Cheap Talking Algorithms*

Daniele Condorelli† Massimiliano Furlan‡

December 11, 2023

Abstract

We simulate behaviour of independent reinforcement learning algorithms playing the Crawford and Sobel (1982) game of strategic information transmission. We show that a sender and a receiver training together converge to strategies approximating the ex-ante optimal equilibrium of the game. Communication occurs to the largest extent predicted by Nash equilibrium. The conclusion is robust to alternative specifications of the learning hyperparameters and of the game. We discuss implications for theories of equilibrium selection in information transmission games, for work on emerging communication among algorithms in computer science, and for the economics of collusions in markets populated by artificially intelligent agents.

*We thank Françoise Forges, Alkis Georgiadis-Harris, Balazs Szentes, ChatGPT4, and seminar participants at the University of Warwick for their valuable comments. We thank Sophia Skenderis for her research assistance. All computations have been performed using the Julia programming language. The code is available at `https://github.com/massimilianofurlan/rl_cheap_talk`.

†Department of Economics, University of Warwick, UK. Email: d.condorelli@warwick.ac.uk.

‡Department of Economics, University of Warwick, UK. Email: massimiliano.furlan@warwick.ac.uk.

1 Introduction

Consider the classic signalling game: a sender is informed about a payoff-relevant state of the world drawn from a known distribution and takes one of several possible actions; an uninformed receiver observes the action but not the state, and makes a decision. In a landmark paper, Crawford and Sobel (1982) (henceforth CS) showed that, even if the payoff of both agents is independent of the sender's action, there are equilibria where the action transmits information about the state, as long as the conflict of interest between the agents about the ideal receiver's decision is not too large. By interpreting the payoff-irrelevant actions of the sender as "cheap talk", CS delivers a powerful formal theory of communication. Non-committal and purely symbolic behaviour can convey information and help coordinate subsequent interactions even if rational agents do not share identical goals.

In this paper, we compute stationary points of independent reinforcement learning algorithms playing the CS's game of information transmission.[1] These algorithms work roughly as follows. For each of a finite set of states, the sender keeps track of a vector, which stores its current estimates of the value of taking each action in that state. The receiver, instead, holds a vector for each of the signals the sender may send. Any such vector contains the receiver's estimate of the value of each action following a given signal. In each period, the algorithms select actions following a softmax policy. Most likely, they take the highest-reward action according to their estimates, but with some probability they experiment with different actions. Such probability decays over time, depending on a hyper-parameter (i.e., the temperature-decay factor). After both agents have moved, the relevant estimates are updated to account for the payoffs received. Another hyper-parameter (i.e., the learning rate) establishes how much the current experience is weighted vis-a-vis the past.[2]

Our main finding is that a sender and a receiver training together converge to behaviour with sizeable information transmission. The mutual information between the distribution of the state and that of the action taken by the sender (i.e., the informativeness of the sender's cheap talk) is very close to the level arising in the maximally informative and Pareto optimal equilibrium in CS, for any given level of the bias that parameterises the conflict of interest. Both the sender and the receiver (nearly) best respond to each other and obtain payoffs close to the theoretical benchmark. Hence, the receiver is not misled by the sender, nor is the sender forgiving communication opportunities. Despite the fact that Nash equilibria are focal points of convergence for reinforcement learners, this result is not *a priori* obvious since there are many equilibria in the CS game, including an uninformative "babbling" one.

Language is, eminently, a social phenomenon. Therefore, it is natural to ask whether the success of communication in one-on-one settings extends to environments where multiple agents learn by

[1]Computational techniques are necessary because finding limit points of independent learning algorithms training together is, to date, an intractable problem. The available methods, which rely on approximation through systems of differential equations are not applicable here (e.g., see Börgers and Sarin (1997) and Banchio and Mantegazza (2022)).

[2]The machine learning literature has proposed numerous learning algorithms. Given the simplicity of the task at hand, we chose one of the purest forms of reinforcement learning. Since results align with the best game-theoretical benchmark, we do not expect more complex algorithms to perform less successfully.

interacting in a casual way. We confirm this is the case by considering a scenario where multiple senders and receivers are, at each iteration in the learning process, randomly matched. We therefore keep track of the value estimates of all the agents in our population and stop the learning algorithms when they all have converged. Our simulations show that, despite requiring more time to converge, agents are able to learn a common language. All senders encode information in the same way, by using an identical mapping from states to signals, and all receivers decode signals in a similar manner, leading them to choose nearly the same actions given any signal. This common language delivers payoffs analogous to those in our baseline scenario to all agents, no matter who interacts with whom once policies have been learned.

Having outlined our main findings, in the remainder of the introduction we elaborate on the motivations for this work and the significance of the results by discussing how we aim to contribute to the literature in three distinct fields: computer science, economics and game theory.

Computer science. While experimental evidence shows that informative communication in cheap talk games with partial conflict of interests is achieved by human subjects (e.g., see Blume et al. (2020) for a survey), to our knowledge an analogous conclusion has not yet been robustly established for artificially intelligent agents (AI agents). Most of the machine learning literature has focused on games with common interest, observing that AI agents learn to communicate successfully (e.g., see Lazaridou et al. (2016), Havrylov and Titov (2017), Foerster et al. (2016)). Instead, mostly negative results have been obtained in games where there is scope for information exchange but agents have conflicting interests (e.g., see Cao et al. (2018)). An important exception is Noukhovitch et al. (2021). They consider a CS game played on a circle, for which equilibrium characterization is not available. Employing AI agents controlled by neural networks they show that some degree of communication is achieved even when the bias of the sender is non-zero. We depart from Noukhovitch et al. (2021) by employing simple reinforcement learners and by looking at the original (discretized) CS game. Doing this allows us to compare simulation outcomes to the theoretical benchmark and establish that communication takes place at the highest level predicted by theory even when a very simple model of learning is adopted.

Economics. The observation that private information can be successfully communicated between AI agents, opens up new questions within a growing literature in economics which, motivated by policy concerns, looks at AI agents playing various market games. Contributions to this recent literature include Calvano et al. (2020), Banchio and Skrzypacz (2022), Asker et al. (2022), Johnson et al. (2023) and Decarolis et al. (2023).[3] A central theme of this research agenda is showing that AI agents learn to play strategies that deliver supra-equilibrium profits, which would be deemed implicitly collusive if played by humans.

Two questions come to mind in light of our findings, which we hope will stimulate further work. First, since communication expands the equilibrium set in a game-theoretic sense (e.g., see Aumann and Hart (2003)), what outcomes should we expect in market games played by algorithms

[3]The literature on market games played by AI agents was initiated by computer scientists with early contributions including Waltman and Kaymak (2008) and Tesauro and Kephart (2002) among others.

if collusion can be explicit? This is not a moot concern, even when a direct communication channel is not part of market design. In fact, as auction practice has shown, bidders learn to exchange information in very imaginative ways, for instance by using the last digits of their submitted bids.[4] Since we expect sophisticated AI agents to exploit all communication opportunities, our results suggest that explicit collusion between algorithms with a sufficiently large state space and a long history of interaction may be as worrisome as the implicit one uncovered by the existing literature.

Second, would collusion emerge when agent valuations for the goods being sold are private information and potentially change period by period? While the existing literature has focused on market games with complete information, it is well known that asymmetric information hinders collusion but does not necessarily eliminate it, especially if bidders can communicate.[5] Our results indicate that AI bidders might be able to implement successful collusive schemes even under asymmetric information if they are able to identify a channel for cheap information exchange.

