# Information Design in Multi-Agent Reinforcement Learning

**Yue Lin, Wenhao Li, Hongyuan Zha, Baoxiang Wang**[*]
The Chinese University of Hong Kong, Shenzhen
`linyue3h1@gmail.com, {liwenhao, zhahy, bxiangwang}@cuhk.edu.cn`

## Abstract

Reinforcement learning (RL) is inspired by the way human infants and animals learn from the environment. The setting is somewhat idealized because, in actual tasks, other agents in the environment have their own goals and behave adaptively to the ego agent. To thrive in those environments, the agent needs to influence other agents so their actions become more helpful and less harmful. Research in computational economics distills two ways to influence others directly: by providing tangible goods (*mechanism design*) and by providing information (*information design*). This work investigates information design problems for a group of RL agents. The main challenges are two-fold. One is the information provided will immediately affect the transition of the agent trajectories, which introduces additional non-stationarity. The other is the information can be ignored, so the sender must provide information that the receiver is willing to respect. We formulate the *Markov signaling game*, and develop the notions of signaling gradient and the extended obedience constraints that address these challenges. Our algorithm is efficient on various mixed-motive tasks and provides further insights into computational economics. Our code is publicly available at `https://github.com/YueLin301/InformationDesignMARL`.

## 1 Introduction

Reinforcement learning (RL) studies how a world-agnostic agent makes sequential decisions to maximize utility. It gains increasing popularity and achieves milestone progress in Atari [46, 22], Go [48], Poker [8, 9, 39], video games [2, 6], and bioinformatics [26], economics [59], etc. The canonical RL formulation requires the environment to be stationary, meaning that other agents could not respond to the ego agent's policy by adapting their own policies [50]. The assumption does not hold in real applications. Instead, the ego agent, in general, cannot dictate other agents and needs to influence others so their adaptation becomes more helpful and less harmful.

A substantial subarea of RL, multi-agent reinforcement learning (MARL), investigates the interaction and influence among multiple RL agents when they are placed in a shared environment. It has obtained promising results thanks to the generality and diversity of its settings. Examples are pure cooperative games and pure competitive games. The pure cooperative setting studies agents that work on a consentaneous goal. In this case, influencing other agents is relatively straightforward. The pure competitive setting studies zero-sum games. In this case, influencing other agents is less likely to be practical or even impossible for two-player games. In between these two extreme cases arises the large, more realistic but less charted area of mixed-motive MARL, where influencing other learning agents becomes a main challenge [31, 37, 30].

---

[*]Corresponding to: Baoxiang Wang

Studies in computational economics have distilled different ways of directly influencing a rational, self-interested agent into two types: by providing tangible goods (*mechanism design*) and by providing information (*information design*) [52, 15]. For the former type, tangible goods align perfectly with RL rewards because rewards are natural and utilitarian. Several works in RL design algorithms for the ego agent to use the reward to incentivize other agents [57, 58, 59, 29]. For the latter type, the ego agent must possess information that is helpful to the other agent while not observed by the other agent. The ego agent then sends a message that partially reveals this information in the hope that the other agents respect the message and act in a way that benefits the ego agent. The key observation in information design is that the message needs to be respected, which indicates that the message, apart from benefiting the ego agent, must also benefit the receiving agent.

Two difficulties persist in modeling and solving information design with reinforcement learning. The first challenge is the issue of non-stationarity. The receiver's environment changes as the sender's signaling scheme is updated. If the sender uses policy gradient, it does not consider how modifications to its scheme affect the receiver's learning [17]. On top of this, the signal impacts not only the updating phase but also the sampling phase. In fact, the receiver uses the signals from the sender as part of its input. The signaling scheme will, therefore, directly affect the trajectory generation. This means that specific techniques in mechanism design, such as hyper-gradient, are unsuitable, and new methods need to be formulated [57].

The second challenge lies in how the message can potentially be respected by the receiver. The most general signal space is the set of all state subsets and is exponentially large. An analysis analogous to the *revelation principle* proves that there is an optimal signaling scheme that uses a signal space of the same size as the action space of the receiver, which leads to the classic *obedience constraints* and the linear program formulation of information design [28, 40]. However, under the revelation principle, a signal suggests an exact action, which means the receiver will be dictated by the sender if it respects the message. When both parties are reinforcement learning agents, such want of dictatorship does not build trust and respect between them and instead inevitably drives them to the equilibrium where all signals are ignored. Counter-intuitively, we find that the revelation principle can be removed under the learning context, and one will need to resort to more persuasive signaling schemes and the corresponding update methods.

This paper investigates how an informative ego sender could learn to influence a self-interested receiver with its informational advantage. We propose *Markov signaling game* for mixed-motive communication, where in each timestep, the sender encodes and sends a message to the receiver. After the signaling step, the receiver will act based on the message and its observation. We derive the *signaling gradient* to learn the signaling scheme that addresses the non-stationarity problem. This gradient considers the additional gradient chain from the receiver's policy and is proved to be an unbiased gradient estimation. It agrees with our intuition that the influence of the sender on the behavior of the receiver should be reflected in the gradient term. Based on the signaling gradient, we design the *extended obedience constraints* for incentive compatibility of the signaling scheme.[2] We further provide an approximation of the gradient of such constraints, which solves the second challenge because it is suitable for learning algorithms while preserving the optimum of the ego agent. Information design in MARL is then end-to-end differentiable and learnable with our algorithm.

Numerical experiments on `Recommendation Letter` and `Reaching Goals` demonstrate the efficacy of our approach. Extended discussions are provided for the method and the empirical results.

## 2 Preliminaries and Related Works

**Information Design** For information design, the core insight is to send messages to change the posterior beliefs of the receiver, which persuades it to take actions that benefit the sender [52, 15]. The canonical formulation considers a task that a sender wants to persuade a myopic receiver for one step [28]. The sender and the receiver share a prior distribution $P(s)$ over the state $s$, which affects both the payoffs of the sender $r^i(s, a)$ and the receiver $r^j(s, a)$ (some recent work lifts this assumption [60]). The sender needs to honestly commit its signaling scheme, which is a policy that determines the distribution of signals, to the receiver before the interaction. This is referred to as the commitment assumption.

---

[2]A mechanism is said to be incentive compatible if every participant's optimal strategy, given the strategies of others and the mechanism's rules, leads to an outcome that is desired by the mechanism designer.

The flow of the persuasion process is as follows: (1) The sender commits a signaling scheme to the receiver; (2) The environment generates a state $s$. The sender observes the state $s$ and then samples a message according to the distribution of the committed signaling scheme; and (3) The receiver receives the message, and then calculates a posterior and chooses an optimal action for itself. Given the current state and the receiver's chosen action, the sender and the receiver get rewards from the environment.

Based on an analysis similar to the revelation principle [28, 19], there is an optimal signaling scheme that does not require more signals than the number of actions available to the receiver. Thus it is without loss of generality for the sender to recommend an action directly to the receiver rather than sending a message. From the self-interested receiver's perspective, as long as it believes that the recommended actions are optimal in its posterior belief, it will follow the sender's advice. This kind of constraints of the sender's signaling scheme is called obedience constraints. When such constraints are satisfied, the signaling scheme will be incentive compatible, meaning that the receiver will follow the sender's advice. In this way, the process of information design can be modeled as a constrained optimization problem

$$\max_{\varphi} \mathbb{E}_{\varphi}[\, w^i(s,a)\,], \;\; \text{s.t.} \sum_{s} P(s) \cdot \varphi(a \mid s) \cdot \big[\, w^j(s,a) - w^j(s,a') \,\big] \geq 0, \forall a, a', \tag{1}$$

where $w^i(s,a)$ and $w^j(s,a)$ are the utility functions of the sender and the receiver respectively, $\varphi(a \mid s)$ is the sender's signaling scheme, and the problem is formulated as a linear program.

Information design is applicable to a vast array of real-world scenarios, including voter coalition formation [1], law enforcement deployment [21], price discrimination [4], etc. See [27] for a summary.

**Learning to Communicate**   Communication learning is a significant subarea of MARL. Existing research primarily focus on fully cooperative settings [16, 49, 42]. Among the proposed methods in MARL with communication, DIAL [16] is the closest work to ours. It highlights the importance of the receiver's feedback to the sender, where the receiver updates its critic network and passes the gradient back to the sender's network. This also implies that the sender is assisting the receiver in evaluating the environment. The altruistic design of the sender is appropriate in fully cooperative scenarios, but might not extend to mixed-motive scenarios.

MARL communication under mixed-motive settings was attempted in [25]. The work uses social influence, which is defined via the Kullback-Leibler divergence of the receiver's policy update, to measure how persuasive is the sender's message. It is a pioneer attempt and demonstrates efficacy in several contexts, but the authors observed that "The listener agent is not compelled to listen to any given speaker. Listeners selectively listen to a speaker only when it is beneficial, and influence cannot occur all the time." This describes a lack of the obedience constraints, and motivates us to propose algorithms that are more effective and more general.

Related works on mechanism design, sequential information design, and emergent communication are deferred to Appendix A.

## 3   Markov Signaling Games

Consider a signaling game involving a sender and a receiver. The sender $i$ is assumed to have access to the global state $s \in S$, while the receiver $j$ makes decisions based only on its local observation $o \in O$ and received message $\sigma \in \Sigma$ from $i$. At each timestep $t$, the observation $o_t \in O$ is sampled by the emission function $q : S \to O$, where the information in $o_t$ is a proper subset of the information in $s_t$. We overload the notation and write $o_t \subset s_t$. The receiver's observation at each timestep is common knowledge between the sender and receiver. The sender's *informational advantage* over the receiver at each timestep $t$ is reflected by $s_t - o_t$. By the sender's informational advantage of the problem setting, assume that $\{s_t - o_t\}_{t \geq 0}$ affects $j$'s payoff. This ensures that the sender has information that the receiver wants to know but does not know.

The sender maintains a stochastic signaling scheme $\varphi_{\eta} : S \times O \to \Delta(\Sigma)$, where $\varphi$ is parameterized by $\eta$ and $\Delta(X)$ denotes the set of all random variables on $X$. The receiver's stochastic action policy is denoted as $\pi_{\theta} : O \times \Sigma \to \Delta(A)$, where $A$ is the receiver's action space and $\theta \in \Theta$ is the policy parameter. Without loss of generality, assume that the sender takes no environmental action. The state

transition function $p : S \times A \to \Delta(S)$ and the reward functions $R^i : S \times A \to \mathbb{R}$ ($R^j : S \times A \to \mathbb{R}$) are dependent on $j$'s chosen action and are not dependent on $i$'s message. In this way, $i$ needs to influence $j$ through its signaling scheme, which indirectly affects its long-term payoff expectation.

Then, a Markov signaling game (MSG, Figure 1) is defined as a tuple

$$\mathcal{G} = \left( \, i, j, S, O, \Sigma, A, R^i, R^j, p, q \, \right).$$

At each timestep $t$ in $\mathcal{G}$, $i$ observes a state $s_t \in S$ and $j$'s observation $o_t \in O$ is sampled by $q$, and then $i$ sends a message $\sigma_t$ based on $\varphi_\eta$. Then $j$ takes action $a_t \in A$ based on its policy $\pi_\theta$ and the environment transits to the next state $s' \in S$ according to the transition function $p$. Meanwhile, player $i$ (respectively, $j$) receives the reward $r_t^i$ ($r_t^j$) via the reward function $R^i$ ($R^j$). The agents and the environment repeat this process until the environment terminates the episode.
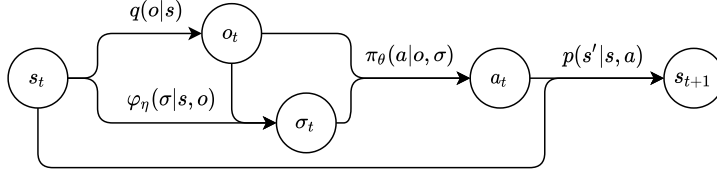


Figure 1: Illustration of the Markov signaling game. The arrows symbolize probability distributions, whereas the nodes denote the sampled variables.

