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

Anonymous authors
Paper under double-blind review

Abstract

We investigate repeated interactions between a decision-making receiver agent and an informed sender agent who cannot directly influence the environment. Our primary focus is to determine whether both agents can learn strategies to maximize joint reward, even when their incentives are not fully aligned. We illustrate that the sender learns an effective signalling strategy that the receiver learns to act upon. We further explore the use of contracts, where the sender sells its information to the receiver. Our findings show that the sender learns to extract surplus reward from the receiver in such scenarios.

1 Introduction

Agents are often faced with the problem of making decisions and taking action in situations where they have incomplete information. Often better-informed agents exist who can provide valuable information without direct intervention. Examples of such scenarios include many services such as navigation and ride-sharing apps where the platform has access to relevant global information while individual users prioritize their own interests. This can lead to misaligned incentives, where an information-rich agent may strategically share information to guide the decision-maker's choices and steer them towards certain outcomes.

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. We show that the sender learns to strategically disclose information to the receiver and that the receiver learns to act in the environment using information, in the form of signals, provided by the sender. We illustrate that these learned policies depend critically on both the alignment or misalignment of the agents' incentives and on the quality of the receiver's information, independent of the sender.

We also study the use of linear contracts, which allows the sender to charge a price for the information they provide. Sender agents quickly learn to extract significant surplus from receivers, raising interesting questions about contract design, fairness, and information design.

1.1 Related Work

Our work is directly influenced by the literature on *Bayesian Persuasion* Kamenica & Gentzkow (2011); Kamenica (2019). Bayesian Persuasion models scenarios where an informed sender influences a receiver's actions, with both parties' rewards dependent on the true "state of the world" and the receiver's chosen action. The sender commits to a signalling strategy, which maps states to signals. The receiver updates their beliefs based on these signals and acts accordingly. This framework, where the sender optimizes their payoff given the receiver's utility, can be solved efficiently.

Recent reinforcement learning research has explored dynamic Bayesian Persuasion. For example, Gan et al. (2022) showed that optimal signalling strategies are computable for myopic receivers but NP-hard to approximate for far-sighted ones in an MDP setting, while Wu et al. (2022)introduced Markovian Persuasion Processes for influencing a stream of myopic receivers. Lin et al. (2023) further advanced this by considering Markov signalling Games where the sender does not commit to a strategy. Instead, sender and receiver learning processes become coupled, aiming for mutually beneficial outcomes, and allowing for richer signal spaces beyond direct action advice.

Our work builds on these dynamic settings by examining how reward misalignment between the sender and receiver impacts signalling strategies and outcomes. We also integrate simple payment-based contracts, specifically linear contracts Dütting et al. (2019); Duetting et al. (2024), with information design. This exploration provides new insights into learning in environments with imperfectly cooperative agents.

2 Model

We consider a setting with two agents, a *Sender*, and a *Receiver*. The environment critically has 3 factors: 1.) The Sender has an informational advantage over the Receiver, 2.) Only the Receiver has agency, and can act in the environment, and 3.) Their rewards may not be fully aligned. We assume these two agents are engaging and interacting in an environment modelled as an MDP: $\mathcal{M} = \langle S, O, A, P, R^S, R^R \rangle$, where S is the state space, $O \subseteq S$ is the observation space visible to the Receiver and A is the action space of the receiver. The transition function $P : S \times A \to \Delta(S)$ specifies the probability distribution of the next state given the current state and executed action. The reward functions for the sender and receiver agents can be different, and are denoted by R^S and R^R ($R^S, R^R : S \times A \to \mathbf{R}$), respectively.