Game Theory. Following pioneering work in psychology (e.g., Bush and Mosteller (1955)) and in game theory (e.g., Erev and Roth (1998)), we can interpret reinforcement learning agents as simplified models of human subjects, in the spirit of the bounded rationality approach to the modelling of economic behaviour.[6] Then, our results complement the game theoretic approach to communication developed by CS. In particular, we show that information transmission in cheap talk games is a robust feature of play, emerging also from alternative modelling approaches to strategic interaction. As was hoped for by Erev and Roth (1998), simple reinforcement learning algorithms seem to fit the experimental data better than equilibrium does. In fact, slight under-communication when the most informative equilibrium predicts perfect information transmission, and tangible over-communication when it predicts partial or no communication, seem to be robust features of experimental implementations of the CS game involving human subjects. (see Dickhaut et al. (1995) and, especially, Cai and Wang (2006)).

In a similar vein, our work may also complement the vast game-theoretic literature on equilibrium selection in games with information transmission. These games, as we have mentioned, normally have multiple Nash equilibria, including a completely uninformative one. Our main result then agrees with the consensus reached in the uniform-quadratic setting around the selection of the most informative and Pareto optimal equilibrium. In fact, focusing on the most informative equilibrium was advocated by CS themselves and in the majority of subsequent work (see Chen et al. (2008) for a key contribution to this literature). Most closely related to our work along this direction is perhaps the evolutionary and learning approach to selection, given the connection between limit points of reinforcement learning and evolutionary dynamics elucidated in Börgers and Sarin (1997). Research using this methodology

[4]There is evidence that in some FCC spectrum auctions bidders used such form of code-bidding to communicate their intentions and avoid competing on the same portions of the spectrum for sale (see Bajari and Yeo (2009)).

[5]On how asymmetric information can reduce collusion see Ortner and Chassang (2018). On collusion with incomplete information see McAfee and McMillan (1992), Marshall and Marx (2012), Che et al. (2018). These papers discuss both explicit collusion (strong cartels) and implicit collusion (weak cartels).

[6]Erev and Roth (1998) wrote: “well-developed, cognitively informed adaptive game theory will complement conventional game theory, both as a theoretical tool and as a tool of applied economics.”

also finds that, when stable outcomes in the CS game exist, they tend to be informative (e.g., see Blume et al. (1993) and Gordon et al. (2022)).

In the next section, we present our simulation design. Section 3 presents the results obtained in a baseline scenario where agents play the classic uniform-quadratic specification of CS with learning hyperparameters that deliver quick convergence. In Section 4 we illustrate the robustness of our findings, both in terms of hyperparameters and the parameters of the game. Section 5 concludes with some avenues for future work.

2 RL agents playing the cheap talk game

We now present the key elements of the environment we study. We start by describing the discretized game of information transmission. Then, we introduce the reinforcement learning algorithms. The details of the simulations we performed and the results are in the next section.

In the cheap talk game we consider, there are two agents, a sender (S) and a receiver (R). At the outset, a state θ is drawn from a known uniform distribution p with support over a finite set Θ, which is composed by n uniformly spaced points in the interval $[0, 1]$. The sender privately observes the realized θ and sends a message $m \in M$ to the receiver, with $|M| = |\Theta|$. Then, the receiver observes message m and takes an action $a \in A$, with A formed by $2n-1$ uniformly spaced points in $[0, 1]$. The receiver wants the action to match θ. Her payoff is $u_R(\theta, a) = -(a - \theta)^2$. Given some bias $b \in [0, \infty)$, the sender wants the action of the receiver to match $\theta + b$. Thus, his payoff is $u_S(\theta, a) = -(a - \theta - b)^2$. The bias parameter measures their divergence of interest.

Frug (2016) (Proposition 2) shows that in the model above, the set of Pareto efficient equilibria is a singleton. This equilibrium, which also exists in the CS's version of the model with a continuous state space, is referred to as the "ex-ante optimal" equilibrium. As this will be useful later, we denote the ex-ante expected utility of receiver and sender in the ex-ante optimal equilibrium, computed using Frug (2016)'s algorithm, with $\bar{U}_R(b)$ and $\bar{U}_S(b)$.[7] A so-called "babbling" equilibrium exists also in the discretized version of the model. In this equilibrium, the sender's strategy is independent of the state and the receiver plays her ex-ante optimal action. We denote with $\underline{U}_R(b)$ and $\underline{U}_S(b)$ the ex-ante expected utilities in the babbling equilibrium. Note that $\underline{U}_R(b)$ does not depend on b.[8]

We let two independent reinforcement learning agents play, as sender and receiver, the discretized cheap talk game. To allow learning, the two agents play the game multiple times, up to a maximum

[7] Frug (2016) endows the receiver with a continuum of actions. This ensures that a single optimal action corresponds to each belief the receiver may have. Given our discretization of Θ and A, the optimal action is unique as long as the strategy of the sender is partitional. Hence, in principle, the receiver might find herself indifferent for some beliefs generated by non-partitional strategies of the sender. In this case, there might exist an equilibrium in our game different from the one identified in Frug (2016) that is also Pareto optimal. The Pareto optimal equilibrium built from Frug (2016) remains an optimal equilibrium in our setting because it is partitional.

[8] The existing literature does not offer a complete characterization of equilibria. Frug (2016) shows that even restricting to partitional equilibria comes with a loss.

of $T = 10^7$ periods (or episodes). Both are programmed to take an action conditional on a state, first the sender and then the receiver. In each period, a state for the sender is drawn from Θ according to p, independently of previous interactions. Then, the sender takes an action from M, which represents the state for the receiver. Finally, the receiver takes an action from A and agents collect their rewards. Because the underlying learning model is the same for both agents (i.e., both take action conditional on some state), we now describe it for a generic agent, with states and actions taking appropriate meaning based on which agent is playing.

Let $\mathcal{S}$ be the finite set of possible states and $\mathcal{A}$ the finite set of actions, for either the sender or the receiver. Each time $t \in \{1, 2, \ldots, T\}$ an agent is called to play in state $s \in \mathcal{S}$, it chooses action $a \in A$ following a parameterized softmax probability distribution

$$\pi_t(a \mid s) = \frac{e^{Q_t(s,a)/\tau_t}}{\sum_{a' \in \mathcal{A}} e^{Q_t(s,a')/\tau_t}},$$

where $Q_t(s, a)$ (discussed in the next paragraph) represents the agent's estimate in period t of the value of taking action a in state s. The parameter τ_t, called temperature, modulates the intensity of exploration: for smaller values of τ_t, the probability mass increasingly concentrates on the action(s) that are most rewarding according to the current estimate $Q_t(s, a)$. We reduce exploration at each interaction by letting the temperature decay according to $\tau_t = \lambda \tau_{t+1}$, where $\lambda \in [0, 1)$ is the decay rate and $\tau_1 = 1$. Hence, the exploration goes to zero in the limit as $t \to \infty$.

The initial estimate, $Q_0(s, a)$, is arbitrarily initialized for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. If the agent takes action a in state s in period t, the estimate associated with that specific state-action pair is updated iteratively according to

$$Q_t(s, a) = Q_{t-1}(s, a) + \alpha \left[r_t(s, a) - Q_{t-1}(s, a) \right],$$

where the step-size parameter $\alpha \in (0, 1]$, called learning rate, regulates how quickly new information replaces the old and $r_t(s, a)$ (discussed in the next paragraph) denotes the reward the agent obtains by playing action a in state s in period t. For all other (s', a') pairs, $Q_t(s', a') = Q_{t-1}(s', a')$.

In multi-agent reinforcement learning, the reward that an agent obtains is not drawn from a stationary distribution, as it generally depends on the action taken by the other agent. In particular, let (a', s') be the pair of state and action taken by the other agent in t. Then, we have $r_t(s, a) = -(a' - b - s)^2$ for the sender's algorithm and $r_t(s, a) = -(a - s')^2$ for the receiver.

