

# Multitasking Inhibits Semantic Drift

**Athul Paul Jacob**  
apjacob@mit.edu  
MIT CSAIL

**Mike Lewis**  
mikelewis@fb.com  
Facebook AI Research

**Jacob Andreas**  
jda@mit.edu  
MIT CSAIL

## Abstract

When intelligent agents communicate to accomplish shared goals, how do these goals shape the agents’ language? We study the dynamics of learning in latent language policies (LLPs), in which instructor agents generate natural-language subgoal descriptions and executor agents map these descriptions to low-level actions. LLPs can solve challenging long-horizon reinforcement learning problems and provide a rich model for studying task-oriented language use. But previous work has found that LLP training is prone to semantic drift (use of messages in ways inconsistent with their original natural language meanings). Here, we demonstrate theoretically and empirically that *multitask training is an effective counter to this problem*: we prove that *multitask training eliminates semantic drift in a well-studied family of signaling games*, and show that *multitask training of neural LLPs in a complex strategy game reduces drift and while improving sample efficiency*.

## 1 Introduction

A major goal in the study of artificial and natural intelligence is to understand how language can scaffold more general problem-solving skills (e.g. Spelke, 2017), and how these skills in turn shape language itself (e.g. Gibson et al., 2017). In NLP and machine learning, **latent language policies** (LLPs; Andreas et al., 2018) provide a standard framework for studying these questions. An LLP consists of *instructor* and *executor* subpolicies: the instructor generates natural language messages (e.g. high-level commands or subgoals), and the executor maps these messages to sequences of low-level actions (Fig. 1). LLPs have been used to construct interactive agents capable of complex reasoning (e.g. programming by demonstration) and planning over long horizons (e.g. in strategy games; Hu et al., 2019). They promise an effective and interpretable interface between planning and control.

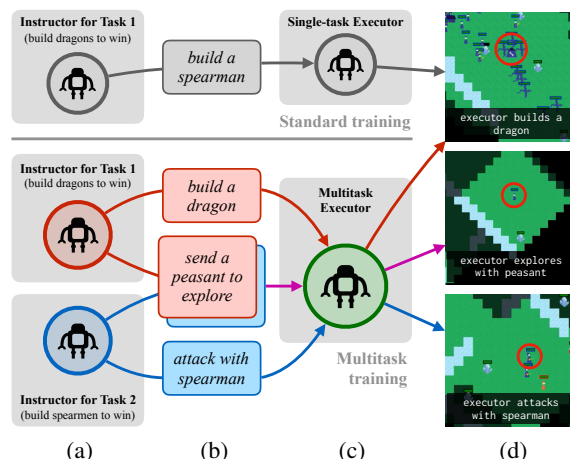


Figure 1: In latent language policies, instructor agents (a) send natural-language commands (b) to executor agents (c), which execute them in an interactive environment (d). Jointly trained instructor–executor pairs learn to use messages in ways inconsistent with their natural language meanings (top shows a real message–action pair from a model described in Section 4). We show that *multitask training with a population of task-specific instructors stabilizes message semantics and in some cases improves model performance*.

However, they present a number of challenges for training. As LLPs employ a human-specified space of high-level commands, they must be initialized with human supervision, typically obtained by pretraining the executor. On its own, this training paradigm restricts the quality of the learned executor policy to that exhibited in (possibly suboptimal) human supervision. For tasks like the real-time strategy game depicted in Fig. 1, we would like to study LLPs trained via **reinforcement learning (RL)**, jointly learning from a downstream reward signal, and optimizing both instructors and executors for task success rather than fidelity to human teachers.

Training LLPs via RL has proven difficult. Past work has identified two main challenges: primarily, the LLP-specific problem of **semantic drift**, in which agents come to deploy messages in ways

inconsistent with their original (natural language) meanings (Lewis et al., 2017; Lee et al., 2019); secondarily, the general problem of **sample inefficiency** in RL algorithms (Kakade et al., 2003; Brunskill and Li, 2013). Model-free deep RL is particularly notorious for requiring enormous amounts of interaction with the environment (Munos et al., 2016; Mnih et al., 2013b). For LLPs to meet their promise as flexible, controllable, and understandable tools for deep learning, better approaches are needed to limit semantic drift and perhaps improve sample efficiency.