The Sender constructs two optimal policies π^S and π^R using the two reward structures of $\mathcal{M}$. The policies $\pi^S, \pi^R : S \to \Delta(A)$ specify probability distributions over the action space A given any state S. This will allow the Sender to potentially share action advice to the Receiver since it has a model of the best actions for both agents. The Sender wants to learn a *signalling policy* where it shares information with the Receiver. In particular, we define the signalling policy of the Sender to be a mapping from a state in S, to a probability over an action recommendation sent to the Receiver. In this work, we impose additional structure on the signalling policy by using a *commitment probability* parameter p, where p is the probability that the Sender will recommend the action specified by $\pi^R(s)$, that is, the best action for the Receiver to take in state $s \in S$. This means that with probability $1 - p$ the Sender will recommend it's preferred action $\pi^S(s)$.

The Sender informs the Receiver of its signalling policy before any action-recommendations. That is, the Receiver knows p. Given p and the action-recommendation, the Receiver can decide to follow the advice of the Sender or take an action on its own. Thus it learns a receiving policy, π^O which, given its current observations, p and the proposed action, a, from the Sender, returns a probability distribution over A. Since both agents wish to maximize their expected discounted sum of future rewards, there is a coupling between the two agents' objectives:

$$
\begin{aligned}
p^* &= \arg\max_{p} \sum_t \gamma^t R^S(s_t, \pi^{O,*}(p, o_t)) \\
\pi^{O,*} &= \arg\max_{\pi^O} \sum_t \gamma^t R^R(s_t, \pi^O(p^*, o_t))
\end{aligned}
$$

where $\gamma < 1$ is the discount factor.

2.1 Contracts and Information Pricing

We introduce the possibility of the Sender charging for information through the use of linear contracts Dütting et al. (2019). In theory this should allow the Sender to increase their expected utility by providing more accurate information to the Receiver. We are interested in understanding whether the Sender can learn to price appropriately.

We expand the policy space and process 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

$$R^{S,*} = R^S + cR^R \tag{1}$$

$$R^{R,*} = (1 - c)R^R. \tag{2}$$

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). 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 (sender) to the recruiter (receiver). If the recruiter hires a strong student, then they receive a reward of +1. Otherwise, they receive a reward of -1. The professor receives a reward of +1 if their student is hired, regardless of the 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.

We model this problem 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 ($P(G|S)$), while p_2 is the probability that the sender provides a good recommendation if the student is weak ($P(G|W)$). 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.

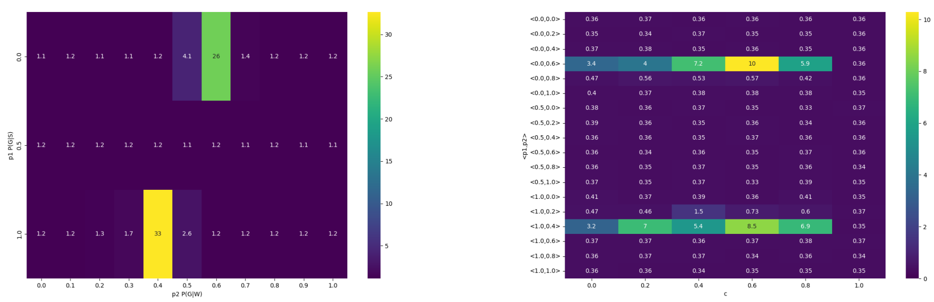
3.1.1 Recommendation Letter Results

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 $\frac{1}{3}$. The theoretically optimal signalling policy of the sender is $\langle 1.0, 0.5 \rangle$.

That is, to truthfully recommend hiring if the student is strong and to recommend hiring half the time if the student is weak. The expected utility of the sender under this strategy is 0.5 while the expected utility of the receiver is 0.0.

We ensure that there is a finite number of arms for the sender's bandit problem by discretizing p_1 and p_2 into 0.1 increments. Each trial consists of 200,000 interactions, and we define an episode to be 50 interactions. In a single episode, the sender commits to a fixed strategy $\langle p_1, p_2 \rangle$. The underlying learning algorithm was a discounted ϵ-greedy algorithm [1]. All of our results are averages computed over 100 trials.