If the distribution of r_t were to depend only on the agent's own actions, existing results would guarantee convergence of the policy $\pi_t(\cdot \mid s)$ to an optimal one. However, because the underlying distributions of rewards the agents face are non-stationary, convergence is not guaranteed. For this reason, we consider agents to have converged and stop the simulation if, before reaching the maximum number of interactions T, the policies of both agents exhibit relative deviations in $L_{2,2}$ norm smaller than 0.1% for $K = 10^4 < T$ consecutive interactions.

Pseudocode for the simulation is given in Algorithm 1 below.

Algorithm 1 Independent reinforcement learning in the discretized cheap talk game

Initialize Q^S and Q^R arbitrarily
for each episode **do**
  $\theta \sim p(\theta)$
  $m \sim \pi^S(m \mid \theta)$
  $a \sim \pi^R(a \mid m)$
  $Q^S(\theta, m) \leftarrow Q^S(\theta, m) + \alpha[u_S(\theta, a) - Q^S(\theta, m)]$
  $Q^R(m, a) \leftarrow Q^R(m, a) + \alpha[u_R(\theta, a) - Q^R(m, a)]$
  if π^S and π^R converged **then break**
end for

3 Baseline Results

In this section, we discuss the baseline simulation we have singled out to present our main results. The robustness of our findings is demonstrated in the next section.

For our base-case, we consider the discretized cheap talk game with $n = 21$ states in $[0, 1]$, so that any two adjacent states are separated by a 0.05 increment. Hence, $\Theta = \{0.00, 0.05, \ldots, 0.95, 1.00\}$, $M = \Theta$ and $A = \{0.00, 0.025, \ldots, 0.975, 1.00\}$.[9]

We implement algorithms for both the sender and the receiver that use the same learning rate $\alpha = 0.1$ and exploration decay $\lambda = 0.99999$. The decay parameter was chosen to deliver the minimal amount of exploration, thus time of play, necessary for the agents to achieve full communication when $b = 0$. The learning rate is the one commonly used in applications.[10] The Q-matrices of the sender and of the receiver have dimensions 21×21 and 21×41, respectively. Their entries are initialized using a uniform distribution in the interval $[\underline{U}_S(b), 0]$ for the sender, and $[\underline{U}_R(b), 0]$ for the receiver.[11]

We study interactions for different levels of bias taking 51 points spaced 0.01 apart from each other in the interval $[0, 0.5]$. For each level of bias b, we repeat the same simulation 1000 times. At the end of each simulation, if the agents' policies have converged, we record the Q-matrices at the point of convergence and compute the implied policies for the sender and receiver, denoted $\pi^S_\infty(\cdot \mid \theta)$ and $\pi^R_\infty(\cdot \mid m)$, respectively. Using these policies we can compute the ex-ante expected rewards of the agents from playing the information transmission game together. These are

$$U_S = -\sum_\theta p(\theta) \sum_m \pi^S_\infty(m \mid \theta) \sum_a \pi^R_\infty(a \mid m)(a - \theta - b)^2,$$

$$U_R = -\sum_\theta p(\theta) \sum_m \pi^S_\infty(m \mid \theta) \sum_a \pi^R_\infty(a \mid m)(a - \theta)^2.$$

[9] We confirmed via additional simulations that endowing the sender with a message space larger than the state space does not affect the final results.

[10] With $\alpha = 0.1$, the weight rewards have in the estimate is less than 1% after 23 interactions and with $\lambda = 0.99999$ the temperature is approximately 10^{-3} after 6.9×10^5 interactions. After that number of iterations the probability mass of policies is concentrated around a few relatively highly rewarding actions.

[11] We confirmed via additional simulations that the initialization of the matrices is irrelevant for the final results.

In the next Figure 1, we compare the average ex-ante payoffs arising from the simulations to the theoretical bounds provided by the babbling equilibrium and the ex-ante optimal equilibrium for the different levels of bias in the discretized $[0, 0.5]$ interval.

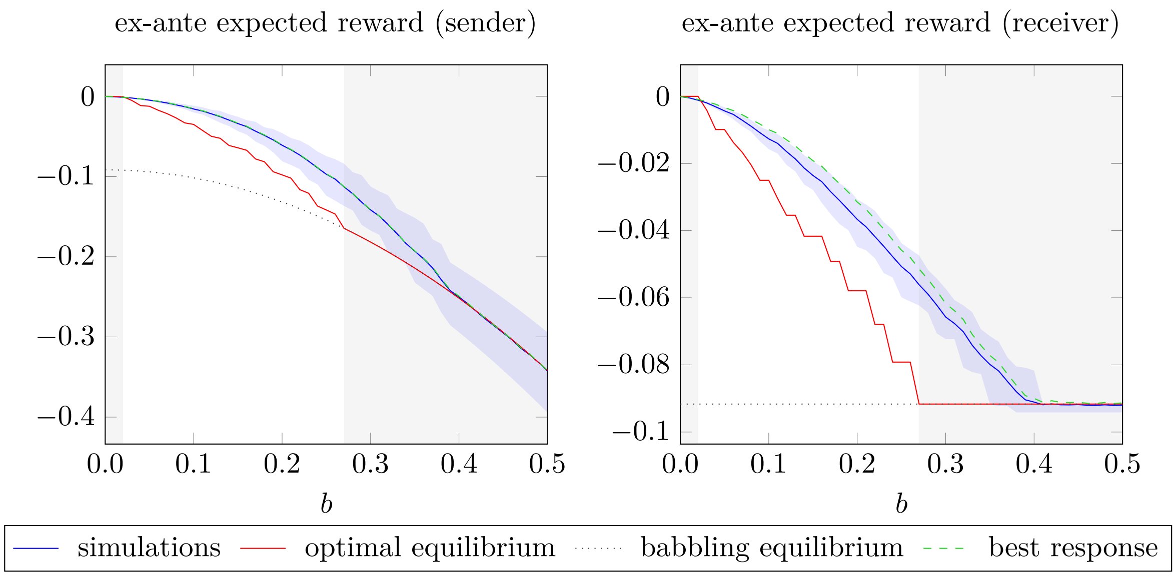

Figure 1: Ex-ante expected reward for the sender (left) and receiver (right) for different levels of bias. The average value across 1000 simulations is shown in blue; 95% of the outcomes fall inside the blue-shaded areas. The value associated with the ex-ante optimal equilibrium is in red and the one associated with the babbling equilibrium is dotted gray.
Also applies to all subsequent figures: The ex-ante optimal equilibrium entails perfect information transmission for biases identified by the shaded grey area to the left, while babbling is the unique equilibrium for biases in the shaded grey areas to the right; Green dotted lines indicate payoffs that agents would get by best-responding.

The two panels illustrate that communication between the sender and the receiver is successful and at the highest level predicted by theory. At any level of the bias, ex-ante payoffs of both the sender and the receiver (blue lines) are in line and often even exceed those arising in the ex-ante optimal equilibrium (red lines). In particular, learned behaviour closely matches equilibrium when the bias is very high (i.e., no communication is the only equilibrium) or very low (i.e., perfect information transmission is the ex-ante optimal equilibrium outcome). When the ex-ante optimal equilibrium entails partial communication, AI agents always tend to exchange more information than Nash equilibrium predicts. However, as we illustrate in Section 4, for given level of exploration, the fewer the states, the closer the agents' payoffs get to the ones in the ex-ante optimal equilibrium.