While semantic change is a constant and well-documented feature of human languages (McMahon and April, 1994), (human) word meanings are on the whole remarkably stable relative to the rate of change in the tasks for which words are deployed (Karjus et al., 2020). In particular, disappearance of lexical items is mitigated by increased population size (Bromham et al., 2015) and increased frequency of use (Pagel et al., 2007). Drawing on these facts about stabilizing factors in human language, we hypothesize that training of machine learning models with latent language variables can be made more robust by incorporating a population of instructors with diverse communicative needs that exercise different parts of the lexicon.

We describe a multitask LLP training scheme in which task-specific instructors communicate with a shared executor. We show that complex long-horizon LLPs can be effectively tuned via *joint reinforcement learning* of instructors and executors using *multitask training*:

- Section 3 presents a formal analysis of LLP training as an iterated Lewis signalling game (Lewis, 1969). By modeling learning in this game as a dynamical system, we completely characterize a class of simple policies that are subject to semantic drift. We show that a particular multitask training scheme eliminates the set of initializations that undergo semantic drift.
- Section 4 evaluates the empirical effectiveness of multitask learning in a real-time strategy game featuring rich language, complex complex dynamics, and LLPs implemented with deep neural networks. Again, we show that multitask training reduces semantic drift (and improves sample efficiency) of LLPs in multiple game variants.

Together, these results show that diverse shared

goals and communicative needs can facilitate (and specifically stabilize) learning of communication strategies.

## 2 Background and Related Work

Deep reinforcement learning (DRL) has recently made impressive progress on many challenging domains such as games (Mnih et al., 2013a; Silver et al., 2016), locomotion (Schulman et al., 2015) and dexterous manipulation tasks (Gu et al., 2016; Rajeswaran et al., 2017). However, even state-of-the-art approaches to reinforcement struggle with tasks involving complex goals, sparse rewards, and long time horizons. A variety of models and algorithms for *hierarchical* reinforcement learning have been proposed to address this challenge (Dayan and Hinton, 1993; Dietterich, 2000; Richard et al., 1999; Bacon et al., 2017) via supervised or unsupervised training of a fixed, discrete set of sub-policies.

Language can express arbitrary goals, and has compositional structure that allows generalization across commands. Building on this intuition, several recent papers have explored hierarchical RL in which natural language is used to parameterize the space of high-level actions (Oh et al., 2017; Andreas et al., 2017; Shu et al., 2018; Jiang et al., 2019; Hu et al., 2019). While there are minor implementation differences between all these approaches, we will refer to them collectively as **latent language policies (LLPs)**. Like other hierarchical agents, an LLP consists of a pair of subpolicies: an instructor  $I(m \mid o)$  and an executor  $E(a \mid m, o)$ . An LLP takes actions by first sampling a string-valued message  $m \sim I$  from the instructor, and then an action  $a \sim E$  from the executor. For these messages to correspond to *natural language*, rather than arbitrary strings, policies need some source of information about what human language users think they mean. This is typically accomplished by pretraining executors via human demonstrations or reinforcement learning; here we focus on the ingredients of effective joint RL of instructors and executors.

Reinforcement learning has been widely used to improve supervised language generation policies, particularly for dialogue (Li et al., 2016; Lewis et al., 2017), translation (Ranzato et al., 2015; Wu et al., 2016) and summarization (Stiennon et al., 2020). Here, we instead focus on models where language is a latent variable as part of a hierarchical policy for a non-linguistic task.

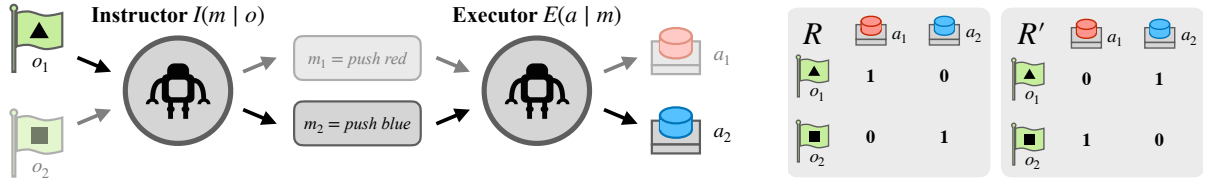


Figure 2: A signaling game with two possible observations, two possible messages, and two possible actions. The instructor observes either a triangle or a square, then sends a message to a executor, who pushes either the red or blue buttons. The players’ reward depends on the observation and the action but not the message. Two possible reward functions,  $R$  and  $R'$ , are shown at right.