We first study the case where the sender learns a signalling policy. The results are shown in Figure 1a (and Figure 5 in the appendix). In particular, we notice that the sender quickly settles on two contracts, $\langle 1.0, 0.4 \rangle$ and $\langle 0.0, 0.6 \rangle$, resulting in average rewards of 0.577 for the sender (professor) and 0.05 for the receiver (recruiter). We observe that the average rewards are close to the theoretical optimal rewards, and that signalling strategy $\langle 1.0, 0.4 \rangle$ is a close approximation to the optimal strategy. $\langle 0.0, 0.6 \rangle$ is technically the same signaling strategy if the two signals are interchanged. We allowed for random tie-breaking in our experiments whereas the Bayesian persuasion literature typically assumes that ties are always broken in favour of the sender.



(a) The average normalized frequency of signalling strategy $\langle p_1, p_2 \rangle$

(b) The average normalized frequency of contract proposals

Figure 1: Results for Recommendation Letter

In our second set of experiments we studied the impact of the addition of contracts. The sender's strategy is enriched to be a vector $\langle p_1, p_2, c \rangle$ where contract $c \in [0, 1]$ is the fraction of the receiver's reward that is paid to the sender if the contract is accepted. As before, in our experiments we discretized the signalling strategy and contract space (into 0.2 increments) resulting in a 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, while Figures 7 and 6 in the appendix present the contract-specific acceptance rates. 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 0.57 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.

[1]We use a discounting rate of 0.9, and a gradually decaying ϵ from 1 to 0.05

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. 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 receiver is able to move and collect objects but has limited observability. We use a parameter v to control the observability, with v defining the Moore neighbourhood around the receiver. 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 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 $\langle \cos\theta, \sin\theta\rangle$. 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 $\langle 0, 1\rangle$. Table 1 contains the reward vectors we experiment with to understand the impact of reward alignment.

We assume that the sender (since it has full information), 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 will use these policies to make action recommendations to the receiver. Given these policies, we are interested in understanding what signalling and contract policies the sender will learn, and how the receiver will learn how to respond. As in the recommendation letter example, we use bandits as the underlying learning mechanism for the sender. 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 contract, 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 $\{0.0, 0.2, \dots, 0.8, 1.0\}^2$, resulting in 36 arms. After an arm is selected, the arm's value is updated with the sender's episodic reward. We use the discounted ϵ-greedy algorithm with a discounting rate of 0.9, and a decaying schedule for ϵ from 1 to 0.05 over the first 75% of the training horizon.

The learning problem of the receiver is more complicated since 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 DQN to learn whether or not to accept a contract.[2] For learning whether to follow the action recommendation or not, we use PPO. 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.

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

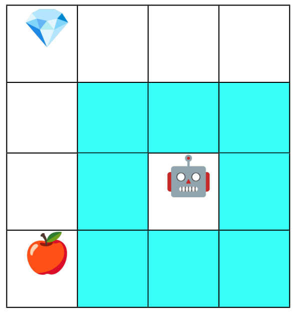


Figure 2: A representative grid world where the cells currently visible to the receiver (visibility, $v = 1$) are shown in blue.

θ (in degrees)	r^S
0	$\langle 1, 0 \rangle$
30	$\langle 0.87, 0.50 \rangle$
45	$\langle 0.71, 0.71 \rangle$
60	$\langle 0.50, 0.87 \rangle$
90	$\langle 0, 1 \rangle$
180	$\langle -1, 0 \rangle$

