results

We first determine whether agents can learn optimal policies for the recommendation letter problem. We instantiate an instance of the problem where the prior probability that a student is strong is $\frac{1}{3}$.

We first study the case where the Sender learns a signalling policy. The results are shown in Figure 1a. In particular, we notice that the Sender quickly settles on two signalling policies, $\langle 1.0, 0.45 \rangle$ and $\langle 0.0, 0.55 \rangle$, resulting in average rewards of 0.54 for the Sender (professor) and 0.06 for the Receiver (recruiter). We observe that the average rewards are close to the theoretical optimal rewards, and that signalling strategy $\langle 1.0, 0.45 \rangle$ is a close approximation to the optimal strategy. This experiment answers **RQ1**, i.e., the Sender (professor here) learns the optimal signalling policy and the Receiver (recruiter) learns to best respond to it.

In our second set of experiments we studied the impact of the addition of contracts to answer **RQ3**. The Sender's strategy is enriched to be a vector $\langle p_1, p_2, c \rangle$ where contract

[2] We use a discounting rate of 0.95.

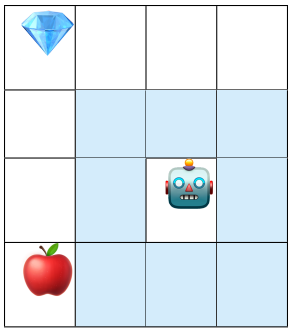


Figure 2: The full environment is a 10×10 grid; for clarity, we display a 4×4 subsection illustrating the key elements. The robot marks the Receiver and objects are shown on the map. The Sender observes the full grid, while the Receiver sees only the highlighted cells (for visibility $v = 1$).

$c \in [0, 1]$ is the fraction of the Receiver's reward that is paid to the Sender if the contract is accepted. We use the same discretization for p_1 as before and discretize both p_2 and c into 0.2 increments, resulting in a $3 \times 6 \times 6 = 108$-arm bandit problem. The Receiver's problem is the same as before, but with an enlarged context (the signalling strategy and the proposed contract), but with the caveat that if the contract is rejected, the Sender sends no signal as to the strength of the student and so the Receiver must make a decision (hire/don't hire) without information.

Figure 1b shows the overall contract proposals made by the Sender. We first observe that the signalling strategy of the Sender quickly converges to the optimal signalling strategies we observed before, but there is more variability around the contract price. While we see that the use of contracts does increase the Sender's average utility to 0.55 while dropping the Receiver's utility to 0.02, we hypothesize that the benefit of contracts is small in this context since there is little surplus to extract from the Receiver.

We note that this is in line with our theoretical analysis. Substituting the expected utility values from this setting, we obtain $\hat{c} = 1$. Thus, the optimal information contract (8) is to implement the same optimal signalling policy as in the non-contract case while charging a cost $c < \hat{c} = 1$.

3.2 Gridworld Experiments

We now explore the possibility of learning signalling policies and contracts in a more complex setting, where we can control both the reward alignment and information asymmetry between the Sender and Receiver.

Environment and Reward Structures

Our environment is shown in Figure 2. It is a simple 10 by 10 grid world with two types of objects: apples and diamonds. The Sender can observe the entire grid, but can not move in the environment. The environment is initialized by placing the Receiver at a cell chosen uniformly at random, and placing one apple and one diamond randomly in the remaining cells. As the objects are collected, they spawn in a new location. The Receiver incurs a movement penalty of -0.05 per step. The Receiver is able to move and collect objects

but has limited observability. To answer **RQ2** we control the observations of the Receiver through a visibility parameter v. This parameter controls the observation radius of the Receiver in terms of the Moore neighbourhood around the Receiver, i.e., the Receiver observes all cells within v steps in each direction. In particular, we consider two scenarios to control for the quality of Receiver observations — low-visibility scenario ($v = 1$, average observability near 10%), and high-visibility scenario ($v = 5$, average observability near 50% of the grid).