The value functions in MSGs are similar to the value functions in Markov decision processes (MDPs). The sender's state value function $V_{\varphi,\pi}^i(s) = \mathbb{E}_{\varphi,\pi} \left[ G_t^i \mid s_t = s \right]$ defines the expected return of the sender at a state, where $G_t^i = \sum_{k=t}^{\infty} \gamma^{k-t} r_{k+1}^i$. The signal value function $Q$ for player $i$ denotes the same expectation but under an additional condition of sending signal $\sigma$ at the state $s$, as $Q_{\varphi,\pi}^i(s,\sigma) = \mathbb{E}_{\varphi,\pi} \left[ G_t^i \mid s_t = s, \sigma_t = \sigma \right]$. The action value function $U$ is the same expectation that conditions on all variables $s_t, \sigma_t, a_t$, as $U_{\varphi,\pi}^i(s,\sigma,a) = \mathbb{E}_{\varphi,\pi} \left[ G_t^i \mid s_t = s, \sigma_t = \sigma, a_t = a \right]$. Notice the transition function and the reward functions are not dependent on $\sigma$. We then define the marginal action value function $W_{\varphi,\pi}^i(s,a) = \mathbb{E}_{\varphi,\pi} \left[ G_t^i \mid s_t = s, a_t = a \right] = U_{\varphi,\pi}^i(s,\sigma,a)$. The receiver's value functions are similarly defined, by replacing $G_t^i$ with $G_t^j$.

Assume that the agents involved are self-interested, i.e., they are focusing solely on optimizing their own payoffs. However, their payoffs are general and are not necessarily the environmental rewards. For instance, the sender may replace its reward function $R^i$ with $(R^i + R^j)$ as its optimization goal, which would result in it "selfishly" optimizing for the social welfare.

MSG is new to the community and the features of MSG are quite distinctive: (1) The sender and receiver are heterogeneous, with the sender having an informational advantage; (2) Their interests may differ and thus they could be in a mixed-motive scenario; (3) The sender's signal does not directly influence anyone's reward. Thus the sender can only influence the receiver's belief to improve its own expected payoffs; (4) The receiver can take actions to directly determine the payoffs for both parties, but lacks sufficient information to estimate the current state and needs to obtain information from the sender.

There are several extensions of the MSGs. For example, (1) One may allow the sender to take environmental actions; (2) The sender may only have a partial observation of the state; (3) There could be multiple senders and receivers. Discussions on these extensions are deferred to Appendix B.

## 4   Method

The fundamental concept of information design is the sender optimizing its payoffs while adhering to the obedience constraints, as referenced in Equation (1). However, persuasion becomes more challenging in sequential and learning scenarios, due to the non-stationarity and the existence of the trivial equilibrium where all messages are arbitrary and are ignored.

The first obstacle is non-stationarity, a recognized issue in MARL, that is especially problematic in communication since it requires bi-level optimization. During the early stages of training, the sender's signaling scheme is mostly random and provides little information, which can quickly break the trust

between mixed-motive agents and lead to the trivial equilibrium. The second difficulty is the want of dictatorship. In the original obedience constraints, the receiver's policy is set to be deterministic by the revelation principle. It means that recommending a wrong action is non-forgiving, and thus suitable for a trial-and-error algorithm. This will also lead to the trivial equilibrium. In this section, we present how these challenges are addressed and how information design is learned.

## 4.1  Signaling Gradient

The proposed signaling gradient is utilized to compute the gradient of the sender's long-term expected payoff w.r.t. its signaling scheme parameters. When calculating this gradient, it explicitly takes into account the chain of the receiver's policy. This helps alleviate the non-stationarity between agents.

In MSGs, the signaling scheme affects the distribution of signals, indirectly impacting the receiver's action, which then determines the payoffs for all agents. Specifically, the sender's expected payoff is expanded as $\mathbb{E}_{\varphi,\pi}\left[V_{\varphi,\pi}^i(s)\right] = \sum\limits_{s,o,\sigma,a} d_{\varphi,\pi}(s) \cdot q(o \mid s) \cdot \varphi_\eta(\sigma \mid s, o) \cdot \pi_\theta(a \mid o, \sigma) \cdot U_{\varphi,\pi}^i(s, \sigma, a)$, where $d_{\varphi,\pi}$ is the state visitation frequency given $\varphi_\eta$ and $\pi_\theta$. Similar to the case of the policy gradient (PG), the relationship between state visitation frequency and $\pi$ cannot be explicitly written. The introduction of communication $\varphi$ further complicates this process as it involves deriving $\nabla_\eta d_{\varphi,\pi}(s)$.

We utilized a method similar to the policy gradient to derive the gradient $\nabla_\eta \mathbb{E}_{\varphi,\pi}\left[V_{\varphi,\pi}^i(s)\right]$ and obtain an unbiased gradient estimation. We call this estimation the *signaling gradient*. The proof of the lemma is deferred to Appendix D.

**Lemma 4.1.** *Given a signaling scheme $\varphi_\eta$ of the sender and an action policy $\pi_\theta$ of the receiver in an MSG $\mathcal{G}$, the gradient of the sender's value function $V_{\varphi,\pi}^i(s)$ w.r.t. the signaling parameter $\eta$ is*

$$\nabla_\eta V_{\varphi,\pi}^i(s) \propto \mathbb{E}_{\varphi,\pi}\left[W_{\varphi,\pi}^i(s, a) \cdot \left[\nabla_\eta \log \pi_\theta(a \mid o, \sigma) + \nabla_\eta \log \varphi_\eta(\sigma \mid s, o)\right]\right]. \tag{2}$$

It is worth noting that $\mathbb{E}\left[W_{\varphi,\pi}^i(s, a)\nabla_\eta \log \pi_\theta(a \mid o, \sigma)\right] \neq 0$ and $W_{\varphi,\pi}^i(s, a)$ takes an action as an input. As a consequence of this derivation, the sender's updated term includes the receiver's policy and action. This result aligns with intuition, as this additional term reflects the sender's consideration of its impact on the receiver.

One may naturally regard the signal as an action and directly apply the policy gradient. If so, one will obtain $\mathbb{E}_{\varphi,\pi}\left[Q_{\varphi,\pi}^i(s, \sigma) \cdot \nabla_\eta \log \varphi_\eta(\sigma \mid s)\right]$. This gradient will then be independent of the receiver's action and is therefore biased.

**Connections to Other MARL Methods**   There are three perspectives to gain insights from the derivation of the signaling gradient. The first perspective is that the signaling gradient can be regarded as policy-based feedback from the receiver instead of value-based feedback in DIAL [16]. The second perspective is that the signaling gradient and LOLA [17] are similar in alleviating the non-stationarity in MARL communication. (More discussions are deferred to Appendix H.4.) Since MSGs consider the coupling decision-making processes of both parties, the signaling gradient involves the sender taking the initiative to consider how to influence the receiver. In contrast, LOLA involves an agent proactively adapting to other people's updates. The third perspective is that the signaling gradient can be seen as a first-order gradient that is absent in LIO [57]. This new gradient chain can be explicitly derived because the signaling scheme directly affects the receiver's sampling phase.

## 4.2  Policy Gradient for the Receiver

From the receiver's perspective, the decision process can be modeled as a partially observable MDP (POMDP), in which its observation is $O \times \Sigma$ in the corresponding MSG. Therefore, the receiver can optimize its payoff expectation $V_{\varphi,\pi}^j(o, \sigma)$ by calculating the gradient $\mathbb{E}_{\varphi,\pi}\left[A_{\varphi,\pi}^j(o, \sigma, a) \cdot \nabla_\theta \log \pi_\theta(a \mid o, \sigma)\right]$, where $A(o, \sigma, a) = Q(o, \sigma, a) - V(o, \sigma)$ is the advantage function [51]. Compared to the receiver in Bayesian persuasion ([28], see Section 4.5.1 for more details), the policy of the receiver in MSGs is now stochastic rather than deterministic, which allows a larger capacity in taking different actions and a larger capacity in decoding the received information.

## 4.3 Extended Obedience Constraints

In the learning algorithm context, we consider the sender's informational advantage on the current state $s$ and investigate the incentive compatibility of the receiver. As an analogous of (1), the prior of such information is then the occupancy measure $d_{\varphi,\pi}(s)$ of the state condition on the current signaling scheme and action policy. The payoff function $w^j(s, a)$ corresponds to the action value function $W^j_{\varphi,\pi}(s, a)$ in Markov signaling games.

It amounts to deciding the signal space $\Sigma$. The revelation principle states that there is an optimal signaling scheme that does not require more signals than the number of actions available to the receiver. If one follows the revelation principle, one would reasonably use $\Sigma = A$, resulting in the following obedience constraints.

$$\sum_s d_{\varphi,\pi}(s) \cdot \varphi_\eta(a \mid s) \cdot \left[ W^j_{\varphi,\pi}(s, a) - W^j_{\varphi,\pi}(s, a') \right] \geq 0, \quad \forall a, a' \in A. \tag{3}$$

These constraints are technically correct. However, in the learning context, having only $|A|$ possible signals means that the receiver either completely follows the suggested action or completely ignores the message. The former happens only if the obedience constraints are satisfied. In sequential interactions, the constraints will, of course, be violated occasionally. However, the dictatorship nature of the signaling scheme fails to sway the receiver without consistent satisfaction with the constraints.

Moreover, consider the trivial equilibrium between the sender and the receiver: The sender does not reveal useful information, and the receiver ignores the message. At this point, neither side could escape from the equilibrium alone. This means it is likely to fail once the learning algorithm converges to the trivial equilibrium. One natural choice is to send the information to the receiver instead of dictating actions. In this way, the receiver, as a learning agent, is more likely to be able to utilize the information and is more likely to respect the message.

Therefore, we extend the obedience constraints to general, continuous signal space $\Sigma$ that describes the state. A common choice is $\Sigma = S$. The following lemma asserts that the extended obedience constraints impose the same optimum as with the obedience constraints.

**Lemma 4.2.** *Given a receiver's observation $o$, the extended obedience constraints (4) in MSGs yield the same optimum as the obedience constraints (3).*

$$\sum_s d_{\varphi,\pi}(s) \cdot \varphi_\eta(\sigma \mid s, o) \cdot \sum_a \left[ \pi_\theta(a \mid o, \sigma) - \pi_\theta(a \mid o, \sigma') \right] \cdot W^j_{\varphi,\pi}(s, a) \geq 0, \tag{4}$$

*for all $\sigma, \sigma' \in \Sigma$.*

The lemma assumes the sender can access the receiver's policy and observation. Otherwise, the sender may use inferring methods to maintain an estimation of that (an example is [35]). This lemma is proved using the same argument in [3]. For convenience, in later sections, the left-hand side of (4) is denoted as $C_\varphi(\sigma, \sigma')$.

## 4.4 Learning Markov Signaling Games

Given a joint policy $\pi$, the self-interested sender attempts to optimize its payoff expectation in an MSG while satisfying the extended obedience constraints. This optimization problem is

$$\max_\eta \mathbb{E}_{\varphi,\pi} \left[ V^i_{\varphi,\pi}(s) \right], \quad \text{s.t.} \quad C_\varphi(\sigma, \sigma') \geq 0, \quad \forall \sigma, \sigma'. \tag{5}$$

Since we are employing a learning-based approach, it is necessary to calculate the gradient $\nabla_\eta C_\varphi(\sigma, \sigma')$. In this way, our method is model-free and does not require the prior knowledge of $P(s)$ (the occupancy measure $d_{\varphi,\pi}(s)$). Unfortunately, when calculating the gradient of the extended obedience constraints, the technique used in the signaling gradient cannot be applied to reveal the dependency of $d_{\varphi,\pi}$ on $\varphi$. Instead, the gradient is estimated using the biased sampling method as below.

$$\nabla_\eta \hat{C}_\varphi(\sigma, \sigma') = \frac{1}{T} \sum_{s_t \in \tau} \left[ \sum_a \left( \pi_\theta(a \mid o_t, \sigma) - \pi_\theta(a \mid o_t, \sigma') \right) \cdot W^j_{\varphi,\pi}(s_t, a) \cdot \nabla_\eta \varphi_\eta(\sigma \mid s_t, o_t) \right],$$
$$\tag{6}$$

where $\tau$ is a sampled trajectory with $T$ timesteps, and $\sigma'$ is randomly sampled, instead of being sampled from the signaling scheme. Despite that $\sigma$ is from the data that is generated from $\varphi$, the purpose of having these $\sigma, \sigma'$ is to find the violation of the $\forall \sigma, \sigma'$ constraints. Therefore in extended obedience constraints $\nabla_\eta \pi_\theta(a \mid o, \sigma) = 0$, which is different from the signaling gradient.