This finding can be reinforced by looking at a direct measure of communication implied by the sender's policy. To this end, we compute the mutual information between state and message, normalised by the entropy of the message. This is equivalent to the percentage reduction in the entropy of the state arising from knowing the message. Formally,

$$I = \left(\sum_{\theta} p(\theta) \log \left(\frac{1}{p(\theta)} \right) \right)^{-1} \sum_{\theta} \sum_{m} \pi_{\infty}^{S}(m \mid \theta) p(\theta) \log \left(\frac{\pi_{\infty}^{S}(m \mid \theta)}{\sum_{\theta} \pi_{\infty}^{S}(m \mid \theta) p(\theta)} \right).$$

This metric takes value 1 if knowledge of the message implies knowledge of the state, as in the perfectly informative equilibrium. It takes value 0 when state and message are statistically independent, as in the babbling equilibrium.

As Figure 2 below illustrates, at each level of the bias, the average normalised mutual information from our simulations (in blue) is in line with and often exceeds the one obtained for the ex-ante optimal equilibrium (in red). An analogous result is obtained by measuring the normalised mutual information between states and actions, indicating that at convergence the receiver correctly decodes the information contained in the messages.

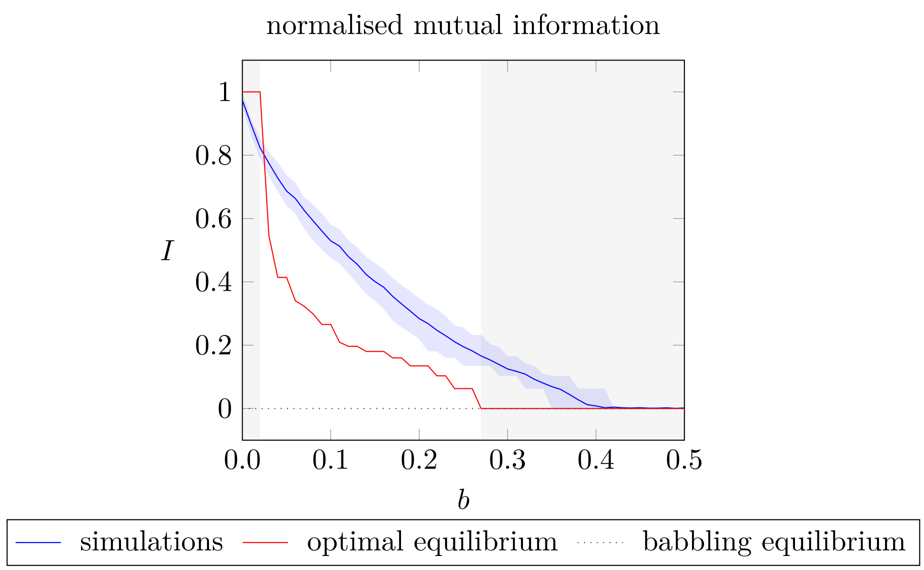

Figure 2: Mutual information between the distribution of messages induced by the sender's policy and the distribution of states of the world. Average across 1000 simulations is shown in blue; 95% of simulation outcomes fall inside the shaded area. The value associated with the optimal equilibrium is in red and the one associated with the worst equilibrium is dotted gray.

The observation that both the payoff and the mutual information implied by the sender's strategy at convergence are close to those in the ex-ante optimal equilibrium indicates that the strategy played at convergence is also close, in terms of the distribution of posteriors it generates, to that played in the ex-ante optimal equilibrium benchmark. This, together with the fact that the receiver obtains a payoff in line with that arising from play of the ex-ante optimal equilibrium, also suggests that the receiver's strategy at convergence is near, in terms of response to the implicit posterior induced by a message, to the one played in the theoretical benchmark.

These results paint a rosy picture for algorithmic communication. Artificially intelligent agents often learn to communicate more than in the ex-ante optimal equilibrium. However, when this happens, they are not best responding to each other. Then, the question that arises is how close to equilibrium are the sender and receiver playing. In fact, it may be argued that agents are not learning robustly to communicate unless they are playing close to an equilibrium. To address this issue, in Figure 3 we measure how distant the sender and the receiver are playing from Nash equilibrium. We compute the additional ex-ante expected reward they would achieve if, instead of playing the learned policy, they best replied to the policy learned by the opponent.

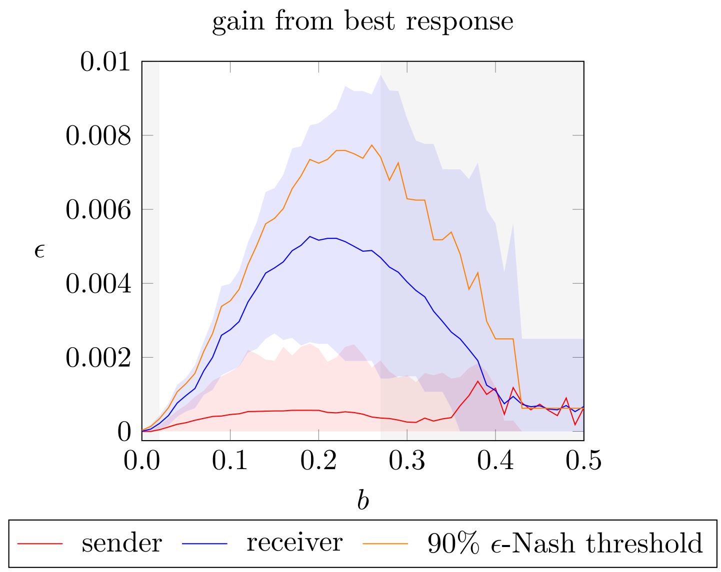

Figure 3: Potential gains from best responding to the opponent. Average value over 1000 simulations; 95% of simulation outcomes fall inside the shaded areas.

Consistently with our previous observations, Figure 3 indicates that agents are playing further away from equilibrium at intermediate levels of the bias. The maximal gain the receiver (sender) obtains on average from best-responding is when the bias is around 0.2 (0.4). At that level of bias, the receiver (sender) could gain around 0.005 (0.001) on average from best-responding, which is about 10% (5%) of her payoff given the learned policy. This suggests the loss of payoff from playing the learned policy compared to best-responding is not large. In 90% of our simulations, agents converge to play, in the worst case scenario, an ϵ-equilibrium (Radner, 1980) with ϵ equal to 0.008.

In addition, as we show in Section 4, allowing for more exploration and a longer time to convergence results in agents getting closer to equilibrium play. As a theoretical matter, the result that agents often do better than equilibrium play should not come as a surprise. It is a common phenomenon, which can be explained by the complex dynamic system generated by the two algorithms learning together (see Banchio and Mantegazza (2022)).

We conclude this section by presenting the results of simulations in which 10 senders are randomly matched to 10 receivers at each iteration of the learning process. This setup introduces additional learning difficulties. For example, senders who have been given positive feedback regarding a certain policy while training with some receivers might find themselves interacting with other receivers with a very different interpretation of the same messages given their own past history of interactions. Nonetheless, we show that agents are able to learn a common language. That is, at any level of the bias, all senders within any given simulation employ the same policy, mapping states to messages, and all receivers take approximately the same action for any given message received. Consider the following figures.

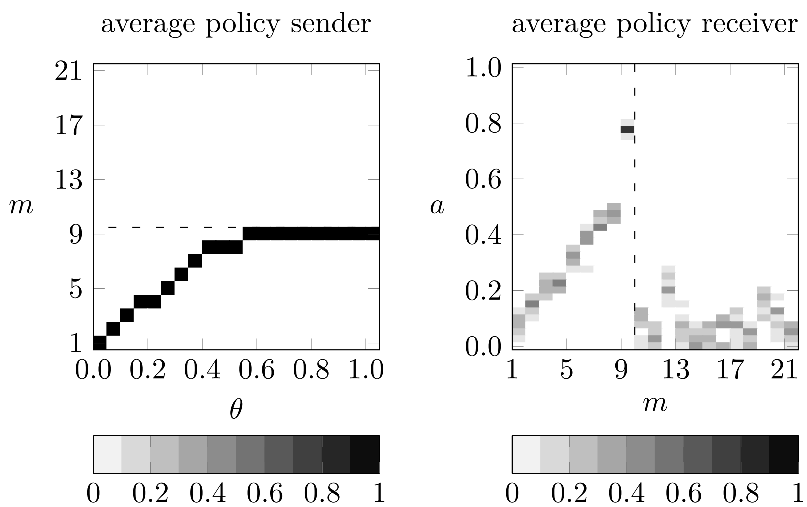

Figure 4: Average policy across the 10 senders (left) and 10 receivers (right) in a single simulation with $b = 0.1$. The mean deviation from the modal policy averages 0 across senders and 0.017 across receivers. Messages above or to the right of the dotted line are played with negligible probabilities and are considered off-the-path.