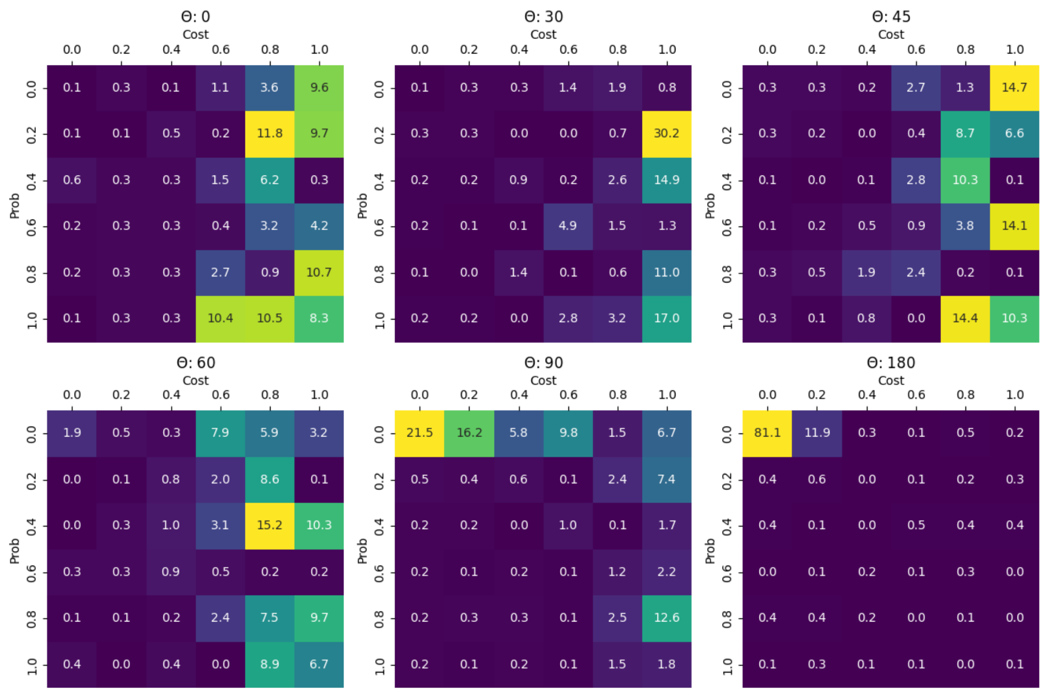
Table 1: The sender reward vector r^S for various values of θ while the receiver reward vector r^R is set as $< 1, 0 >$.

We train the agents for 2000 episodes and each episode consists of 500 timesteps. We consider two scenarios to control for the information gap — low-visibility scenario ($v = 1$, average observability near 10%), and high-visibility scenario ($v = 5$, average observability near 50% of the grid). To account for misaligned incentives, we vary the angle between the sender and receiver reward vectors θ from 0 to 180 degrees. The reward vector for the receiver is $\langle 1, 0 \rangle$ and the corresponding values of r_S can be seen in Table 1. Further, both agents receive a negative reward of -0.05 for each step, as is common in most RL environments. All results reported are averaged over 10 trials, where each trial consists of 2000 episodes.