There are various methods available to solve the constrained optimization problem (5) iteratively, e.g., the Lagrangian method, the dual gradient descent method (DGD) [7]. We tested both methods in the experiments and we found that the Lagrangian method has better performance. See Appendix H.5 for the details. Taking the Lagrangian method as an example, The update of the signaling scheme parameters $\eta^{(k)}$ for the $k$-th iteration is

$$\eta^{(k+1)} \leftarrow \eta^{(k)} + \nabla_\eta \mathbb{E}_{\varphi, \pi} \left[ V^i_{\varphi, \pi}(s) \right] + \sum_{\sigma, \sigma'} \lambda_{\sigma, \sigma'} \cdot \nabla_\eta \left( \hat{C}_\varphi(\sigma, \sigma') \right)^-, \tag{7}$$

where $\lambda_{\sigma, \sigma'}$ denotes the non-negative Lagrangian multipliers (predefined as hyperparameters), and $(\cdot)^- = \min\{0, \cdot\}$.

## 4.5 Discussions on Method

### 4.5.1 Lift to the Commitment Assumption

The most controversial but reasonable assumption in information design is the commitment assumption. In Bayesian persuasion (one-to-one persuasion) [28], the sender will commit to a signaling scheme first. The sender determines its signaling scheme before the games start and honestly tells it to the receiver. A justification of the commitment assumption is that, in a repeated game where a long-term sender interacts with a sequence of short-term receivers, the commitment will naturally emerge in equilibria. This is due to the sender's need to establish its reputation for credibility, which is essential for optimizing its long-term payoff expectations [45].

However, the reputation system between the receivers still needs to be well-defined, so the receivers that have previously interacted cannot convey information about the sender to the receivers that will interact later. Without a reputation system, the sender can optimize its payoff by claiming a respectful signaling scheme while taking an exploitable one. Instead, RL allows for organic and repeated interactions between senders and receivers in a given environment, more closely resembling real-world scenarios. This unique feature enables the learning framework to capture policy evolution and better replicate phenomena in human society.

### 4.5.2 Lift to the Revelation Principle

The extended obedience constraints remove the revelation principle analysis from the obedience constraints, thereby reverting the sender's behavior from "action recommending" to "signal sending". In this way, the sender's set of signals becomes redundant. The redundancy renders the signaling scheme more general and amenable to learning-based approaches. Previously, the signaling scheme required a one-to-one mapping to recommend a particular action, where recommending an undesired action can be non-forgiving. With the introduction of redundancy, the sender can now learn many-to-one mappings to refer to the wanted action distribution, which is a more lenient way for a trial-and-error method. This redundancy is similar to other areas of learning algorithms. For example, one could increase the size of the neural network beyond the information theory necessity to better encode and represent the mapping.

### 4.5.3 Far-sighted Receiver

The nature of RL determines that the receiver in MSGs is considering the cumulative reward. This lifts the commitment assumption and evolves the trustworthiness of the sender's signaling scheme. But meanwhile, the cumulative reward formulates that the receiver must be far-sighted, which is different from the common assumption of a myopic receiver in information design. In fact, a far-sighted receiver is inevitable once we lift the commitment assumption.

A far-sighted receiver is, in general, regarded hard in the literature. Gan et al. [18] prove that information design with a far-sighted receiver is NP-hard. One could intuitively see this from the `Recommendation Letter` example (See Appendix F). The HR can deliberately choose not to hire any students, even when the signaling scheme satisfies the obedience constraints, hoping to force

the professor to be more honest (i.e., reveal more information) in the future. The optimality of the receiver's policy is undefined in this setting.

One way to empirically prevent this is to set an additional constraint $\int_{\sigma, \sigma'} C_\varphi(\sigma, \sigma') d\sigma d\sigma' \geq \epsilon$ apart from the obedience constraints $C_\varphi(\sigma, \sigma') \geq 0$ in Equation (5), where $\epsilon > 0$. A more substantial improvement in reward will incentivize the receiver to establish trust in the long run.

### 4.5.4 Hyper Gradient

Our approach shares similarities with LIO (discussed in Appendix A), as both methods allow agents to alter their parameters to indirectly enhance their payoff expectations by influencing others. In their cases, agents achieve influence by offering rewards to others. The main difference between rewarding and signaling is that the former solely impacts others' policy updates (as the gained rewards are used exclusively for updating). In contrast, the latter additionally affects the sample generation.

In methods to incentivize others, the gradients of the receiver's one-step policy update w.r.t. the sender's rewarding network parameters are required to capture the sender's influence on the receiver. This kind of gradient is second-order and can be viewed as a hyper-gradient.

Unlike incentive-based interventions, in communication methods, the sender's outputs are the inputs of the receiver's actor. The sender can achieve influence while generating trajectories. Hence, our primary focus is on studying the first-order gradient of the receiver's policy w.r.t. the sender's signaling parameters, i.e., $\nabla_\eta \pi_\theta(a \mid o, \sigma)$. The second-order gradients can also be computed, as shown in Equation (21) in the appendix. The effect of the hyper gradient is left for future work.

### 4.5.5 Sender's Access to Receiver

In signaling gradient, the updating of the sender's signaling scheme needs to backpropagate through the receiver's policy during training. Accessing the opponent in mixed-motive settings might sound counterintuitive at a glance, but this setting is actually reasonable and technically feasible.

In MARL algorithms, the Centralized Training with Decentralized Execution (CTDE) framework is commonly used [35]. This means that in the simulated environment, data is centralized and accessible during training, which allows us to use all the required quantities. Once the training is completed, the agent no longer has access to other agents.

From the information design perspective, we may consider the receiver as a "dummy" learning agent that is auxiliary in centralized training. Once a sender agent is trained, it finds the (coarse correlated) equilibrium that is no longer connected to specific receivers. The sender is supposed to persuade any rational, self-interested agents in subsequent interactions with them.

Technically, in communication methods, it is common for gradients to pass through other agents during training. We adopted the commonly used Gumbel-Softmax technique in the field of emergent communication research to allow for end-to-end differentiation, which is used to retain the gradients of sampled signals [24, 20].

## 5 Experiments

The method proposed in this paper is validated in `Recommendation Letter` and `Reaching Goals`. The receiver's action policy is implemented by the advantage actor-critic (A2C) [38]. Each curve in the experimental result graphs is drawn with at least $15$ random seeds. All the seeds are included in the results (including those failed ones). Running $4$ seeds with $2$ NVIDIA GeForce RTX 3090, the longest time is `Reaching Goals` with $5 \times 5$ map, which takes up to a day.

Our designed algorithm takes the perspective of a self-interested sender. Therefore, the measure of success for our algorithm primarily depends on the sender's reward $r^i$. However, we also present the curve of $r^i + r^j$ (social welfare) in an attempt to demonstrate that even when the sender is self-interested and the signaling scheme is somewhat deceptive, it does not harm social welfare, and it might even enhance social welfare in some scenarios.

## 5.1 Recommendation Letter

`Recommendation Letter` is a classic example in information design [15, 28]. In this task, a professor will write recommendation letters for a number of graduating students, and a company's human resources department (HR) will receive the letters and decide whether to hire the students. The professor and the HR share a prior distribution of the candidates' quality, with a probability of $1/3$ that the candidate is strong and a probability of $2/3$ that the candidate is weak. The HR does not know exactly what each student's quality is but wants to hire strong students, while the letters are the only source of information. The HR will get a reward of $1$ for hiring a strong candidate, a penalty of $-1$ for hiring a weak candidate, and a reward of $0$ for not hiring. The professor gets a $1$ reward for each hire. In this section, we focus on analyzing the experimental results, while a classic analysis of the three situations is presented in Appendix F.

In the experiments, we primarily compared the performance of policy gradient class algorithms (PG), PG class algorithms with obedience constraints (PGOC), DIAL [16], signaling gradient (SG, our ablation), and signaling gradient with obedience constraints (SGOC, our proposed method) in learning a signaling scheme. More specifically, the PG class algorithm utilized in experiments refers to A2C. Furthermore, SG also employed A2C techniques, including the use of the actor-critic framework, target critic, and advantage function (adapted to $W^i(s, a) - V^i(s)$ in MSGs). The algorithms with obedience constraints are also required to maintain an extra critic for estimating $W^j(s, a)$. We let $\varphi_\eta(\sigma \mid s, o) = \varphi_\eta(\sigma \mid s)$ and $\Sigma = \{0, 1\}$. The performance comparisons are shown in Figure 2 (a-c).
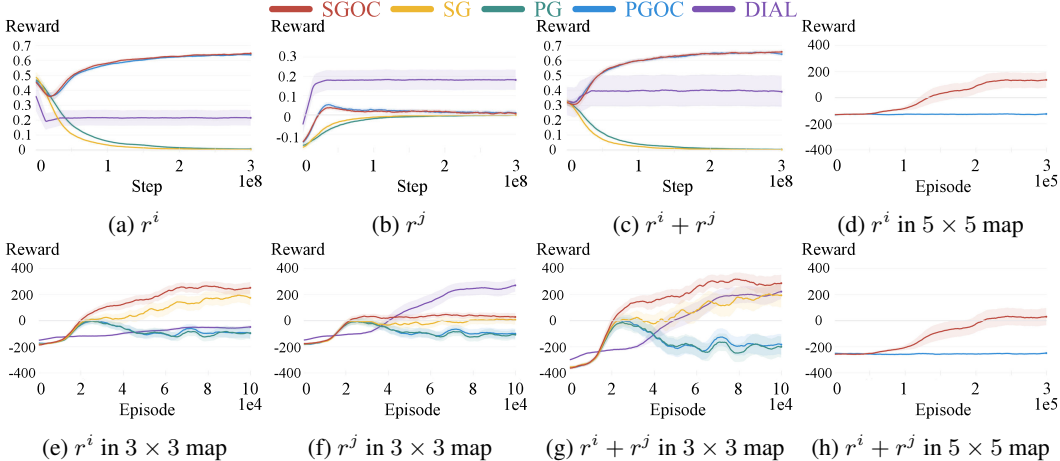


Figure 2: Comparisons of the performance. (a-c) The results of `Recommendation Letter`. (d-h) The results of `Reaching Goals`. The rewards and penalties are amplified by $20$ and $5$ ($12$ and $3.5$) respectively in $3 \times 3$ ($5 \times 5$) map.

The experiment results show that the three examples (two equilibria plus the honest signaling scheme) introduced in Section 1 have emerged. The algorithms with obedience constraints (PGOC and SGOC) reach the third equilibrium (the best for the sender), DIAL reaches the second example (the best for the receiver), and PG and SG reach the first equilibrium (the worst for both parties). Given that the `Recommendation Letter` task is not a problem with interdependent states, it is justifiable to expect that the performance of PGOC and SGOC (PG and SG) would be similar.

## 5.2 Reaching Goals

To reflect the inconsistency of the interests and the information asymmetry, we evaluate the methods in the `Reaching Goals` task. In this challenging task, the sender and receiver have different goals: the sender wants the red apple, while the receiver wants the green apple. At any given time, there is one of each type of apple on the map. Only the receiver is able to move around to reach apples in the grid-world map. The conflict of interest arises because these two types of apples are independently generated at random, and their locations are typically not the same. The receiver will not get a reward for reaching the red apple, but it does not know where the green apple is. The sender, on the other hand, knows the location of both types of apples, but can only get a reward when the receiver reaches

9

the red apple. The sender can only indirectly optimize its payoff by sending messages to influence the receiver's actions. Moreover, this conflict can be exacerbated by designing distance penalties. After each timestep, both agents will get a penalty based on the distance between the receiver and its desired goal. And we set a fixed horizon of $50$ for each episode in this scenario. Example maps of `Reaching Goals` are shown in Figure 6.