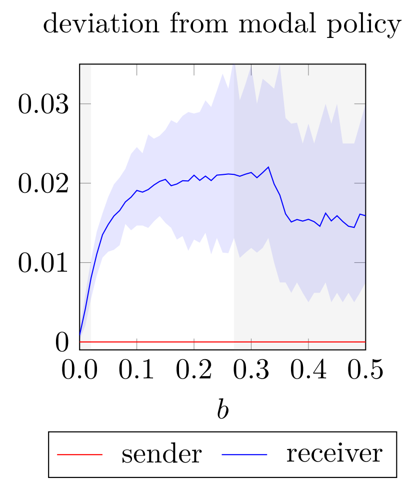

Figure 5: Mean deviation from modal policy. Average over 100 simulations; 95% of outcomes fall inside the shaded areas. Computed with only on-the-path messages.

Figure 4 above demonstrate a high level of homogeneity in the policies learned by senders and receivers present within a single simulation. There is hardly any variability in senders' behaviour at any given state. There is only minor variability in the receivers' decoding of messages, once messages from 10 to 21 are excluded because they are not sent by senders following convergence. Figure 5 shows that, on average across 100 simulations for each bias level, the mean deviation from the modal action played by the set of receivers within a simulation following a given message (played with non-negligible probability) is at most 0.02. Roughly speaking, this corresponds to receivers differing in their reaction to a message by, on average, taking the nearby action to that played by the majority of them. Notably, the less communication takes place, that is the higher the bias, the more vague the language.

In addition, when playing together after learning is completed, any two agents achieve payoffs analogous to those obtained in the baseline scenarios. Therefore, also in this case communication is at the highest level possible predicted by equilibrium. We omit to visually present the results as the difference with the baseline case (Figure 1 and Figure 2) is not noticeable.

4 Robustness

In this section, we demonstrate that communication emerges robustly in CS games played by AI agents. To do so, we report the results of simulations obtained for a wide variety of alternative assumptions. We first keep the game fixed and we look at the effect of employing different learning hyperparameters. Then, we look at different specifications of the information transmission game. We consider a higher and lower number of states, non-uniform distributions of the state, and utility functions that are not linear-quadratic.

4.1 Learning parameters

We run our simulations of the cheap talk game for a grid of reinforcement learning hyperparameters. We consider 10 uniformly spaced learning rates in $[0.05, 0.5]$ and 10 different exploration decay rates in $[0.99998, 0.999998]$. The exploration decay rates are spaced such that the number of interactions required to converge scales linearly.[12] Figure 6 below superposes the average ex-ante expected rewards of the agents over 100 simulations for each of the 100 (α, λ) pairs in the discretized $[0, 0.5] \times [0.99998, 0.999998]$ grid.

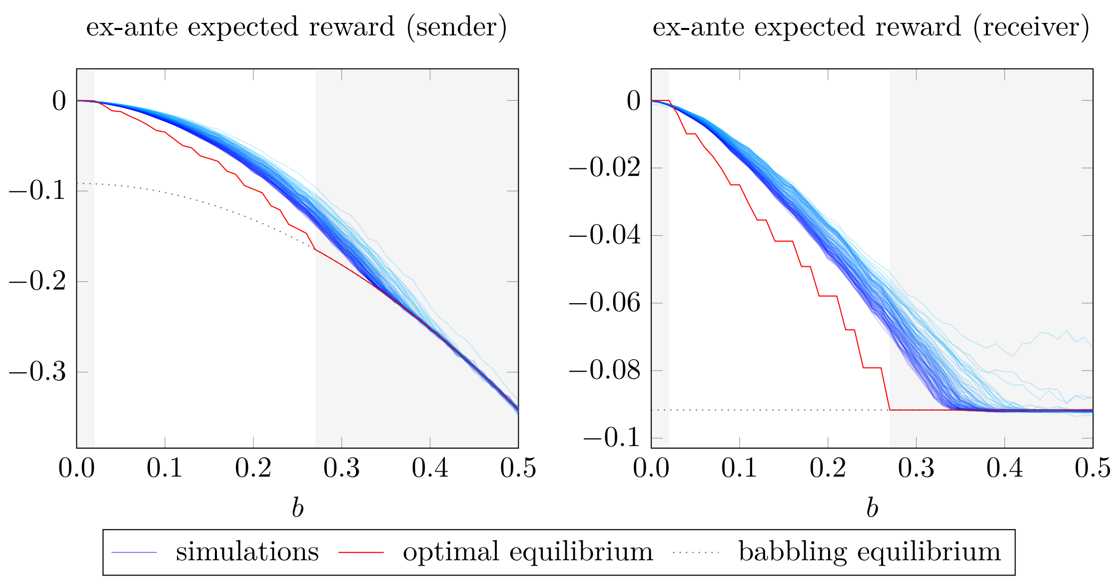

Figure 6: Ex-ante expected reward for the sender (left) and receiver (right) for different levels of bias. Each blue-toned line is the average across 100 simulations with a different learning rate, α, and exploration decay, λ. The lines' hue gets darker as λ gets closer to 1 and agents' exploration increases.

The figure shows that the results described in the previous section extend to a range of different

[12]In practice, with $\lambda = 0.99998$ it takes approximately 5×10^5 interactions for the agents' policies to converge, and with $\lambda = 0.999998$ it takes approximately 5×10^6 interactions.

reinforcement learning hyperparameters' configurations. Moreover, it highlights that letting agents explore more extensively yields outcomes that are progressively closer to the ex-ante optimal equilibrium. The same trend naturally extends to the normalised mutual information between messages and states of the world.

Figure 7 shows, for different combinations of reinforcement learning hyperparameters, the threshold level for ϵ such that 90% of simulations outcomes (across all bias levels) are ϵ-Nash equilibria. The heatmap further confirms that more exploration results in agents playing closer to exact equilibrium behaviour. Conversely, limited exploration results in larger mistakes and overcommunicative outcomes.

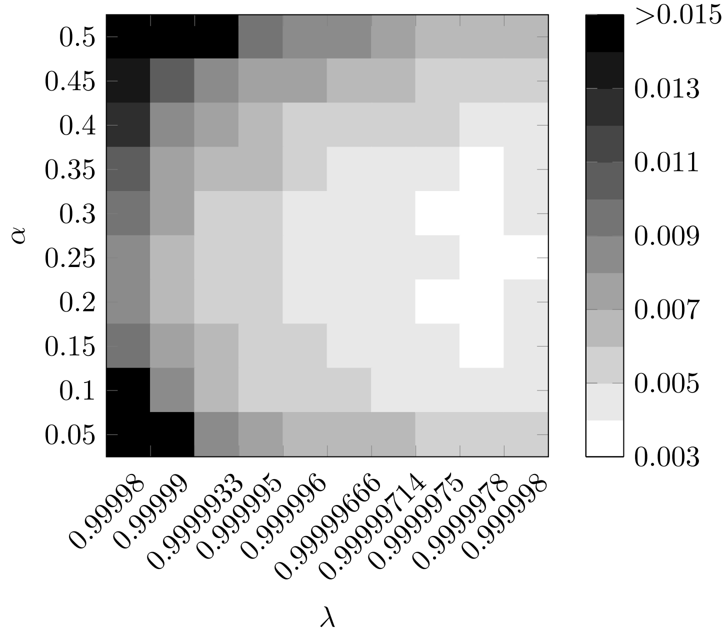

Figure 7: Required level of ϵ to have at least 90% of simulations over all bias levels at an ϵ-Nash equilibrium.