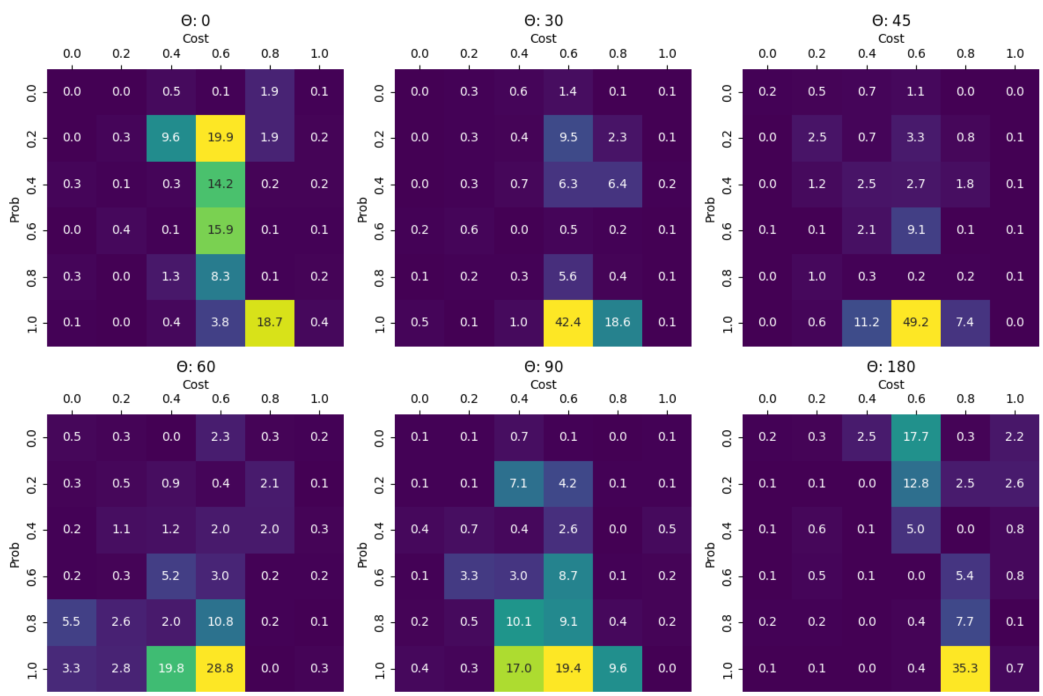
Signalling Strategies: First, we look at the case where the sender does not charge a price for information. The average rewards for both agents and the number of objects collected on average in an episode are shown in Table 2. When the angle between their reward vectors, θ is 0, they are fully aligned, and therefore, they are interested in apples only and receive the same reward. We note that as the difference between the two agents' reward structures increases (i.e. the sender prefers diamonds while the receiver prefers apples), the number of collected diamonds increases. However, if the receiver can observe more of the environment it collects more apples. This is the result of a change of signalling policy on the side of the sender (see appendix, Figure 4a and Figure 4b). If the receiver has low observability the sender learns to use signalling strategies with $p = 0.0$, meaning that it always recommends actions in its own interest, not the receiver. If the receiver can observe more of the environment, then the learned signalling strategy uses higher values of p, though the actual value appears to depend on how aligned or misaligned the agents are.

Contract Strategies: We now explore whether the sender will learn to use contracts to price the information sent to the receiver. The average episodic rewards and the objects collected per episode are shown in Table 3. Similarly, the bandit arm pull frequencies are shown as heatmaps in Figure 3.

We observe a qualitative difference in the strategies learned by the sender, as they focus on extracting surplus from the receiver and thus benefit from the collection of apples. Overall, there is an increase in the receiver's overall utility (at a cost to the sender). The contract offered depends on the alignment of the agents. When the two agents are well aligned ($theta < 90$) the sender sends useful information to the receiver but charges a high amount for it ($c > 0.8$). Once $\theta > 90$, the two agents are no longer well 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. Again we observe that if the receiver can observe the environment, then it is less reliant on the receiver which again results in the receiver supplying better quality information.



(a) Low visibility setting $v = 1$



(b) High visibility setting $v = 5$

Figure 3: Contract Setting | Heatmaps depicting average normalized arm pull frequencies over 10 trials for different values of reward alignment θ. In each map, row values are the commitment probability p while column values are the payment fraction c. Lighter colors indicate higher frequency.

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

Table 2: Information Design Setting | Average episodic rewards for receiver, sender and the average objects collected of each type. These are averaged over 10 trials, each consisting of 2000 episodes. Values in parentheses are the standard deviation.

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

Table 3: Contract Setting | Average episodic rewards for receiver, sender and the average objects collected of each type. These are averaged over 10 trials, each consisting of 2000 episodes. Values in parentheses are the standard deviation.

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 settings, we find that the sender improves its cumulative rewards by learning signalling policies to influence the receiver. The receiver learns to use its own partial observation along with the sender's signal to better navigate the environment. These learned policies depend on the degree of alignment of their incentives and the quality of receiver's observations. Further, we also explore the use of linear contracts, which allow the sender to fix a price for the signals. We observe that the sender learns to extract the surplus from the receiver. Future work could explore other mechanisms and contract designs that enables fairer outcomes for the receiver.