To evaluate the efficacy of SGOC, we conducted experiments on maps of $3 \times 3$ and $5 \times 5$, and the results are presented in Figure 2. Let the signal space $\Sigma = S^1$, where $S^1$ is one channel out of three channels (the locations of two apples and the receiver's position) of the image. The receiver can only see its position. The experiments of having a larger $o^j$ are included in Appendix H.7. The smaller the $o^j$, the greater the sender's persuasion ability. From the empirical results, the signaling gradient is shown to be an essential factor in sequential communication scenarios.

### 5.3 Discussions on Experiments

#### 5.3.1 Symmetricity of the Signaling Schemes

An interesting phenomenon is observed in the `Recommendation Letter` experiments: training with different random seeds may result in different pairs of encoders and decoders. In other words, the professor may signal $1$ to indicate a strong student (or recommend this student) in some seeds, while in others, this is signaled by $0$. However, regardless of which case it is, the paired HR can always understand the semantics of the signals (reflected in the evolution of the receiver's policy). Based on the outcomes, the seeds are divided into two parts (A seeds and B seeds).

This phenomenon is reasonable since we do not make any prior assumption about signaling semantics. The symmetric results of `Recommendation Letter` experiments are shown in Figure 3.
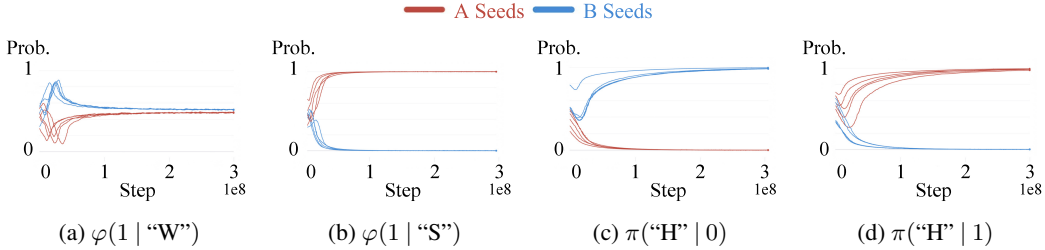


Figure 3: The signaling schemes and the action policy by SGOC. (a) The prob. of signaling $1$ for **W**eak students. (b) The prob. of signaling $1$ for **S**trong students. (c) The prob. of choosing to **H**ire when signaled $0$. (d) The prob. of choosing to **H**ire when signaled $1$.

#### 5.3.2 Honesty of the Sender

As discussed in Section 4.5.3, the obedience constraints can be associated with an additional constraint $\int_{\sigma,\sigma'} C_\varphi^j(\sigma, \sigma') d\sigma d\sigma' \geq \epsilon$. The hyperparameter $\epsilon > 0$ is set to improve the credibility of signaling schemes in practical situations and to cope with the RL capacity of the receiver. Define the honesty metric as $|\varphi(1 \mid \text{"S"}) - \varphi(1 \mid \text{"W"})|$. The experiments show that the larger the $\epsilon$ and the Lagrangian $\lambda$, the more honest the signaling scheme will be, as shown in Figure 8 in the appendix.

## 6 Conclusion

We investigate information design, a substantial and open area, for MARL that discusses mixed-motive communication. Technically, we propose the Markov signaling games to describe the problem and provide its characterizations. We then prove the signaling gradient lemma, which gives an unbiased way to estimate the gradient and update the sender's signaling network. To learn the incentive compatibility of the signaling scheme, we propose extended obedience constraints. The new constraints are more suitable for learning algorithms and practically promote mutually beneficial signaling schemes. The commitment assumption and the revelation principle are lifted by investigating information design with MARL. Experiments and extended discussions are presented to demonstrate the efficacy of our framework and algorithm.

## Acknowledgments

## References

[1] Ricardo Alonso and Odilon Câmara. Persuading voters. *American Economic Review*, 106(11): 3590–3605, 2016.

[2] Kai Arulkumaran, Antoine Cully, and Julian Togelius. Alphastar: An evolutionary computation perspective. In *Proceedings of the genetic and evolutionary computation conference companion*, pages 314–315, 2019.

[3] Dirk Bergemann and Stephen Morris. Information design: A unified perspective. *Journal of Economic Literature*, 57(1):44–95, 2019.

[4] Dirk Bergemann, Benjamin Brooks, and Stephen Morris. The limits of price discrimination. *American Economic Review*, 105(3):921–957, 2015.

[5] Martino Bernasconi, Matteo Castiglioni, Alberto Marchesi, Nicola Gatti, and Francesco Trovò. Sequential information design: Learning to persuade in the dark. In *Advances in Neural Information Processing Systems*, 2022.

[6] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

[7] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[8] Noam Brown and Tuomas Sandholm. Superhuman ai for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018.

[9] Noam Brown and Tuomas Sandholm. Superhuman ai for multiplayer poker. *Science*, 365 (6456):885–890, 2019.

[10] Kris Cao, Angeliki Lazaridou, Marc Lanctot, Joel Z Leibo, Karl Tuyls, and Stephen Clark. Emergent communication through negotiation. *arXiv preprint arXiv:1804.03980*, 2018.

[11] Andrea Celli, Stefano Coniglio, and Nicola Gatti. Private bayesian persuasion with sequential games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1886–1893, 2020.

[12] E Choi, A Lazaridou, and N de Freitas. Compositional obverter communication learning from raw visual input. *arXiv preprint arXiv:1804.02341*, 2018.

[13] Vincent Conitzer and Tuomas Sandholm. Complexity of mechanism design. In *UAI*, 2002.

[14] Alvin W Drake. *Observation of a Markov process through a noisy channel*. PhD thesis, Massachusetts Institute of Technology, 1962.

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

[16] Jakob Foerster, Ioannis Alexandros Assael, Nando De Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 2016.

[17] Jakob Foerster, Richard Y Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with opponent-learning awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 122–130, 2018.

[18] Jiarui Gan, Rupak Majumdar, Goran Radanovic, and Adish Singla. Bayesian persuasion in sequential decision-making. In *AAAI Conference on Artificial Intelligence*, 2022.

[19] Jiarui Gan, Rupak Majumdar, Goran Radanovic, and Adish Singla. Sequential decision making with information asymmetry. In *International Conference on Concurrency Theory*, 2022.

[20] Serhii Havrylov and Ivan Titov. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. *Advances in Neural Information Processing Systems*, 30, 2017.

[21] Penélope Hernández and Zvika Neeman. How bayesian persuasion can help reduce illegal parking and other socially undesirable behavior. *American Economic Journal: Microeconomics*, 14(1):186–215, 2022.

[22] Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, 2018.

[23] Dmitry Ivanov, Iskander Safiulin, Igor Filippov, and Ksenia Balabaeva. Optimal-er auctions through attention. *Advances in Neural Information Processing Systems*, 35:34734–34747, 2022.

[24] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[25] Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, DJ Strouse, Joel Z Leibo, and Nando De Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International Conference on Machine Learning*, pages 3040–3049. PMLR, 2019.

[26] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 2021.

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

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

[29] Raphael Koster, Jan Balaguer, Andrea Tacchetti, Ari Weinstein, Tina Zhu, Oliver Hauser, Duncan Williams, Lucy Campbell-Gillingham, Phoebe Thacker, Matthew Botvinick, et al. Human-centred mechanism design with democratic ai. *Nature Human Behaviour*, 6(10): 1398–1407, 2022.

[30] Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 464–473, 2017.

[31] Joel Z Leibo, Edgar A Dueñez-Guzman, Alexander Vezhnevets, John P Agapiou, Peter Sunehag, Raphael Koster, Jayd Matyas, Charlie Beattie, Igor Mordatch, and Thore Graepel. Scalable evaluation of multi-agent reinforcement learning with melting pot. In *International Conference on Machine Learning*, 2021.

[32] David Lewis. *Convention: A philosophical study*. John Wiley & Sons, 2008.

[33] Changxi Li, Fenghua He, Hongsheng Qi, Daizhan Cheng, Longbiao Ma, Yonghong Wu, and Shuo Chen. Potential games design using local information. In *2018 IEEE Conference on Decision and Control*, pages 1911–1916. IEEE, 2018.

[34] Na Li and Jason R Marden. Designing games for distributed optimization. *IEEE Journal of Selected Topics in Signal Processing*, 7(2):230–242, 2013.

[35] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in Neural Information Processing Systems*, 2017.

[36] Donald McCloskey and Arjo Klamer. One quarter of gdp is persuasion. *The American Economic Review*, 85(2):191–195, 1995.

[37] Kevin R McKee, Ian Gemp, Brian McWilliams, Edgar A Duèñez-Guzmán, Edward Hughes, and Joel Z Leibo. Social diversity and social preferences in mixed-motive reinforcement learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pages 869–877, 2020.

[38] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, 2016.

[39] Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisỳ, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):508–513, 2017.

[40] Roger B Myerson. Incentive compatibility and the bargaining problem. *Econometrica: Journal of the Econometric Society*, pages 61–73, 1979.

[41] Roger B Myerson. Optimal auction design. *Mathematics of Operations Research*, 6(1):58–73, 1981.

[42] Peng Peng, Ying Wen, Yaodong Yang, Quan Yuan, Zhenkun Tang, Haitao Long, and Jun Wang. Multiagent bidirectionally-coordinated nets: Emergence of human-level coordination in learning to play starcraft combat games. *arXiv preprint arXiv:1703.10069*, 2017.

[43] Xue Bin Peng, Angjoo Kanazawa, Sam Toyer, Pieter Abbeel, and Sergey Levine. Variational discriminator bottleneck: Improving imitation learning, inverse rl, and gans by constraining information flow. In *International Conference on Learning Representations*, 2019.

[44] Lillian J Ratliff and Tanner Fiez. Adaptive incentive design. *IEEE Transactions on Automatic Control*, 66(8):3871–3878, 2020.

[45] Luis Rayo and Ilya Segal. Optimal information disclosure. *Journal of Political Economy*, 118 (5):949–987, 2010.

[46] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 2020.

[47] Tianmin Shu and Yuandong Tian. M3̂RL: Mind-aware multi-agent management reinforcement learning. In *International Conference on Learning Representations*, 2019.

[48] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[49] Sainbayar Sukhbaatar, Rob Fergus, et al. Learning multiagent communication with backpropagation. *Advances in Neural Information Processing Systems*, 2016.

[50] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[51] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in Neural Information Processing Systems*, 1999.

[52] Ina Taneva. Information design. *American Economic Journal: Microeconomics*, 11(4):151–85, 2019.

[53] Leigh Tesfatsion and Kenneth L Judd. *Handbook of computational economics: Agent-based computational economics*. Elsevier, 2006.

[54] Richard H Thaler. Anomalies: The ultimatum game. *Journal of economic perspectives*, 2(4): 195–206, 1988.

[55] Mycal Tucker, Roger Levy, Julie A Shah, and Noga Zaslavsky. Trading off utility, informativeness, and complexity in emergent communication. *Advances in Neural Information Processing Systems*, 35:22214–22228, 2022.

[56] 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*, pages 471–472, 2022.

[57] Jiachen Yang, Ang Li, Mehrdad Farajtabar, Peter Sunehag, Edward Hughes, and Hongyuan Zha. Learning to incentivize other learning agents. *Advances in Neural Information Processing Systems*, 2020.

[58] Jiachen Yang, Ethan Wang, Rakshit Trivedi, Tuo Zhao, and Hongyuan Zha. Adaptive incentive design with multi-agent meta-gradient reinforcement learning. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pages 1436–1445, 2022.

[59] Stephan Zheng, Alexander Trott, Sunil Srinivasa, David C Parkes, and Richard Socher. The ai economist: Taxation policy design via two-level deep multiagent reinforcement learning. *Science Advances*, 8(18):eabk2607, 2022.

[60] You Zu, Krishnamurthy Iyer, and Haifeng Xu. Learning to persuade on the fly: Robustness against ignorance. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 927–928, 2021.

# A    Other Related Works

**Mechanism Design**    In parallel to the setting at hand, mechanism design[3] addresses situations in which agents possess distinct and private preferences [41, 13, 47]. The crux of this approach entails devising regulations that incentivize agents to candidly disclose their preferences in their own self-interest, culminating in a collectively optimal outcome. The community has extensively investigated the analytical paradigm of mechanism design. However, this paradigm is constrained by several factors such as linear agent cost and planner incentive functions [44], finite single-stage games [34], and state-based potential games [33]. Consequently, these simplifications restrict its applicability in nonlinear, temporally-extended environments [58].