4.2 Game form

We now keep fixed the reinforcement learning hyperparameters as in our baseline configuration and consider variations of the cheap talk game. We show for each case the average ex-ante expected reward of the agents over 1000 simulations. We look at cases with different numbers of states of the world, different utility specifications and different distributions over the states of the world.

In Figure 8 we consider simulations with $n = 6$, $n = 11$ and $n = 41$ states of the world, so that any two adjacent states are spaced 0.2, 0.1 and 0.025 from each other, respectively. The figure shows that agents play closer to the theoretical benchmark when the number of states is small. With a large number of states instead, communication tends to exceed the theoretical benchmark, especially when babbling is the unique equilibrium. This is explained by the relative increase (reduction) in exploration due to the change in size of the agents' Q-matrices. As we keep λ fixed to the base case configuration, each state-action pair is on average visited more (less) often depending on the size of the agent's Q-matrix. This eventually results in improving (worsening) the agent's learning. We see that when n is smaller than our base case, agents explore more in relative terms and are closer to equilibrium behaviour. The opposite is true when n is larger. For the latter case, letting agents explore more extensively at the expense of longer times of play gives back outcomes closer to the theoretical benchmark.

For cases where we vary the utility functions and the distribution of states, existence of the ex-ante optimal equilibrium is no longer guaranteed. Nonetheless, Frug (2016, Proposition 1) shows that the receiver-optimal equilibrium is partitional if the action space is sufficiently large, the utilities are concave and the sender is upwardly biased.[13] Hence, in cases where these three assumptions are satisfied, we rely on the ex-ante receiver-optimal equilibrium as our benchmark (see Figure 9). Instead, in the case of non-uniform distributions of states, where the first of those three assumptions fails, we rely on the receiver-preferred partitional equilibrium, which turns out to exist, as comparator (see Figure 10).

Figure 9 shows simulation outcomes with different utility specifications. We consider the case of a fourth-power loss function and the case of an absolute loss function. To ensure results are not determined by the magnitude of rewards we also consider a scaled-up quadratic utility by a factor of 10. The figure confirms that the high level of communication is not dependent on the specific forms of the utility function. All three scenarios show similar results, in line with our benchmark case. While we have not run further cases because it is hard to identify the right comparator, we strongly suppose that the assumptions of concavity and upward bias are not crucial for the emergence of a high level of communication.

Finally, in Figure 10 we show outcomes for different distributions over the states of the world; namely, a bell-shaped distribution, a probability distribution with linearly increasing probability mass, and one with linearly decreasing probability mass. In this case results also indicate communication in line with the most optimistic theoretical benchmark. A surprising result is obtained in the case of the decreasing distribution. While in all our simulations agents do better than babbling, here the receiver obtains a payoff lower than the babbling one. We think this finding is interesting because the sender seems able to manipulate the receiver even when theoretically it should not be possible and the receiver is losing out from not just playing the ex-ante optimal action. Unfortunately, we do not have a cogent explanation for this result.

[13] The sender is upwardly biased if for all $\theta \in \Theta$ and $a, a' \in \mathbb{R}$ with $a > a'$, if $u_S(\theta, a') \geq u_S(\theta, a)$ then $u_R(\theta, a') > u_R(\theta, a)$.

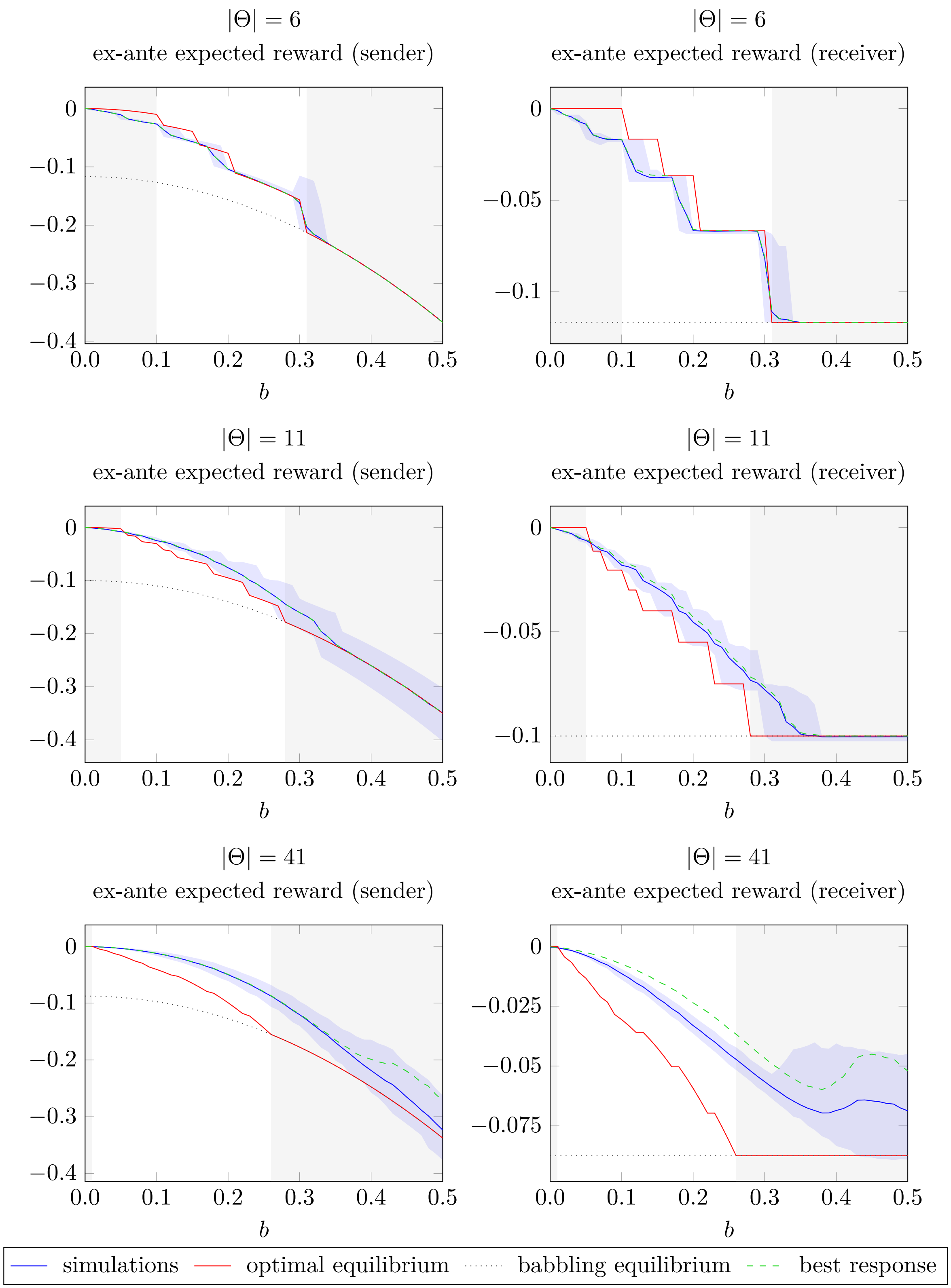

Figure 8: Ex-ante expected reward for the sender (left) and receiver (right) for different levels of bias. Cases with 6 states (top), 11 states (middle) and 41 states (bottom).