While the Receiver can collect both apple and diamond objects, we structure the rewards of the agents so that their interests are potentially misaligned. In particular, the reward functions of the agents are a vector $\langle r_a, r_d \rangle$ where r_a is the reward an agent receives for a collected apple while r_d is the reward per collected diamond. We set the reward vector for the Receiver agent, r_R to be $\langle 1, 0 \rangle$ (i.e. it only cares about collecting apples). We capture the degree of misalignment between the Receiver and the Sender by a parameter θ, the angle between two reward vectors, and set the reward vector of the Sender agent to be $r_S = \langle \cos\theta, \sin\theta \rangle$. Thus, fully aligned agents ($\theta = 0$) have the same reward vectors while fully misaligned agents ($\theta = 180$) have reward vectors $\langle 1, 0 \rangle$ and $\langle -1, 0 \rangle$. We consider $\theta \in \{0°, 30°, 45°, 60°, 90°, 180°\}$, covering aligned, partially divergent, orthogonal, and diametrically opposed objectives. These experiments address **RQ1**: whether the Sender learns optimal signalling policies and how these policies change under varying degrees of reward misalignment.

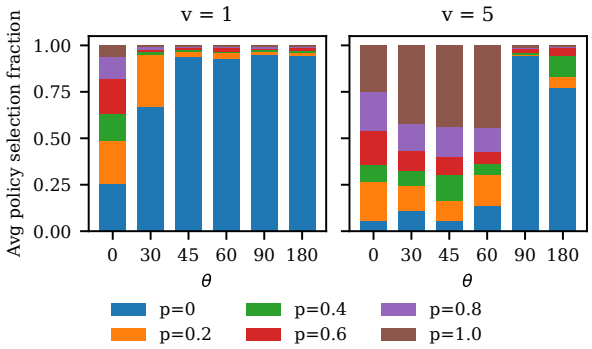


Figure 3: Gridworld: Average selection rates of signalling strategies (p) for different reward alignment (θ) and visibility levels (v)

Learning Processes

As described in Section 2, we assume that the Sender can compute optimal policies for moving in the grid world from its own perspective (π_S') and from the perspective of the Receiver(π_R'). It uses these policies to make action recommendations to the Receiver. Given these policies, we are interested in understanding what signalling and contract policies the Sender learns, and how the Receiver learns how to respond. As in the recommendation letter example, we use bandits as the underlying learning mechanism for the Sender. We consider two settings: the signalling setting without payments and the contract setting as described in Section 2.3. The Sender's policy takes the form of a tuple $\langle p, c \rangle$, $p, c \in [0, 1]$, where p is the probability that the Sender recommends the action according to π_R' (and with probability $1 - p$ it recommends the best action from its perspective, according to π_S'). Parameter c is the cost, which specifies what fraction of the reward collected by the Receiver should be shared with the Sender. For example, if $p = 1$ and $c = 0$ then the Sender always sends optimal action information for the Receiver and asks for no compensation, while if $p = 0$ and $c = 1$ then the Sender always recommends the best action for itself and demands all the Receiver's rewards. We discretize the strategy space into $\{0.0, 0.2, \ldots, 0.8, 1.0\}^2$, resulting in 36 arms. After an arm is selected, the arm's value is updated with the Sender's episodic reward. We use discounted ϵ-greedy with discount factor 0.9 and an ϵ schedule decaying linearly from 1 to 0.05 over the first 75% of training.

The learning problem of the Receiver is more complicated since it must learn whether to accept or reject the contract and, if the contract is accepted, whether to accept the action recommendation or act on its own. We use tabular Q-learning to learn whether or not to accept a contract, where the state space is all possible Sender proposals $\langle p, c \rangle$, the action space is to accept or reject the contract, and the Receiver's episodic reward is used for updates.[3] We use PPO from Stable-Baselines3 to learn whether the Receiver should follow the Sender's recommended action (with $n_{steps} = 256$, batch size 64, $n_{epochs} = 10$). If the action recommendation is not followed, then the Receiver follows a simple heuristic strategy that greedily moves towards the closest observed apple or takes an action at random if no apples are observable. Agents are trained for 2000 episodes and evaluated on 100 test episodes. All reported values represent averages across 10 independent trials, each using a different random seed.