To overcome these limitations, recent works adopt the agent-based simulation [53] and utilize SOTA agent learning methods such as MARL for mechanism design. While this approach sacrifices analytical tractability, it offers greater flexibility and applicability in complex environments [58]. As a method of mechanism design, providing rewards has been applied to MARL. For example, LIO allows the RL agents to send rewards directly to others, which can be used to solve first-order social dilemmas (e.g. iterated prisoner's dilemma and tragedy of the commons) and improve social welfare [57, 58]. The other perspective of influencing by rewards is taxation. By adopting RL, the AI economist improves utilitarian social welfare in one-step co-adaptive learning scenarios [59].

Besides, to elicit the desired social choice (the aggregation of the preferences of all the agents), the method of mechanism design is not necessarily to be providing rewards, but also more general mechanisms, such as distribution rules of public goods. For applications, [29] proposed a method that designs mechanisms by RL for voting. Their designed mechanisms successfully won the majority vote at the human level in a public goods social dilemma.

**Sequential Information Design**    Information design has recently been extended to sequential scenarios [19]. To model the coupling decision processes of the sender and the receiver, Markov persuasion processes (MPPs) are proposed in [18], and [56][4]. [18] proved that persuading a far-sighted receiver in MPPs is NP-hard, and [56] proposed a learning method for persuading a bunch of one-shot myopic receivers in MPPs. On the other hand, [5] proposed a learning method for a sender to persuade a far-sighted receiver without knowing the prior belief. Besides, [11] proposed a variant of obedience constraints for persuading multiple receivers in sequential interactions. The studies above have provided a solid theoretical foundation. However, all current discussions on this topic still rely on the commitment assumption and the revelation principle and need more algorithms that work in practical scenarios.

Compared to applying mechanism design to reinforcement learning, applying information design approaches to reinforcement learning presents a more challenging task. This is because the signals directly influence the agents' interactions, not only in their update phase but also in the generation of trajectory data. In contrast, the reward in mechanism design only affects the agent's update phase since it is solely used during updates. Therefore, the existing mechanism design in MARL methods cannot be directly applied to the case of information design, and alternative analyses are required.

**Reference Games**    Reference games are usually cooperative tasks (e.g. Lewis signaling games [32]). The core of these tasks is how the sender efficiently conveys its information to the receiver, enabling the receiver to take a specific action desired by both the sender and receiver. The crucial aspect lies in how the sender conveys information to enable the receiver to understand the intended semantics. Therefore, the sender aims to transmit as much of its available information as possible. While information design focuses on mixed-motive communication tasks. The receiver's action will determine the payoffs for both parties, but there may not be a specific action that is desired by both. In fact, the self-interested sender's goal is not to maximize the transmission of its known information, but to leverage its informational advantage to confuse the self-interested receiver, so as to maximize

---

[3]In the literature, the term *mechanism design* sometimes refers to a border definition that includes direct influence methods (e.g. providing tangible goods and information), and indirect methods (e.g. building reputations systems and infrastructure systems).

[4]Although both works name their models as MPPs, they are different models. In [18], the sender's informational advantage is separated from the states and has no impact on the state transitions. In [56], a new myopic receiver interacts with the sender every timestep. Both are different from our model.

its own payoffs. The sender aims to persuade the receiver to take actions that are advantageous to the sender. In fact, senders typically learn deceptive strategies.

**Emergent Communication**    Another domain closely related to our work is emergent communication [12, 10, 20, 55]. A close connection is found in [10], as they mention a negotiation environment which is indeed a mixed-motive task. The communication protocol we study employs the linguistic channel they mentioned, satisfying two primary characteristics: **(1)** Non-bindingness: Messages conveyed through this channel don't bind the sender to any action or state, given the sender's empty action space, rendering the communication as purely cheap talk; **(2)** Unverifiability: The absence of an intrinsic connection between the linguistic expression and the actual state of affairs implies a potential for deceit by the sender. Contrary to their methodology where self-interested agents fall short in achieving favorable outcomes via a linguistic channel, our approach facilitates good performance in Markov signaling games.

# B    Extensions of Markov Signaling Games

The MSG defined above can be extended to more general settings. Some of these extended models are compatible with our methods, requiring only minor modifications. Some extensions, though, require further investigation and are left for future work.

**The Sender's Actions**    It is immediate to allow the sender to take action at the same time as sending signals. In cases where the sender is permitted to take environmental actions beyond signaling, the sender $i$ chooses actions $a^i \in A^i$ according to its policy $\pi_{\theta^i}^i : S \times \Sigma \to \Delta(A^i)$. Notably, the sender's action policy considers the signals it sends to the receiver in the same round. This is necessary to enable the adaptation to a variety of receiver responses induced by the dispatched signals.

**Partial Observability of the Sender**    In some scenarios, the sender may only have access to a partial observation $o^i$. Under an informational advantage ($o^i - o^j \neq \varnothing$, and $\{o_t^i - o_t^j\}_{t \geq 0}$ affects $j$'s payoff) condition, there are 4 possible cases for $o_t^i$ and $o_t^j$, as shown in Figure 4. Case 2 can be



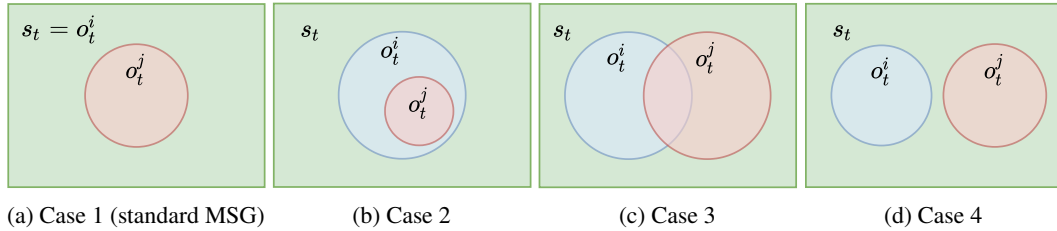(a) Case 1 (standard MSG)    (b) Case 2    (c) Case 3    (d) Case 4

Figure 4: Sender $i$ has informational advantage over receiver $j$. The informational advantage is reflected by $o_t^i - o_t^j$. The receiver's observation in standard MSGs is turned to $o_t^i \cap o_t^j$ in every case.

handled well in practice. The sender could estimate the current state from its observation sequence, similar to methods in partially observable Markov decision processes (POMDPs) [14]. In Cases 3 and 4, the sender needs to estimate $o_t^j - o_t^i \neq \varnothing$ (information that the receiver know but the sender does not) and consider the effect of it. We leave Cases 3 and 4 for future work.

**Multiple Senders**    One possible approach to model multiple senders is by modeling a separate MSG for each sender. However, it treats each sender independently and overlooks the interplay among them. The general setting should model this as a game among senders and further specify the receiver's decision based on multiple signals received from different senders. This is an important topic for future works because, more generally, more than one agent possesses such information.

**Multiple Receivers**    Markov signaling games with multiple receivers are an immediate extension. The conclusions of the signaling gradient Lemma 4.1 can also be applied. However, if the sender wants to persuade a group of receivers simultaneously, additional considerations need to be taken into account for extended obedience constraints. Since the reward is determined by all receivers' actions, the sender must simultaneously consider the effects of signaling to all receivers, the dependencies of

which are not yet clear. This issue is left for future work. In this section, we redefine the model for clarity. Subsequent discussions in the appendix are based on this extension.

Consider a signaling game involving 1 sender and $N$ receivers, where $J = \{0, 1, \ldots, N-1\}$ denotes the set of receivers. The signaling channel between the sender and each receiver is private, so the messages sent through each channel are only observable to the sender and the corresponding receiver. $\Sigma^j$ denotes the message set of receiver $j \in J$, and the joint message set is defined as $\boldsymbol{\Sigma} = \prod_j \Sigma^j$. Each receiver $j \in J$ makes decisions based only on received messages and its observation $o^j \in O^j$. At each timestep, the environment generates a joint observation $\boldsymbol{o}_t \in \prod_j O^j$ according to the emission function $q : S \to \prod_j O^j$, where each signal $o_t^j \subset s_t$ is a proper subset of $s_t$. The sender's informational advantage over receiver $j$ is reflected by $s_t - o_t^j$, where $\{s_t - o_t^j\}_{t \geq 0}$ affects $j$'s payoff.

The sender maintains a stochastic signaling scheme $\varphi_\eta : S \to \Delta(\boldsymbol{\Sigma})$. The stochastic action policy of receiver $j$ is denoted as $\pi_{\theta^j}^j : O^j \times \Sigma^j \to \Delta(A^j)$. Specifically, $\pi_{\theta^j}^j(a^j \mid o^j, \sigma^j)$ represents the probability of receiver $j$ choosing an action $a^j$ given the message $\sigma^j$ and the observation $o^j$ received. The joint action space is then defined as $\boldsymbol{A} = \prod_j A^j$, where $A^j$ is the action space of receiver $j$ and $\theta^j \in \Theta^j$ is the corresponding policy parameter. And the joint policy of all agents is defined as $\boldsymbol{\pi_\theta}(\boldsymbol{a} \mid \boldsymbol{o}, \boldsymbol{\sigma}) = \prod_j \pi_{\theta^j}^j(a^j \mid o^j, \sigma^j)$, where $\boldsymbol{a} \in \boldsymbol{A}$, and $\boldsymbol{\theta} \in \prod_j \Theta^j$. When the context is clear, we will drop the subscripts for the parameters and let $\pi_\theta^j$ denote $\pi_{\theta^j}^j$.
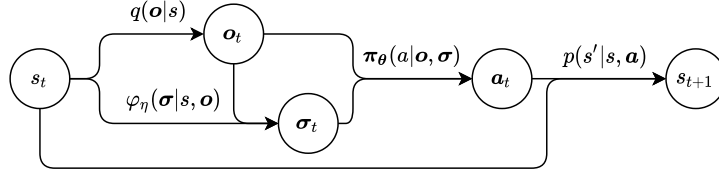


Figure 5: An illustration of the Markov signaling game with multiple receivers. The arrows symbolize probability distributions, whereas the nodes denote the sampled variables.

Then, a Markov signaling game with multiple receivers(Figure 5) is defined as a tuple

$$\mathcal{G}' = \left( i, J, S, \{O^j\}_{j \in J}, \{\Sigma^j\}_{j \in J}, \{A^j\}_{j \in J}, R^i, \{R^j\}_{j \in J}, p, q \right).$$

In $\mathcal{G}'$, the sender $i$ observes a state $s \in S$ and sends messages $\boldsymbol{\sigma}$ based on $\varphi_\eta$, and the environment generates joint observations $\boldsymbol{o}$ according to the emission function $q$. Then all agents take actions $\boldsymbol{a}$ based on the joint policy $\boldsymbol{\pi_\theta}$ and the environment transits to the next state $s'$ according to the transition function $p : S \times \boldsymbol{A} \to \Delta(S)$. Meanwhile, the sender $i$ (respectively, a receiver $j$) receives the reward $r^i$ ($r^j$) via the reward function $R^i : S \times \boldsymbol{A} \to \mathbb{R}$ ($R^j : S \times \boldsymbol{A} \to \mathbb{R}$). The agents and the environment repeat this process until the environment terminates the episode.

The definition of value functions for the sender's signaling process in MSGs with multiple receivers can be immediately obtained from the definition in MSGs. The sender's state value function is defined as $V_{\varphi,\boldsymbol{\pi}}^i(s) = \mathbb{E}_{\varphi,\boldsymbol{\pi}}\left[ G_t^i \mid s_t = s \right]$. The signal value function $Q$ for the sender of taking signal $\sigma$ at the state $s$ is $Q_{\varphi,\boldsymbol{\pi}}^i(s, \boldsymbol{\sigma}) = \mathbb{E}_{\varphi,\boldsymbol{\pi}}\left[ G_t^i \mid s_t = s, \boldsymbol{\sigma}_t = \boldsymbol{\sigma} \right]$. The action value function $U$ is defined as $U_{\varphi,\boldsymbol{\pi}}^i(s, \boldsymbol{\sigma}, \boldsymbol{a}) = \mathbb{E}_{\varphi,\boldsymbol{\pi}}\left[ G_t^i \mid s_t = s, \boldsymbol{\sigma}_t = \boldsymbol{\sigma}, \boldsymbol{a}_t = \boldsymbol{a} \right]$. And still, $W_{\varphi,\boldsymbol{\pi}}^i(s, \boldsymbol{a}) = U_{\varphi,\boldsymbol{\pi}}^i(s, \boldsymbol{\sigma}, \boldsymbol{a})$.