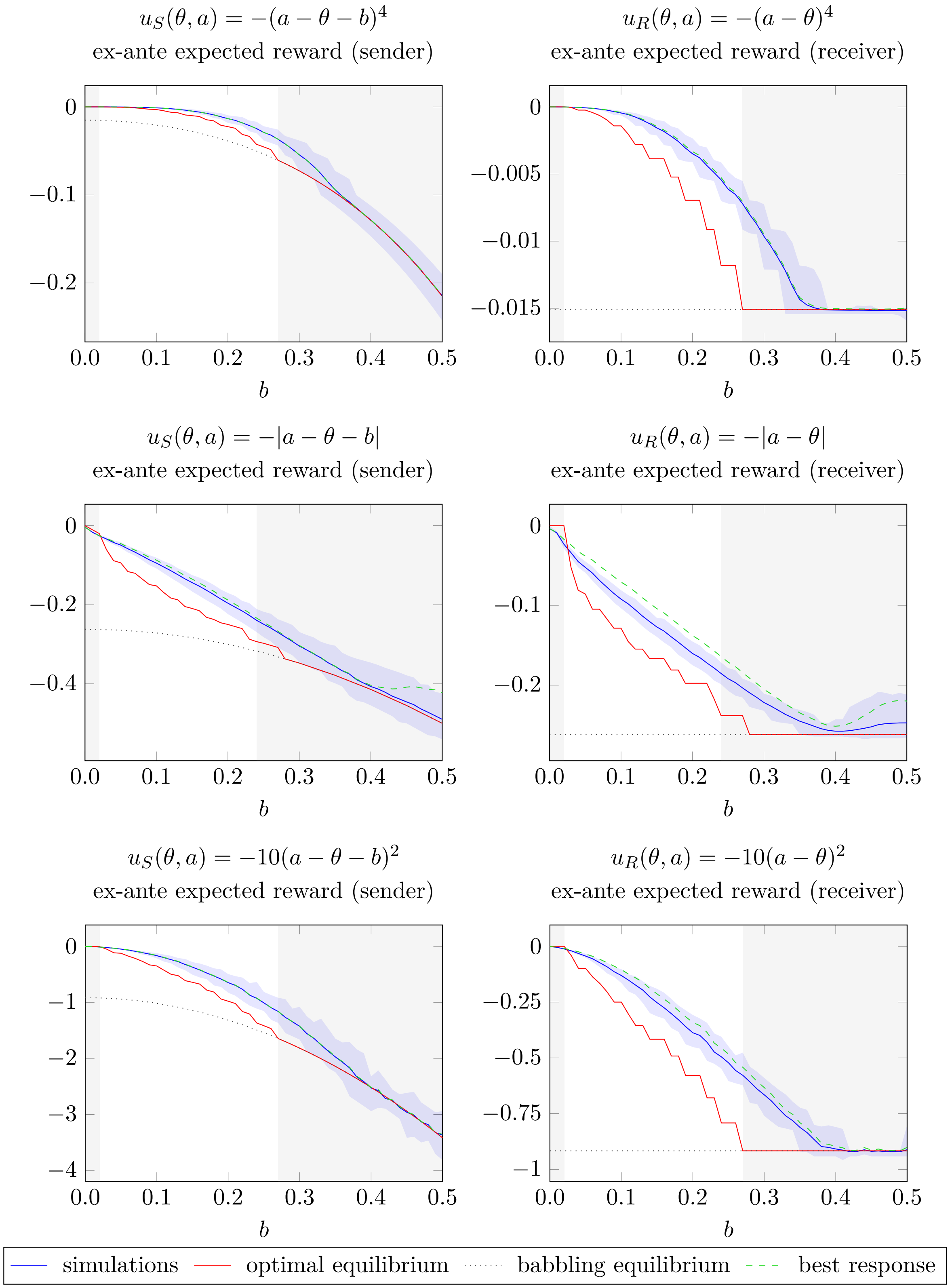

Figure 9: Ex-ante expected reward for the sender (left) and receiver (right). Fourth-power loss (top), absolute loss (middle), scaled quadratic loss (bottom)

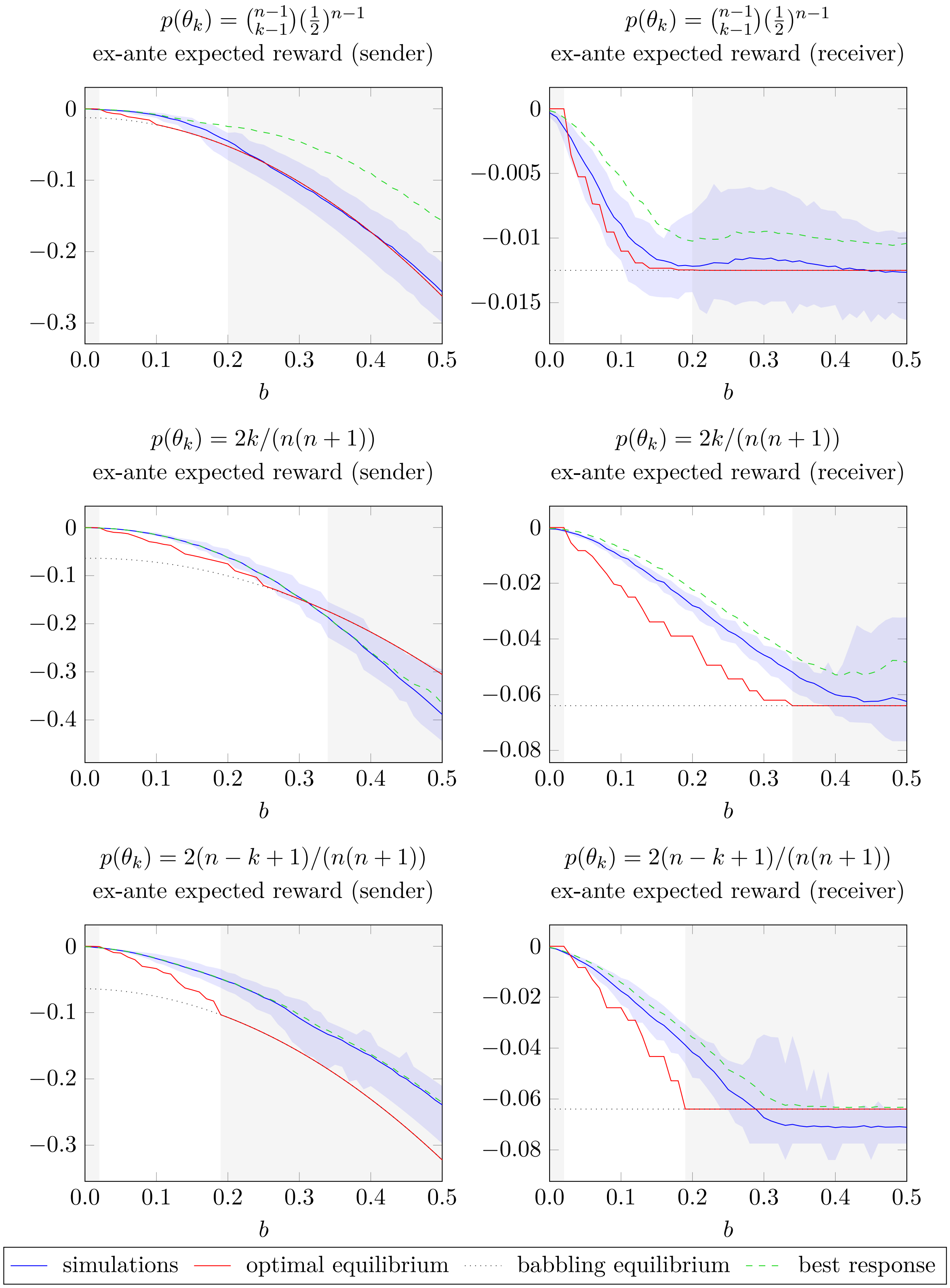

Figure 10: Ex-ante expected reward for the sender (left) and receiver (right) for different levels of bias. Cases with a bell-shaped distribution distribution (top), linearly increasing distribution (middle) and linearly decreasing distribution (bottom). We use $p(\theta_k)$ to indicate the probability mass on the k-th state in $\Theta = \{\theta_1, \ldots, \theta_n\}$. There are $n = 21$ states as in the base-case simulations.