References

Paul Duetting, Michal Feldman, and Inbal Talgam-Cohen. Algorithmic contract theory: A survey, 2024. URL https://arxiv.org/abs/2412.16384.

Shaddin Dughmi. Algorithmic information structure design: a survey. *SIGecom Exch.*, 15 (2):2–24, February 2017. DOI: 10.1145/3055589.3055591. URL https://doi.org/10.1145/3055589.3055591.

Paul Dütting, Tim Roughgarden, and Inbal Talgam-Cohen. Simple versus optimal contracts. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, EC '19, pp. 369–387, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450367929. DOI: 10.1145/3328526.3329591. URL https://doi.org/10.1145/3328526.3329591.

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. DOI: 10.1609/aaai.v36i5.20434. URL https://ojs.aaai.org/index.php/AAAI/article/view/20434.

Emir Kamenica. Bayesian persuasion and information design. *Annual Review of Economics*, 11(Volume 11, 2019):249–272, 2019. ISSN 1941-1391. DOI: https://doi.org/10.1146/annurev-economics-080218-025739. URL https://www.annualreviews.org/content/journals/10.1146/annurev-economics-080218-025739.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, October 2011. DOI: 10.1257/aer.101.6.2590. URL https://www.aeaweb.org/articles?id=10.1257/aer.101.6.2590.

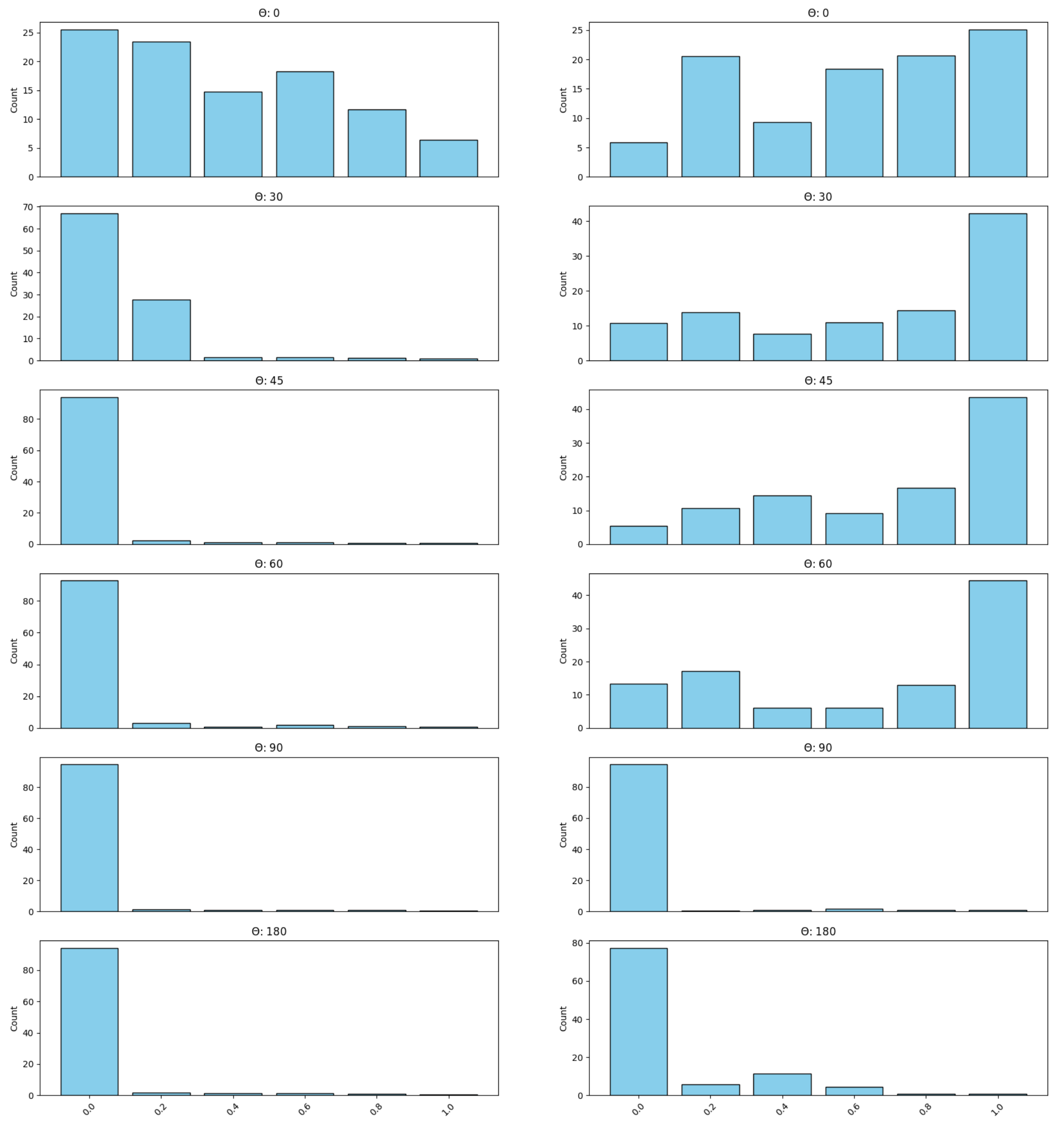
Yue Lin, Wenhao Li, Hongyuan Zha, and Baoxiang Wang. Information design in multi-agent reinforcement learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