## C  Bellman Equations in Markov Signaling Games

According to the definitions of the value functions in MSGs (defined in Section 3) and the law of total expectation, it can immediately derive a variant of Bellman equations as

$$
\begin{aligned}
V_{\varphi,\boldsymbol{\pi}}^i(s) &= \sum_{\boldsymbol{o}} \Pr(\boldsymbol{o} \mid s) \sum_{\boldsymbol{\sigma}} \Pr(\boldsymbol{\sigma} \mid s, \boldsymbol{o}) \sum_{\boldsymbol{a}} \Pr(\boldsymbol{a} \mid s, \boldsymbol{o}, \boldsymbol{\sigma}) \cdot U_{\varphi,\boldsymbol{\pi}}^i(s, \boldsymbol{\sigma}, \boldsymbol{a}) \\
&= \sum_{\boldsymbol{o}} q(\boldsymbol{o} \mid s) \sum_{\boldsymbol{\sigma}} \varphi_\eta(\boldsymbol{\sigma} \mid s) \sum_{\boldsymbol{a}} \boldsymbol{\pi_\theta}(\boldsymbol{a} \mid \boldsymbol{o}, \boldsymbol{\sigma}) \cdot U_{\varphi,\boldsymbol{\pi}}^i(s, \boldsymbol{\sigma}, \boldsymbol{a}).
\end{aligned}
\tag{8}
$$

17

In particular, in our current model, we set the environment to not give any reward to the signaling processes. And there is no cost for sending a message. Thus,

$$
\begin{aligned}
U_{\varphi,\pi}^i(s, \boldsymbol{\sigma}, \boldsymbol{a}) &= \mathbb{E}_{\varphi,\pi}\left[G_t^i \mid s_t = s, \boldsymbol{\sigma}_t = \boldsymbol{\sigma}, \boldsymbol{a}_t = \boldsymbol{a}\right] \\
&= \mathbb{E}_{\varphi,\pi}\left[G_t^i \mid s_t = s, \boldsymbol{a}_t = \boldsymbol{a}\right] = W_{\varphi,\pi}^i(s, \boldsymbol{a}),
\end{aligned}
\tag{9}
$$

$$
\begin{aligned}
V_{\varphi,\pi}^i(s) &= \mathbb{E}_{\varphi,\pi}\left[G_t^i \mid s_t = s\right] = \mathbb{E}_{\varphi,\pi}\left[r_{t+1}^i + \gamma \cdot G_{t+1}^i \mid s_t = s\right] \\
&= \mathbb{E}_{\varphi,\pi}\left[r_{t+1}^i + \gamma \cdot V_{\varphi,\pi}^i(s_{t+1}) \mid s_t = s\right],
\end{aligned}
\tag{10}
$$

and

$$
\begin{aligned}
U_{\varphi,\pi}^i(s, \boldsymbol{\sigma}, \boldsymbol{a}) &= \mathbb{E}_{\varphi,\pi}\left[G_t^i \mid s_t = s, \boldsymbol{\sigma}_t = \boldsymbol{\sigma}, \boldsymbol{a}_t = \boldsymbol{a}\right] \\
&= \mathbb{E}_{\varphi,\pi}\left[r_{t+1}^i + \gamma \cdot G_{t+1}^i \mid s_t = s, \boldsymbol{\sigma}_t = \boldsymbol{\sigma}, \boldsymbol{a}_t = \boldsymbol{a}\right] \\
&= \mathbb{E}_{\varphi,\pi}\left[r_{t+1}^i + \gamma \cdot V_{\varphi,\pi}^i(s_{t+1}) \mid s_t = s, \boldsymbol{\sigma}_t = \boldsymbol{\sigma}, \boldsymbol{a}_t = \boldsymbol{a}\right] \\
&= R^i(s, \boldsymbol{a}) + \gamma \sum_{s'} p(s' \mid s, \boldsymbol{a}) \cdot V_{\varphi,\pi}^i(s').
\end{aligned}
\tag{11}
$$

## D    Proof of the Signaling Gradient Lemma

Firstly, let $\Pr_{\varphi,\pi}(s \to x, k)$ denote the probability of transferring from state $s$ to state $x$ in $k$ steps, given the signaling scheme $\varphi$ and the joint policy $\pi$. Its recursive relationship is similar to the situation in MDPs:

$$
\Pr_{\varphi,\pi}(s \to s, 0) = 1, \Pr_{\varphi,\pi}(s \to s', 1) = \sum_{\boldsymbol{\sigma}, \boldsymbol{o}, \boldsymbol{a}} q(\boldsymbol{o} \mid s) \cdot \varphi_\eta(\boldsymbol{\sigma} \mid s) \cdot \pi_\theta(\boldsymbol{a} \mid \boldsymbol{o}, \boldsymbol{\sigma}) \cdot p(s' \mid s, \boldsymbol{a}),
$$

$$
\Pr_{\varphi,\pi}(s \to x, k+1) = \sum_{s'} \Pr_{\varphi,\pi}(s \to s', k) \cdot \Pr_{\varphi,\pi}(s' \to x, 1).
\tag{12}
$$

Then according to 8,

$$
\nabla_\eta V_{\varphi,\pi}^i(s) = \nabla_\eta \sum_{\boldsymbol{o}} q(\boldsymbol{o} \mid s) \sum_{\boldsymbol{\sigma}, \boldsymbol{a}} \varphi_\eta(\boldsymbol{\sigma} \mid s) \cdot \pi_\theta(\boldsymbol{a} \mid \boldsymbol{o}, \boldsymbol{\sigma}) \cdot U_{\varphi,\pi}^i(s, \boldsymbol{\sigma}, \boldsymbol{a})
\tag{13}
$$

$$
= \sum_{\boldsymbol{o}} q(\boldsymbol{o} \mid s) \sum_{\boldsymbol{\sigma}, \boldsymbol{a}} \nabla_\eta \left[\varphi_\eta(\boldsymbol{\sigma} \mid s) \cdot \pi_\theta(\boldsymbol{a} \mid \boldsymbol{o}, \boldsymbol{\sigma})\right] \cdot U_{\varphi,\pi}^i(s, \boldsymbol{\sigma}, \boldsymbol{a})
\tag{14}
$$

$$
+ \sum_{\boldsymbol{o}} q(\boldsymbol{o} \mid s) \sum_{\boldsymbol{\sigma}, \boldsymbol{a}} \varphi_\eta(\boldsymbol{\sigma} \mid s) \cdot \pi_\theta(\boldsymbol{a} \mid \boldsymbol{o}, \boldsymbol{\sigma}) \cdot \nabla_\eta U_{\varphi,\pi}^i(s, \boldsymbol{\sigma}, \boldsymbol{a}).
\tag{15}
$$

In the following many steps of derivation, we will expand 15, leaving 14 unchanged. Since 14 is a function of the state $s$, for brevity, we will let $f_{\varphi,\pi}(s)$ denote it:

$$
\nabla_\eta V_{\varphi,\pi}^i(s) = f_{\varphi,\pi}(s) + \sum_{\boldsymbol{o}} q(\boldsymbol{o} \mid s) \sum_{\boldsymbol{\sigma}, \boldsymbol{a}} \varphi_\eta(\boldsymbol{\sigma} \mid s) \cdot \pi_\theta(\boldsymbol{a} \mid \boldsymbol{o}, \boldsymbol{\sigma}) \cdot \nabla_\eta U_{\varphi,\pi}^i(s, \boldsymbol{\sigma}, \boldsymbol{a})
$$

$$
= f_{\varphi,\pi}(s) + \sum_{\boldsymbol{o}} q(\boldsymbol{o} \mid s) \sum_{\boldsymbol{\sigma}, \boldsymbol{a}} \varphi_\eta(\boldsymbol{\sigma} \mid s) \cdot \pi_\theta(\boldsymbol{a} \mid \boldsymbol{o}, \boldsymbol{\sigma}) \cdot \nabla_\eta \left[R^i(s, \boldsymbol{a}) + \gamma \sum_{s'} p(s' \mid s, \boldsymbol{a}) \cdot V_{\varphi,\pi}^i(s')\right]
$$

$$
= f_{\varphi,\pi}(s) + \gamma \sum_{\boldsymbol{o}} q(\boldsymbol{o} \mid s) \sum_{\boldsymbol{\sigma}, \boldsymbol{a}} \varphi_\eta(\boldsymbol{\sigma} \mid s) \cdot \pi_\theta(\boldsymbol{a} \mid \boldsymbol{o}, \boldsymbol{\sigma}) \cdot \sum_{s'} p(s' \mid s, \boldsymbol{a}) \cdot \nabla_\eta V_{\varphi,\pi}^i(s')
$$

$$
= f_{\varphi,\pi}(s) + \gamma \sum_{s'} \Pr_{\varphi,\pi}(s \to s', 1) \cdot \nabla_\eta V_{\varphi,\pi}^i(s').
$$

$$
\tag{16}
$$

18

Keep unrolling recursively,

$$\nabla_\eta V^i_{\varphi,\boldsymbol{\pi}}(s) = f_{\varphi,\boldsymbol{\pi}}(s) + \gamma \sum_{s'} \mathrm{Pr}_{\varphi,\boldsymbol{\pi}}(s \to s', 1) \cdot \nabla_\eta V^i_{\varphi,\boldsymbol{\pi}}(s')$$

$$= f_{\varphi,\boldsymbol{\pi}}(s) + \gamma \sum_{s'} \mathrm{Pr}_{\varphi,\boldsymbol{\pi}}(s \to s', 1) \cdot \left[ f_{\varphi,\boldsymbol{\pi}}(s') + \gamma \sum_{s''} \mathrm{Pr}_{\varphi,\boldsymbol{\pi}}(s' \to s'', 1) \cdot \nabla_\eta V^i_{\varphi,\boldsymbol{\pi}}(s'') \right]$$

$$= f_{\varphi,\boldsymbol{\pi}}(s) + \gamma \sum_{s'} \mathrm{Pr}_{\varphi,\boldsymbol{\pi}}(s \to s', 1) \cdot f_{\varphi,\boldsymbol{\pi}}(s') + \gamma^2 \sum_{s''} \mathrm{Pr}_{\varphi,\boldsymbol{\pi}}(s \to s'', 2) \cdot \nabla_\eta V^i_{\varphi,\boldsymbol{\pi}}(s'') \quad (17)$$

$$\cdots$$

$$= \sum_{x \in S} \sum_{k=0}^{\infty} \gamma^k \cdot \mathrm{Pr}_{\varphi,\boldsymbol{\pi}}(s \to x, k) \cdot f_{\varphi,\boldsymbol{\pi}}(x).$$

Let $h_{\varphi,\boldsymbol{\pi}}(x)$ denote $\sum_{k=0}^{\infty} \gamma^k \cdot \mathrm{Pr}_{\varphi,\boldsymbol{\pi}}(s \to x, k)$. Then the stationary distribution $d_{\varphi,\boldsymbol{\pi}}(s)$ is defined as

$$d_{\varphi,\boldsymbol{\pi}}(s) = \frac{h_{\varphi,\boldsymbol{\pi}}(s)}{\sum_{x \in S} h_{\varphi,\boldsymbol{\pi}}(x)}. \quad (18)$$