θ	Receiver	Sender	Apple	Diamond
		v = 1		
0	53.72 (0.48)	53.72 (0.48)	74.79 (0.46)	3.88 (0.15)
30	45.24 (0.66)	50.31 (0.87)	65.57 (0.73)	27.71 (2.51)
45	33.24 (0.49)	49.54 (0.51)	53.32 (0.49)	45.13 (0.74)
60	11.00 (0.69)	49.35 (0.82)	31.32 (0.70)	62.36 (1.15)
90	-6.26 (1.31)	47.67 (2.51)	14.59 (1.23)	68.50 (2.38)
180	-6.22 (1.09)	-35.38 (1.06)	14.63 (1.07)	68.33 (2.00)
		v = 5		
0	53.67 (0.54)	53.67 (0.54)	74.74 (0.52)	3.91 (0.18)
30	48.64 (2.26)	43.05 (1.51)	69.78 (2.24)	7.49 (2.62)
45	46.59 (3.41)	32.20 (1.62)	67.81 (3.37)	7.72 (3.35)
60	43.29 (4.35)	18.16 (1.46)	64.64 (4.26)	8.27 (3.01)
90	35.34 (2.49)	-3.42 (1.55)	56.62 (2.33)	17.82 (1.39)
180	31.83 (3.50)	-74.83 (2.96)	53.35 (3.23)	16.45 (2.81)

Table 1: Signalling Setting: Average episodic rewards and the mean number of apples and diamonds collected (SD in parentheses).

Signalling Strategies

First, we look at the case where the Sender does not charge a price for information. The average rewards for both agents and the number of objects collected on average in an episode are shown in Table 1. When the angle between their reward

[3]We use a discounting rate of 0.9, a learning rate of 0.1, and an exploration constant of 0.05.

θ	Receiver	Sender	Apple	Diamond
		v=1		
0	6.31 (3.66)	78.49 (22.90)	64.04 (11.39)	3.22 (0.37)
30	3.44 (4.85)	83.07 (8.40)	64.95 (4.72)	14.71 (6.12)
45	5.52 (3.61)	69.26 (6.90)	58.77 (7.67)	23.01 (11.29)
60	3.76 (3.70)	50.46 (6.50)	46.23 (12.06)	31.30 (15.33)
90	-3.42 (3.20)	34.78 (8.89)	24.63 (12.13)	49.32 (16.81)
180	-8.37 (0.91)	-33.91 (0.81)	13.13 (0.97)	63.07 (2.70)
		v=5		
0	20.13 (5.02)	80.95 (4.75)	71.78 (0.67)	3.35 (0.16)
30	17.45 (3.87)	70.52 (6.00)	68.34 (3.25)	6.01 (3.13)
45	20.78 (3.97)	56.67 (6.35)	67.83 (2.96)	6.03 (2.89)
60	23.55 (4.17)	34.01 (11.41)	63.71 (7.40)	5.80 (3.23)
90	20.15 (4.25)	6.47 (7.33)	63.09 (6.42)	6.53 (3.26)
180	11.26 (0.89)	-55.03 (1.14)	55.08 (10.98)	7.09 (4.55)

Table 2: Contract Setting: Average episodic rewards and the mean number of apples and diamonds collected (SD in parentheses).

vectors, θ is 0, they are fully aligned, and therefore, they are interested in apples only and receive the same reward. As the degree of reward misalignment increases (i.e., as the Sender's preference shifts toward diamonds while the Receiver continues to prefer apples), the outcomes depend strongly on the Receiver's visibility. Under low visibility ($v = 1$), the number of collected diamonds rises with θ, reflecting the Sender's increasing influence on the Receiver's actions. But, under high visibility ($v = 5$), the Receiver collects more apples, as improved observability allows it to rely more on its own policy rather than the Sender's signals. This shift in outcomes is also reflected in the Sender's learned signalling policy (see Figure 3). When the Receiver's observability is low, the Sender typically adopts signalling strategies with $p = 0.0$, meaning that it consistently recommends actions aligned with its own interests rather than the Receiver's. As visibility increases, the Sender's learned strategy uses higher p values (until $\theta < 90$). The exact value of p depends on the degree of reward alignment: the more misaligned the agents, the lower the likelihood of Sender's advice being sampled from the Receiver-optimal policy.