5 Conclusions

We showed that simple reinforcement learning algorithms training together in the classic Crawford and Sobel (1982) cheap-talk game engage in proficuous information transmission and develop a common language. Communication is substantial and matches the level predicted by the most informative equilibrium of the cheap-talk game. This result is robust and extends to the case of a population of agents randomly interacting with each other.

Equilibria in CS exhibit a nice structure. Both sender and receiver unambiguously benefit from more communication. This raises the question of what would happen in games with multiple equilibria that are not Pareto ranked, with some more favourable to the receiver and others to the sender. Will communication break down? Or will one of the two agents lead the other to their favourite equilibrium? While our results in Section 4 suggest that communication will persist and favour the sender, we believe extending the analysis to more general games with communication is an interesting avenue for future work.

Another natural extension of the present framework would be looking at how populations learn a common language when agents are heterogeneous (e.g., senders may have different biases) or the frequency of interactions is not driven by random matching (e.g., agents may be arranged in a network where a number of receivers interact with a single sender). Would agents still be able to learn a common language? Will there be winners and losers depending on the level of bias or the network architecture of interactions? What population structures facilitate learning?

A more speculative next step would be to consider the interaction between humans and algorithms. Suppose we let a human train with an algorithm, or we let a multitude of humans randomly interact with multiple algorithms. Will they learn a common language? Will there be more communication than in human-to-human experiments? Would humans be manipulated or maybe the other way around? Human-algorithm play also raises interesting questions regarding the interaction between strategic signalling and natural language. How would reinforcement learning algorithms endowed with natural language processing abilities, such as those currently possessed by chatGPT and other large language models, perform? Will the use of natural language result in more or less information transmission? Will human agents be more easily deceived? We think that human-AI experiments show promise well beyond the questions raised above.

Finally, it may be worth revisiting some of the existing findings in the economics of AI agents playing market games. For instance, will the sort of code-bidding collusion described in the introduction emerge in market played by AI agents? Since a large state space would be required to handle this sort of "non-verbal" exchange, our finding that communication emerges suggests it may be worth looking at the behaviour of more complex agents, such as those endowed with deep neural networks. It would not be surprising to see collusion sustained at higher levels than those already observed with simple learning algorithms. Such a finding would suggest the need for market design to mitigate communication possibilities, especially when AI agents interact frequently.

References

Asker, J., Fershtman, C., and Pakes, A. (2022). Artificial intelligence, algorithm design, and pricing. *AEA Papers and Proceedings*, 112:452–56.

Aumann, R. J. and Hart, S. (2003). Long cheap talk. *Econometrica*, 71(6):1619–1660.

Bajari, P. and Yeo, J. (2009). Auction design and tacit collusion in fcc spectrum auctions. *Information Economics and Policy*, 21(2):90–100. Special Section on Auctions.

Banchio, M. and Mantegazza, G. (2022). Adaptive algorithms and collusion via coupling.

Banchio, M. and Skrzypacz, A. (2022). Artificial intelligence and auction design.

Blume, A., Kim, Y.-G., and Sobel, J. (1993). Evolutionary stability in games of communication. *Games and Economic Behavior*, 5(4):547–575.

Blume, A., Lai, E. K., and Lim, W. (2020). *Handbook of Experimental Game Theory*, chapter 13, pages 311–347. Edward Elgar Publishing.

Börgers, T. and Sarin, R. (1997). Learning through reinforcement and replicator dynamics. *Journal of Economic Theory*, 77(1):1–14.

Bush, R. R. and Mosteller, F. (1955). *Stochastic models for learning.* John Wiley & Sons, Inc.

Cai, H. and Wang, J. T.-Y. (2006). Overcommunication in strategic information transmission games. *Games and Economic Behavior*, 56(1):7–36.

Calvano, E., Calzolari, G., Denicolò, V., and Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10):3267–97.

Cao, K., Lazaridou, A., Lanctot, M., Leibo, J. Z., Tuyls, K., and Clark, S. (2018). Emergent communication through negotiation. *CoRR*, abs/1804.03980.

Che, Y.-K., Condorelli, D., and Kim, J. (2018). Weak cartels and collusion-proof auctions. *Journal of Economic Theory*, 178:398–435.

Chen, Y., Kartik, N., and Sobel, J. (2008). Selecting cheap-talk equilibria. *Econometrica*, 76(1):117–136.

Crawford, V. P. and Sobel, J. (1982). Strategic information transmission. *Econometrica*, 50(6):1431–1451.

Decarolis, F., Rovigatti, G., Rovigatti, M., and Shakhgildyan, K. (2023). DP18009 artificial intelligence & data obfuscation: Algorithmic competition in digital Ad auctions. (mimeo).

Dickhaut, J. W., McCabe, K. A., and Mukherji, A. (1995). An experimental study of strategic information transmission. *Economic Theory*, 6(3):389–403.

Erev, I. and Roth, A. E. (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *The American Economic Review*, 88(4):848–881.

Foerster, J. N., Assael, Y. M., de Freitas, N., and Whiteson, S. (2016). Learning to communicate with deep multi-agent reinforcement learning. *CoRR*, abs/1605.06676.

Frug, A. (2016). A note on optimal cheap talk equilibria in a discrete state space. *Games and Economic Behavior*, 99:180–185.

Gordon, S., Kartik, N., Pei-yu Lo, M., Olszewski, W., and Sobel, J. (2022). Effective communication in cheap talk games. (mimeo).

Havrylov, S. and Titov, I. (2017). Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. *CoRR*, abs/1705.11192.

Johnson, J. P., Rhodes, A., and Wildenbeest, M. (2023). Platform design when sellers use pricing algorithms. *Econometrica*, 91(5):1841–1879.

Lazaridou, A., Peysakhovich, A., and Baroni, M. (2016). Multi-agent cooperation and the emergence of (natural) language. *CoRR*, abs/1612.07182.

Marshall, R. C. and Marx, L. M. (2012). *The Economics of Collusion: Cartels and Bidding Rings*. The MIT Press.

McAfee, R. P. and McMillan, J. (1992). Bidding rings. *The American Economic Review*, 82(3):579–599.

Noukhovitch, M., LaCroix, T., Lazaridou, A., and Courville, A. C. (2021). Emergent communication under competition. *CoRR*, abs/2101.10276.

Ortner, J. and Chassang, S. (2018). Making corruption harder: Asymmetric information, collusion, and crime. *Journal of Political Economy*, 126(5):2108–2133.

Radner, R. (1980). Collusive behavior in noncooperative epsilon-equilibria of oligopolies with long but finite lives. *Journal of Economic Theory*, 22(2):136–154.

Tesauro, G. and Kephart, J. O. (2002). Pricing in agent economies using multi-agent q-learning. *Autonomous Agents and Multi-Agent Systems*, 5(3):289–304.

Waltman, L. and Kaymak, U. (2008). Q-learning agents in a cournot oligopoly model. *Journal of Economic Dynamics and Control*, 32(10):3275–3293.