As noted in Section 1, an observed shortcoming of reinforcement learning in all these settings is its susceptibility to **semantic drift**. In the literature on human language change (Blank, 1999), semantic drift refers to a variety of phenomena, including specific terms becoming more general, general terms becoming specific, and parts coming to refer to wholes. In machine learning, it refers broadly to the use of messages inconsistent with their natural language meanings in language-generation policies (Lazaridou et al., 2020).

Lee et al. (2019) mitigate semantic drift in pivot-based machine translation by using visual grounding, whereas Lu et al. (2020) periodically update a student model on data generated by an RL teacher. Work in emergent communication has found that reinforcement learning tends not to learn policies with natural language-like properties (Kottur et al., 2017), although population-based training has been found to be helpful (Gupta et al., 2019). Most relatedly to our work, Lazaridou et al. (2020) train speaker-listener agents jointly in a visual referential communication task and introduce auxiliary loss functions for stabilizing training. Our work focuses on a more general setting where the interactions are temporally extended, have large action spaces and is partially observable. Agarwal et al. (2019) use populations of agents to reduce semantic drift in visual dialogue. We view the current paper’s analysis of multitask learning as complementary to these approaches from the emergent communication literature; future work might consider ways of combining the two.

A great deal of recent work in both RL (e.g. Jaderberg et al., 2016; Shelhamer et al., 2016) and language processing (e.g. Clark et al., 2019; Gururangan et al., 2020) has observed that carefully designed training objectives can serve as a source of model-agnostic inductive bias. Our results bring these two lines of work together: multitask train-

ing improves the faithfulness and adaptability of learned language understanding models, even when optimizing for a downstream reward.

### 3 Multitask Communication in Theory: Lewis Signaling Games

We begin our analysis with the simple **signaling game** depicted in Fig. 2. In this game, one agent receives an observation, then sends a message to another agent, which then performs an action. Signaling games like this one are widely studied in NLP as models of reference resolution and language generation (Frank and Goodman, 2012). The instructor-executor pair may together be viewed as the simplest LLP of the kind described in Section 2.

Formally, a (2-observation, 2-message) **Lewis signalling game** is defined by:

- a set of **observations**  $O = \{o_1, o_2\}$
- a set of **messages**  $M = \{m_1, m_2\}$
- a set of **actions**  $A = \{a_1, a_2\}$
- a **reward function**  $R : O \times A \rightarrow \mathbb{R}$

The game is played between two agents: a **instructor** (with parameters  $\theta_I$ ), which receives an observation and samples an observation-specific message from a distribution  $I(m | o; \theta_I)$ ; and a **executor** (with parameters  $\theta_E$ ), which receives the instructor’s message and uses it to sample an action from a distribution  $E(a | m; \theta_E)$ . The agents then receive a reward  $R(o, a)$  that depends on the observation and action but not on the message sent. This policy’s expected reward is given by:

$$\sum_{\substack{o \in O \\ m \in M \\ a \in A}} p(o) I(m | o; \theta_I) E(a | m; \theta_E) R(o, a) . \quad (1)$$

Gradient ascent on Eq. (1) with respect to  $\theta_I$  and  $\theta_E$  (e.g. using a policy gradient algorithm; Williams, 1992) can be used to improve the expected reward obtained by an LLP.

As an example, the right portion of Fig. 2 shows two reward functions  $R$  and  $R'$ . In both, each observation is paired with a single action, and the executor must take the action corresponding to the observation to receive a positive reward. For  $R$ , two strategies obtain the optimal expected reward of 1: one in which  $I(m_1 | \blacktriangle) = 1$  and  $E(\text{red} | m_1) = 1$ , and one in which  $I(m_1 | \blacksquare) = 1$  and  $E(\text{blue} | m_1) = 1$ . Almost every initialization of  $\theta_E$  and  $\theta_I$  (excluding a set of *pooling equilibria*; see e.g. [Huttenberger et al., 2010](#)) converges to one of these two strategies when agents are jointly trained to optimize Eq. (1).