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, pp. 471–472, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391504. DOI: 10.1145/3490486.3538313. URL https://doi.org/10.1145/3490486.3538313.

Supplementary Materials

The following content was not necessarily subject to peer review.

Additional Figures for Gridworld



(a) Low visibility setting $v = 1$

(b) High visibility setting $v = 5$

Figure 4: Information Design Setting | Average normalized arm pull frequencies over 10 trials. Each bar represents how many times each value of p was chosen by the sender.

Additional Results for Rec Letter

Here, we list out some additional figures and results for the recommendation letter experiments.

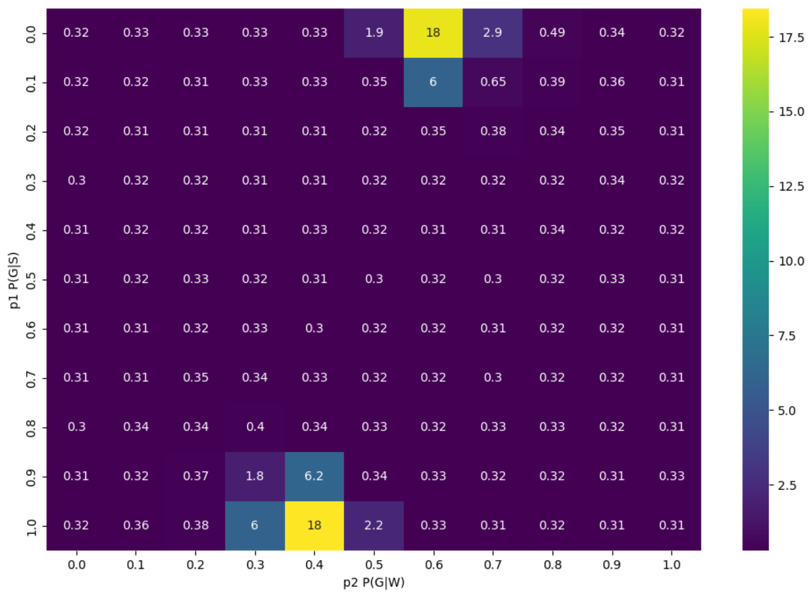


Figure 5: Information Design Setting: The normalized frequency of signalling strategies $\langle p_1, p_2 \rangle$ for the full range of discretized values for p_1 and p_2.