Considering the objective of signaling gradient is to optimize $V^i_{\varphi,\boldsymbol{\pi}}(s_0)$, then we have

$$\nabla_\eta V^i_{\varphi,\boldsymbol{\pi}}(s_0) = \sum_s h_{\varphi,\boldsymbol{\pi}}(s) \cdot f_{\varphi,\boldsymbol{\pi}}(s) = \left( \sum_s h_{\varphi,\boldsymbol{\pi}}(s) \right) \sum_s \frac{h_{\varphi,\boldsymbol{\pi}}(s)}{\sum_s h_{\varphi,\boldsymbol{\pi}}(s)} \cdot f_{\varphi,\boldsymbol{\pi}}(s)$$

$$\propto \sum_s \frac{h_{\varphi,\boldsymbol{\pi}}(s)}{\sum_s h_{\varphi,\boldsymbol{\pi}}(s)} \cdot f_{\varphi,\boldsymbol{\pi}}(s) = \sum_s d_{\varphi,\boldsymbol{\pi}}(s) \cdot f_{\varphi,\boldsymbol{\pi}}(s)$$

$$= \sum_s d_{\varphi,\boldsymbol{\pi}}(s) \cdot \sum_{\boldsymbol{o}} q(\boldsymbol{o} \mid s) \sum_{\boldsymbol{\sigma},\boldsymbol{a}} U^i_{\varphi,\boldsymbol{\pi}}(s, \boldsymbol{\sigma}, \boldsymbol{a}) \cdot \nabla_\eta [\varphi_\eta(\boldsymbol{\sigma} \mid s) \cdot \boldsymbol{\pi_\theta}(\boldsymbol{a} \mid \boldsymbol{o}, \boldsymbol{\sigma})] \quad (19)$$

$$= \sum_{s,\boldsymbol{o},\boldsymbol{\sigma},\boldsymbol{a}} d_{\varphi,\boldsymbol{\pi}}(s) \cdot q(\boldsymbol{o} \mid s) \cdot U^i_{\varphi,\boldsymbol{\pi}}(s, \boldsymbol{\sigma}, \boldsymbol{a}) \cdot \varphi_\eta(\boldsymbol{\sigma} \mid s) \cdot \boldsymbol{\pi_\theta}(\boldsymbol{a} \mid \boldsymbol{o}, \boldsymbol{\sigma}) \cdot \frac{\nabla_\eta \varphi_\eta(\boldsymbol{\sigma} \mid s)}{\varphi_\eta(\boldsymbol{\sigma} \mid s)}$$

$$+ \sum_{s,\boldsymbol{o},\boldsymbol{\sigma},\boldsymbol{a}} d_{\varphi,\boldsymbol{\pi}}(s) \cdot q(\boldsymbol{o} \mid s) \cdot U^i_{\varphi,\boldsymbol{\pi}}(s, \boldsymbol{\sigma}, \boldsymbol{a}) \cdot \varphi_\eta(\boldsymbol{\sigma} \mid s) \cdot \boldsymbol{\pi_\theta}(\boldsymbol{a} \mid \boldsymbol{o}, \boldsymbol{\sigma}) \cdot \frac{\nabla_\eta \boldsymbol{\pi_\theta}(\boldsymbol{a} \mid \boldsymbol{o}, \boldsymbol{\sigma})}{\boldsymbol{\pi_\theta}(\boldsymbol{a} \mid \boldsymbol{o}, \boldsymbol{\sigma})}$$

$$= \mathbb{E}_{\varphi,\boldsymbol{\pi}} \left[ U^i_{\varphi,\boldsymbol{\pi}}(s, \boldsymbol{\sigma}, \boldsymbol{a}) \cdot [\nabla_\eta \log \varphi_\eta(\boldsymbol{\sigma} \mid s) + \nabla_\eta \log \boldsymbol{\pi_\theta}(\boldsymbol{a} \mid \boldsymbol{o}, \boldsymbol{\sigma})] \right].$$

Finally, by substituting $U^i_{\varphi,\boldsymbol{\pi}}(s, \boldsymbol{\sigma}, \boldsymbol{a})$ by $W^i_{\varphi,\boldsymbol{\pi}}(s, \boldsymbol{a})$ (as analyzed in 9), the deriving result of the signaling gradient is

$$\nabla_\eta V^i_{\varphi,\boldsymbol{\pi}}(s_0) = \mathbb{E}_{\varphi,\boldsymbol{\pi}} \left[ W^i_{\varphi,\boldsymbol{\pi}}(s, \boldsymbol{a}) \cdot [\nabla_\eta \log \varphi_\eta(\boldsymbol{\sigma} \mid s) + \nabla_\eta \log \boldsymbol{\pi_\theta}(\boldsymbol{a} \mid \boldsymbol{o}, \boldsymbol{\sigma})] \right]. \quad (20)$$

## E  Hyper Gradient in the Signaling Gradient

Note that in every $\nabla_\eta \log \pi^j_\theta(a^j \mid o^j, \sigma^j)$, it can be decomposed to two parts:

$$\nabla_\eta \log \pi^j_\theta(a^j \mid o^j, \sigma^j) = \frac{\partial \log \pi^j_\theta(a^j \mid o^j, \sigma^j)}{\partial \pi^j_\theta(a^j \mid o^j, \sigma^j)} \cdot \frac{\partial \pi^j_\theta(a^j \mid o^j, \sigma^j)}{\partial \eta}$$

$$= \frac{1}{\pi^j_\theta(a^j \mid o^j, \sigma^j)} \left[ \frac{\partial \pi^j_\theta(a^j \mid o^j, \sigma^j)}{\partial \sigma^j} \cdot \frac{\partial \sigma^j}{\partial \eta} + \frac{\partial \pi^j_\theta(a^j \mid o^j, \sigma^j)}{\partial \theta^j} \cdot \frac{\partial \theta^j}{\partial \eta} \right]. \quad (21)$$

Similar to LIO, it can be considered that $\eta$ affects the update process of $\theta^j$. In this way,

$$\frac{\partial \theta^j}{\partial \eta} \approx \frac{\partial \Delta \theta^j}{\partial \eta}, \tag{22}$$

where $\Delta \theta^j$ is the difference for one-step update $\theta^j \leftarrow \theta^j + \Delta \theta^j$.

## F    Details and Analysis of Recommendation Letter

To better understand the information design problem, we illustrate it with an example of `Recommendation Letter` and its Bayesian persuasion solution [15, 28]. In the example, a professor will write recommendation letters for a number of graduating students, and a company's human resources department (HR) will receive the letters and decide whether to hire the students. The professor and the HR share a prior distribution of the candidates' quality, with a probability of $1/3$ that the candidate is strong and a probability of $2/3$ that the candidate is weak. The HR does not know exactly what each student's quality is but wants to hire strong students, while the letters are the only source of information. The HR will get a reward of $1$ for hiring a strong candidate, a penalty of $-1$ for hiring a weak candidate, and a reward of $0$ for not hiring. The professor gets a $1$ reward for each hire. There are three types of outcomes between the professor and the HR:

- Since there are more weak students than strong students, the HR tends not to hire anyone if the professor does not write letters.

- If the professor honestly reports the qualities of the students, then the HR will accurately hire those strong candidates. Then their payoff expectations are both $1/3$;

- The professor reports the qualities of the strong students honestly and lying with a probability of $(1/2 - \epsilon)$ for weak students, for some arbitrarily small $\epsilon$. The optimal policy of HR is to respect the professor's recommendations. In this way, the professor's and the HR's payoff expectations are $(2/3 - 2\epsilon/3)$ and $2\epsilon/3$, respectively.

The critical insight from the example is that the information provider (the professor) needs to "lie" about its information to get the best interest. This lie, meanwhile, must still reveal part of the truth so that the information receiver (the HR) respects the information because it will benefit from the posterior belief in its best interest. The condition that the receiver benefits from the message is known as the *obedience constraints* in information design, which implies the incentive compatibility of the receiver. The sender must subtly balance the benefits of both parties under this condition and carefully design the information to be sent.

It is easy to analyze that HR can only make decisions based on the prior probability distribution if the professor does not write a recommendation letter, and its best policy is to refuse to hire any student. In this way, the professor and the HR payoffs are both $0$, which is obviously a bad situation for both parties. If the professor tells the HR the student's quality honestly (i.e., the professor gives up its informational advantage), then the best strategy for the HR is to hire strong students and not weak students. In the case of the honest signaling scheme, the payoff expectations of the professor and the HR are $1/3$.

The professor can change its signaling scheme to make its payoff expectation higher, which is exactly the primary concern of information design. If the current student is strong, the professor will report it honestly; otherwise, the professor tells the HR that it is strong with a probability of $(1/2 - \epsilon)$, where $\epsilon \in (0, 1/2]$. When HR heard the professor say this was a weak student, it knew the student must be weak, so it would refuse to hire her. And the HR can calculate that $1/3$ of the students are strong, and the professor will call them strong, and $(1/3 - 2\epsilon/3)$ of students are weak, but the professor will call them strong still. So when the professor says that the current student is strong, the probability of being a strong student is $1/(2 - 2\epsilon)$, and the probability of being a weak student is $(1 - 2\epsilon)/(2 - 2\epsilon)$. Then, when the professor recommends the student, the payoff expectation of the HR of choosing to hire is $\epsilon/(1 - \epsilon)$, and the payoff expectation of choosing not to hire is $0$. A rational HR will select the action that can maximize its payoff expectation. That is to say, when the professor says that the current student is strong, the HR will choose to hire. In this case, the payoff expectation of the professor is $(2/3 - 2\epsilon/3)$, and the payoff expectation of the HR is $2\epsilon/3$. It can be found that when epsilon takes $1/2$, the signaling scheme degenerates into the honest one.

## G Details of Reaching Goals

The basics of `Reaching Goals` are introduced in Section 5.2. This section aims to offer more specific details. Example maps of `Reaching Goals` are shown in Figure 6.

At any given time in the map, there is only one target goal for the sender and one for the receiver, both uniformly distributed and randomly generated. Once the receiver reaches a goal, it will be regenerated. And the regenerated goal and the receiver will not be in the same position. An episode will only end when the specified step limit is reached. The receiver's actions consist of moving up, down, left, or right by one square. When the receiver's position coincides with a particular goal, the apple will be automatically harvested.

When the receiver's decision to pursue the green apple would mean moving away from the red goal. And in a fixed-length episode, this will reduce the receiver's goal harvesting efficiency. Since the respawn locations of goals are randomly and uniformly distributed, the positions of goals are highly likely to be non-coincident. As the map size increases, the conflict of interest between the sender and the receiver increases.
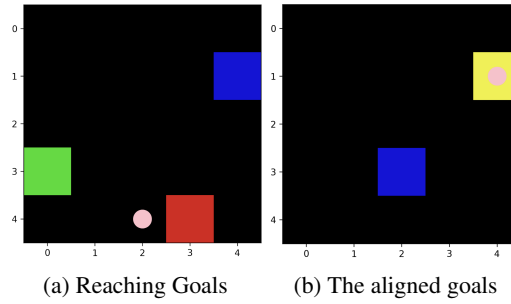


(a) Reaching Goals       (b) The aligned goals

Figure 6: Maps $5 \times 5$ of `Reaching Goals`. The blue, red, and green squares represent the receiver, the sender's goal, and the receiver's goal, respectively. If the red square and the green square overlap, it will turn yellow, meaning that the goals of agents are aligned. The pink dots represent the messages sent by the sender. The sender is out of the map.

## H More Discussions and Results

### H.1 Incentive Compatibility of the Obedience Constraints

The obedience constraints are the core of information design and it significantly improves the performance in experiments. Assuming that the sender's signal set is equal to the receiver's action set, the sender's signals can be interpreted as recommending the receiver to take a specific action (This common assumption is without loss of generality according to the revelation principle). Under this premise, obedience constraints ensure that the receiver will definitely follow the sender's recommendations.

To explain why is it that as long as the sender's signaling scheme satisfies obedience constraints, the receiver will definitely follow the sender's recommendations. We provide a simple derivation adapted from [3] as follows:

$$\sum_s \mu_0(s) \cdot \varphi(a \mid s) \cdot \left( r^j(s,a) - r^j(s,a') \right) \geq 0$$

$$\Leftrightarrow \sum_s \frac{\mu_0(s) \cdot \varphi(a \mid s)}{\sum_{s'} \mu_0(s') \cdot \varphi(a \mid s')} \cdot \left( r^j(s,a) - r^j(s,a') \right) \geq 0, \quad \forall a' \in A$$

$$\Leftrightarrow \sum_s \mu(s \mid a) \cdot \left( r^j(s,a) - r^j(s,a') \right) \geq 0, \qquad \forall a' \in A$$

$$\Leftrightarrow \sum_s \mu(s \mid a) \cdot r^j(s,a) \geq \sum_s \mu(s \mid a) \cdot r^j(s,a'), \qquad \forall a' \in A$$

where $\mu_0$ represents the prior probability, and $\mu$ represents the posterior probability. Therefore, a self-interested and rational receiver will definitely follow the sender's recommendations, because the posterior expected payoff of the action recommended by the sender is greater than or equal to the posterior expected payoffs of all other actions.