As a baseline, we consider the case where the Receiver navigates the environment without any interaction with the Sender. In the low visibility scenario ($v = 1$), its average episodic reward (over 100 episodes) is -20.81, whereas, in the high visibility scenario, it increases to 11.69. This result addresses **RQ1**, showing that the Sender learns good signalling policies and the Receiver learns to use the signals to improve its rewards. These signalling policies are sensitive to the degree of reward misalignment. This experiment also contributes to answering **RQ2**, showing that the Sender's signalling strategies tend to align more closely with the Receiver's interests when the Receiver's private observations are of higher quality.

Contract Strategies

We now explore whether the Sender learns to effectively use contracts to price the information sent to the Receiver. The average episodic rewards and the mean number of collected objects per episode are presented in Table 2. We also analyze the distribution of contract strategies learned by the Sender for different values of reward misalignment (θ) and Receiver observability (v). Detailed visualizations are provided in the full version of the paper [Sreenivas and Larson, 2026].

We observe a qualitative difference in the strategies learned by the Sender, as they focus on extracting surplus from the Receiver and thus indirectly benefit from the collection of apples. Comparing these results with the signalling setting in Table 1, overall, there is an increase in the Sender's overall utility (at a cost to the Receiver). The contract offered depends on the alignment of the agents. When the two agents are well aligned ($\theta < 90$) the Sender sends information to the Receiver but charges a high amount for it ($c \geq 0.8$). Once $\theta > 90$, the two agents are no longer reasonably aligned and the Sender shifts to a signalling policy with $p = 0$ (i.e. it only sends action advice that is in its own interest, not the Receiver's interest). The contract also drops to $c = 0$ since the Receiver quickly learns that the information provided by the Sender has no value. If the Receiver has better observability ($v = 5$), then it is less reliant on the Sender, which again results in the Sender supplying better-quality information. When the agents are reasonably well aligned, the Sender's proposals promise useful information ($p = 1$) while extracting more than half of the Receiver's rewards ($c = 0.6$).

These results demonstrate that the Sender learns to use contracts strategically, aligning its actions with the Receiver's interests when they are reasonably compatible, and thereby improving its episodic reward compared to the signalling setting. Addressing **RQ4**, the introduction of contracts generally reduces the Receiver's accumulated reward. As in the signalling setting, the learned contract strategies remain sensitive to the Receiver's quality of observations and to the degree of reward misalignment between the two agents.

4 Conclusion

We study repeated interactions between an information-rich Sender agent and a decision-making Receiver agent with misaligned incentives. Through experiments in two different environments, we find that the Sender optimizes its cumulative rewards by learning effective signalling policies to influence the Receiver (**RQ1**). The Receiver learns to use its own partial observations along with the Sender's signals to navigate the environment more effectively, thus improving its cumulative rewards. These learned policies depend on the degree of alignment of their incentives and the quality of Receiver's observations (**RQ2**). Specifically, the Sender's signalling strategies tend to align more closely with the Receiver's interests when the Receiver's observations are of higher quality, and conversely prioritize the Sender's own interests when the Receiver's observability is limited. In addition, we investigate the use of linear contracts that allow the Sender to price its information. Our results show that the Sender can leverage contracts to extract additional surplus (**RQ3**) and contract usage generally reduces the Receiver's total reward (**RQ4**).

Future work could explore extensions to more complex multi-agent environments such as social dilemmas where multiple agents act in the environment but information asymmetry exists. More generally, given the increasing interest in agentic markets and human-AI collaboration, it is imperative to explore alternative mechanisms and contract designs that promote more balanced outcomes for the interacting agents to mitigate the exploitation of low-information participants.

References

[Bergemann *et al.*, 2018] Dirk Bergemann, Alessandro Bonatti, and Alex Smolin. The design and price of information. *American Economic Review*, 108(1):1–48, 2018.