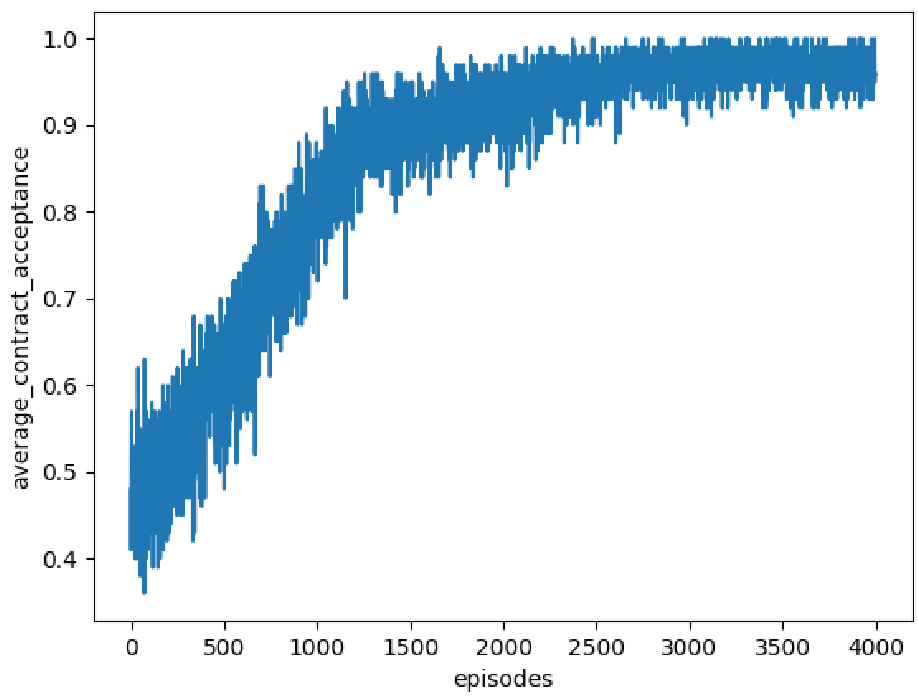


Figure 6: The average contract acceptance rates over 4000 episodes (200,000 interactions). These are averaged over 100 trials.

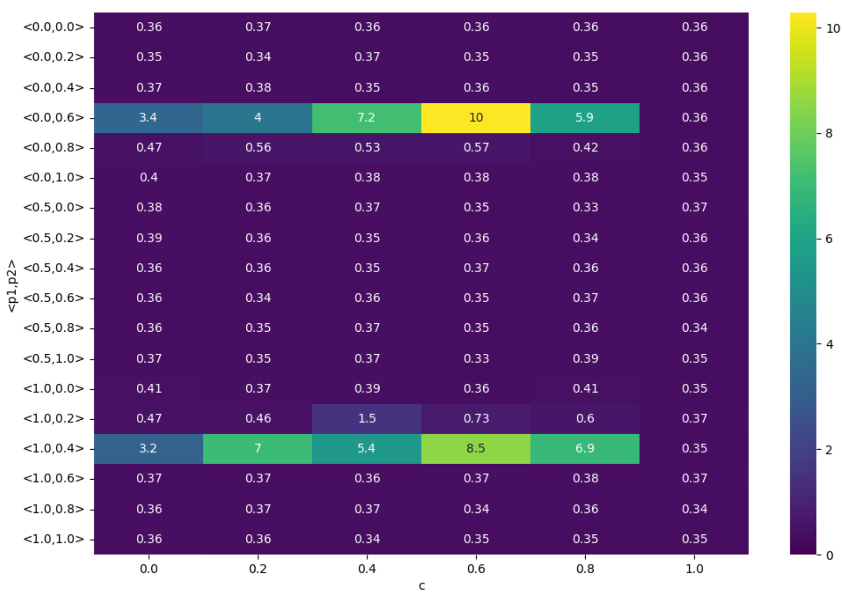


Figure 7: Contract setting: Average contract acceptance rates of all possible contracts averaged over 100 trials with each trial consisting of 4000 episodes.