## H.2 The Non-Stationarity is Alleviated by the Signaling Gradient

This issue is a very important one and involves the insight behind signaling gradient. Non-stationarity is a common problem in the field of MARL. Each RL agent only focuses on its own learning, which means that every agent treats all quantities except itself as part of the environment, treating other agents as part of the environment. As a result, for an agent with policy, when other agents update, its environment changes, and even with the same policy, its value function will change as a result of the updates to other agents' states. We believe this phenomenon arises due to the independence of agents. However, our signaling gradient does not update its network independently; instead, it considers the impact of the receiver's policy. This is precisely the core insight behind signaling gradient.

## H.3 Discussions about DIAL

For the same experiment, if we add a constant to all sender rewards, the sender can still persuade the receiver using the same signaling scheme (as the receiver's reward remains unchanged), and thus the receiver's utility remains unaffected.

In this case, the difference between fully cooperative communication methods (e.g., DIAL) and our method becomes evident. Because DIAL only considers the receiver's reward and does not utilize the sender's reward (according to their concept, their method does not even use $r^i$ and does not reflect the self-interested sender). If we add a constant to all sender rewards, their method's social welfare remains unchanged, while our method's social welfare improves.

## H.4 Discussions about LOLA

When multiple learning agents are involved in a training process, each agent's environment becomes non-stationary. Modeling each agent as a POMDP means treating other people as part of the environment, so when other agents' policies are updated, each agent's environment changes. It can result in unstable training or undesired final results. To alleviate the non-stationarity, LOLA [17] was proposed. This method lets each agent notice the update process of others to adapt itself to this ever-changing environment. This awareness is implemented by accounting for others' gradient when an agent updates its policy. The signaling gradient we proposed is also to solve the non-stationarity problem. From the perspective of conclusion, the sender's signaling gradient also considers the receiver's policy. However, our goal is to improve the sender's ability to influence the receiver, and using LOLA would encourage the agent to adapt to updates from other agents.

Another potential question is whether LOLA can be used as a replacement for obedience constraints. LOLA is not a direct replacement for obedience constraints but rather a technique that can be used to enhance the receiver's algorithm. LOLA emphasizes an agent's consideration of others' updates during its updates to adapt to non-stationarity in multi-agent scenarios. When applied to a receiver's algorithm, LOLA allows the receiver to consider changes in the signaling scheme, which can improve its performance. On the other hand, obedience constraints restrict the sender to ensure that the signals it emits give the impression to the receiver that its rewards are not significantly reduced. Adapting the receiver to changes in signaling may not necessarily make the receiver follow the sender's recommendations. In contrast, obedience constraints provide a more muscular incentive-compatible condition regarding rewards, which can more effectively align the sender's influence.

## H.5 Discussions about the Dual Gradient Descent Method

One reasonable consideration is whether the dual gradient descent method (DGD) can be used to learn Markov signaling games as discussed in Equation (5), as this approach does not require tuning the Lagrangian multipliers. Other works, such as [43] and [23], have also utilized the DGD method.

By applying the dual gradient descent, $\eta$ and $\boldsymbol{\lambda}$ are updated as

$$\eta \leftarrow \eta - \alpha \cdot \nabla_\eta L(\eta, \boldsymbol{\lambda}), \quad \boldsymbol{\lambda} \leftarrow (\boldsymbol{\lambda} + \alpha \cdot \nabla_{\boldsymbol{\lambda}} L(\eta, \boldsymbol{\lambda}))^+, \tag{23}$$

where $\alpha$ is the learning rate, $(\cdot)^+ = \max\{0, \cdot\}$, and $L(\eta, \boldsymbol{\lambda}) = -\mathbb{E}_{\varphi, \boldsymbol{\pi}}\left[V^i(s)\right] - \sum_{j, \sigma^j, \sigma^{j'}} \lambda_{j, \boldsymbol{\sigma}, \boldsymbol{\sigma}'} \cdot$
$C_\varphi^j(\boldsymbol{\sigma}, \boldsymbol{\sigma}')$ is the Lagrangian function of Equation (5). All the gradients required have been discussed. The comparisons of performance in the `Recommendation Letter` and the `Reaching Goals` experiments are shown in Figure 7.
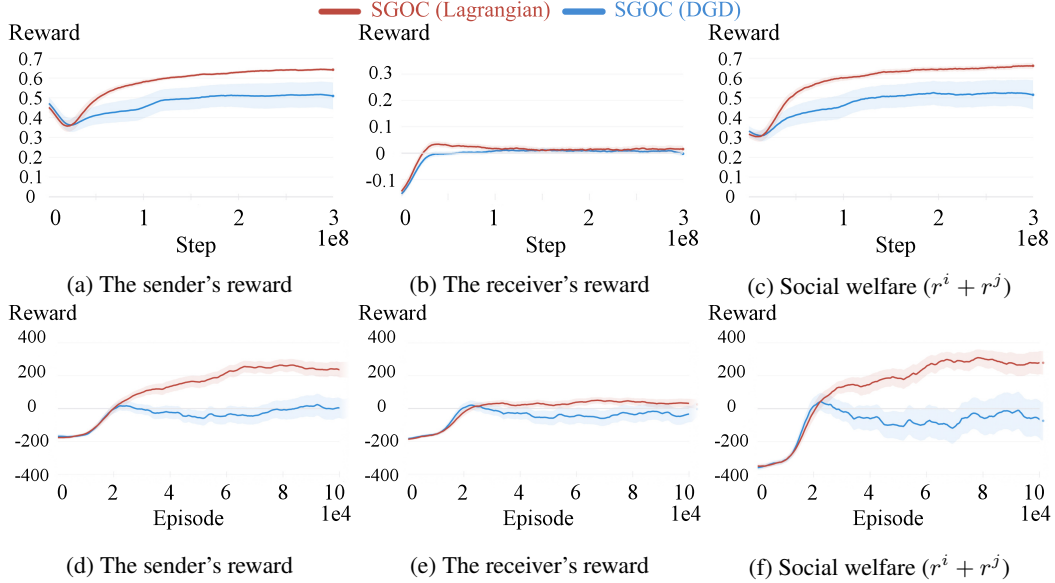


Figure 7: Performance comparisons of SGOC and DGD. (a-c) The results of `Recommendation Letter`. (d-f) The results of `Reaching Goals` with $3 \times 3$ map. The rewards and penalties are amplified by 20 and 5 respectively. The receiver can only see its position.

## H.6  Results with Different Hyperparameters

We plot the heatmap of the honesty metric against $\epsilon$, $\lambda$, as shown in Figure 8. We observe that the honesty metric increases with $\epsilon$ and $\lambda$, which agrees with our intuition. Specifically, when lambda reaches over 3.75 (respectively, $\epsilon$, 0.15), the honesty stops increasing, which means the sender is being very honest in this region. The best value for $\lambda$ is then somewhere between 0 to 5 (respectively, $\epsilon$, 0 to 0.3).
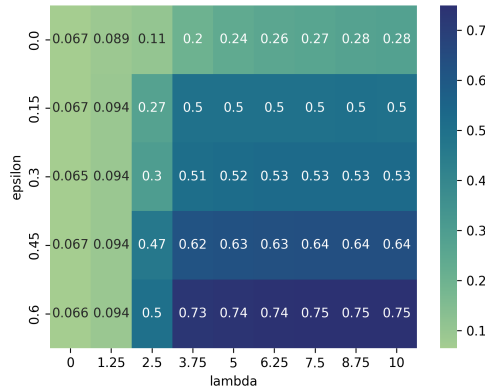


Figure 8: Honesty heatmap of the sender's signaling scheme in `Recommendation Letter`.

## H.7 Results with Different Observation of the Receiver

The algorithm proposed in this paper is suitable for scenarios where the sender has an informational advantage (discussed in Section 3). We conducted various experiments to investigate the impact of the receiver's observation in the SGOC algorithm in the `Reaching Goals` scenarios. The results, shown in Figure 9, compare the situations where the receiver "cannot see anything" (No-obs), "can only see its location" (Pos-obs), and "can see both its location and the location of its preferred apple" (Full-obs). The results indicate that the sender's payoff decreases as the receiver knows more (the sender's informational advantage decreases).
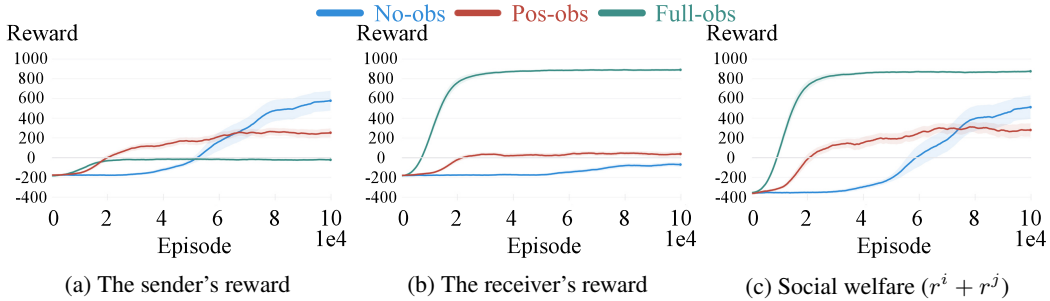


Figure 9: Performance comparisons in `Reaching Goals` with $3 \times 3$ map. The rewards and penalties are amplified by $20$ and $5$ respectively.

# I Broader Impacts

In practical economic scenarios, persuasion is ubiquitous and plays a crucial role. As stated in the title and conclusion of McCloskey and Klamer's paper, "one quarter of the GDP is persuasion" [36]. This kind of persuasion demonstrates that communication is conceivable in mixed-motive scenarios. In this paper, we build upon the classic model of Bayesian persuasion to discuss how RL agents in sequential settings can communicate under mixed motives (most of the related research focused on fully cooperative situations).

In such a setting, the sender optimizes its own expectation while considering the incentive compatibility of the receiver, and this constraint is relatively weak. The rationality of the receiver assumes it follows the sender's instructions even if it leads to just a slightly better posterior payoff. Therefore, the receiver is in a relatively disadvantaged position in equilibrium. We discussed the potential of a far-sighted receiver to protect itself from excessive information exploitation in Section 4.5.3. Furthermore, the receiver's capability can be enhanced through the reputation mechanism in social science. Multiple receivers can evaluate the sender and refuse to cooperate with those who have a poor reputation. This compels the sender to reveal more information based on variants of obedience constraints. The insight behind this process is similar to the ultimatum game [54].

Additionally, since the optimization objective of persuasion is the sender's benefit, its equilibrium may not necessarily be advantageous for social welfare. We discussed situations where the sender can "selfishly" optimize social welfare in Section 3. Practice use of information design should take social welfare into some serious consideration.

# J Limitations and Future Works

Our current model only considers the scenario where one sender persuades one receiver. For more complex situations, additional factors need to be taken into account, which are discussed in Appendix B. The setting of multiple senders is not addressed in this manuscript. The setting of multiple receivers is formulated in the appendix, but $\varphi_\eta : S \to \Delta(\boldsymbol{\Sigma})$ has a very large signal space. If one uses $\Delta(\boldsymbol{\Sigma}) = \Delta(\Sigma_1) \times \cdots \times \Delta(\Sigma_J)$, it reduces to the one-receiver formulation. More subtle factorization of the signaling scheme is yet to be discussed. The setting of multiple receivers is yet to be empirically tested.

Our work invokes many future directions. One problem is to consider multiple senders, where the game between the sender and the receiver and between the sender and other senders co-exist. Extending the results to multiple senders will cover a wider range of real applications. Another direction is to consider the hyper gradient of the receiver's action policy concerning the sender's signaling scheme on the equilibria. This hopefully will provide a more accurate description of the learning process. Additionally, one may also use our framework to investigate a far-sighted receiver. By arming the receiver with the awareness of the sender updates, they could learn not to respect the sender, even if it is more rewarding immediately, for a better equilibrium in the long run.