[Carroll, 2015] Gabriel Carroll. Robustness and linear contracts. *American Economic Review*, 105(2):536–563, 2015.

[Christoffersen *et al.*, 2023] Phillip J.K. Christoffersen, Andreas A. Haupt, and Dylan Hadfield-Menell. Get it in writing: Formal contracts mitigate social dilemmas in multi-agent RL. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, page 448–456, Richland, SC, 2023. International Foundation for Autonomous Agents and Multiagent Systems.

[Da Silva and Costa, 2019] Felipe Leno Da Silva and Anna Helena Reali Costa. A survey on transfer learning for multiagent reinforcement learning systems. *J. Artif. Int. Res.*, 64(1):645–703, January 2019.

[da Silva *et al.*, 2017] Felipe Leno da Silva, Ruben Glatt, and Anna Helena Reali Costa. Simultaneously learning and advising in multiagent reinforcement learning. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '17, page 1100–1108, Richland, SC, 2017. International Foundation for Autonomous Agents and Multiagent Systems.

[Duetting *et al.*, 2019] Paul Duetting, Tim Roughgarden, and Inbal Talgam-Cohen. Simple versus optimal contracts. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, EC '19, page 369–387, New York, NY, USA, 2019. Association for Computing Machinery.

[Dughmi, 2017] Shaddin Dughmi. Algorithmic information structure design: A survey. *SIGecom Exch.*, 15(2):2–24, February 2017.

[Gan *et al.*, 2022] Jiarui Gan, Rupak Majumdar, Goran Radanovic, and Adish Singla. Bayesian persuasion in sequential decision-making. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(5):5025–5033, Jun. 2022.

[Garivier and Moulines, 2011] Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In Jyrki Kivinen, Csaba Szepesvári, Esko Ukkonen, and Thomas Zeugmann, editors, *Algorithmic Learning Theory*, pages 174–188, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[Kamenica and Gentzkow, 2011] Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, October 2011.

[Kamenica, 2019] Emir Kamenica. Bayesian persuasion and information design. *Annual Review of Economics*, 11(Volume 11, 2019):249–272, 2019.

[Lin *et al.*, 2023] Yue Lin, Wenhao Li, Hongyuan Zha, and Baoxiang Wang. Information design in multi-agent reinforcement learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[Liu *et al.*, 2021] Zekun Liu, Dennis J. Zhang, and Fuqiang Zhang. Information sharing on retail platforms. *Manufacturing & Service Operations Management*, 23(3):606–619, 2021.

[Rauba *et al.*, 2026] Paulius Rauba, Simonas Cepenas, and Mihaela van der Schaar. Multi-agent systems should be treated as principal-agent problems, 2026.

[Sreenivas and Larson, 2026] Nanda Kishore Sreenivas and Kate Larson. Information and contract design for repeated interactions between agents with misaligned incentives, 2026. https://arxiv.org/abs/2605.11294.

[Subramanian *et al.*, 2022] Sriram Ganapathi Subramanian, Matthew E Taylor, Kate Larson, and Mark Crowley. Multi-agent advisor Q-learning. *Journal of Artificial Intelligence Research*, 74:1–74, 2022.

[Wu *et al.*, 2022] Jibang Wu, Zixuan Zhang, Zhe Feng, Zhaoran Wang, Zhuoran Yang, Michael I. Jordan, and Haifeng Xu. Sequential information design: Markov persuasion process and its efficient reinforcement learning. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, EC '22, page 471–472, New York, NY, USA, 2022. Association for Computing Machinery.

[Zhang and Nault, 2024] Jian Zhang and Barrie R. Nault. Upstream information sharing in platform-based e-commerce with retail plan adjustment. *Decision Support Systems*, 177:114099, 2024.

[Zhu *et al.*, 2024] Changxi Zhu, Mehdi Dastani, and Shihan Wang. A survey of multi-agent deep reinforcement learning with communication. *Autonomous Agents and Multi-Agent Systems*, 38(4), 2024.