**Semantic drift** Suppose, as shown in Fig. 2, the messages  $m_1$  and  $m_2$  are not arbitrary symbols, but correspond to the natural language expressions  $m_1 = \text{push red}$  and  $m_2 = \text{push blue}$ . In this case, only of the policies described above corresponds to the semantics of natural language—namely, the one in which  $E(a_1 | m_1) = E(\text{red} | \text{push red}) = 1$ . What is needed to ensure that a pair of agents playing converge to the natural language strategy?

In the analysis that follows, we will consider instructor and executor agents (each with a single scalar parameter  $\in [0, 1]$ ):

$$I(m_j | o_i; \theta_I) = \begin{cases} \theta_I & i = j \\ 1 - \theta_I & \text{otherwise} \end{cases} \quad (2)$$

$$E(a_j | m_i) = \begin{cases} \theta_E & i = j \\ 1 - \theta_E & \text{otherwise} \end{cases} \quad (3)$$

For the game depicted in Fig. 2, we would like to avoid any outcome in which, after training,  $\theta_E = E(\text{red} | \text{push red}) < \frac{1}{2}$ . More generally, let us assume that we have an initial set of executor parameters that are possibly suboptimal but correspond to natural language semantics in the sense that  $E(a | m_i; \theta_E^{(0)}) > \frac{1}{2}$  if and only if the meaning of  $m$  is *do a*. In this case, we will say that a parameter initialization  $(\theta_I^{(0)}, \theta_E^{(0)})$  undergoes **executor semantic drift** if, after training, any such  $E(a | m_i; \theta_E) = \theta_E < \frac{1}{2}$ .

To analyze semantic drift in this game, we consider the final values of the parameters  $(\theta_I, \theta_E)$  when optimized from an initialization  $(\theta_I^{(0)}, \theta_E^{(0)})$ . For the reward function  $R$  depicted in Fig. 2, we can perform gradient ascent on Eq. (1) with respect to  $\theta_I$  and  $\theta_E$  in this model by observing that:

$$\frac{\partial J}{\partial \theta_I} = \theta_E - \frac{1}{2} \quad \frac{\partial J}{\partial \theta_E} = \theta_I - \frac{1}{2} \quad (4)$$

By considering the limiting behavior of gradient ascent as step size goes to zero (a **gradient flow**; see Appendix A), it is possible to give a closed-form expression for the value of these parameters as a function of time:

**Proposition 1.** *Suppose  $\theta_E^{(0)} + \theta_I^{(0)} < 1$ . Then two agents optimizing Eq. (1) via Eq. (4) undergo semantic drift (converging to  $\theta_E = 0$ ).*

Proof is given in Appendix A. Note in particular that semantic drift will occur whenever  $\theta_I^{(0)} < 1 - \theta_E^{(0)}$ , which can occur even assuming a well-initialized executor with  $\theta_E > \frac{1}{2}$ . Fig. 5 in the appendix provides a visualization of learning dynamics and these drift-susceptible initializations. However, we will next show that this drift can be eliminated via multitask training.

**Multitask signaling games** Consider a **multitask** version of this game with the **two** reward functions  $R$  and  $R'$  depicted in Fig. 2. As discussed in the introduction and depicted in Fig. 1, our approach to multitask training focuses on sharing a single executor  $E(a | m; \theta_E)$  between multiple task-specific instructors, here  $I(m | o; \theta_{I1})$  and  $I(m | o; \theta_{I2})$ , both parameterized as in Eq. (2). As above, we train  $(\theta_{I1}, \theta_{I2}, \theta_E)$  jointly to optimize:

$$\sum_{o,m,a} p(o) I(m|o; \theta_{I1}) E(a|m; \theta_E) R(o, a) + \sum_{o,m,a} p(o) I(m|o; \theta_{I2}) E(a|m; \theta_E) R'(o, a) \quad (5)$$

We assume that the two instructors share the same initialization, with  $\theta_{I1}^{(0)} = \theta_{I2}^{(0)} = \theta_I^{(0)}$ . In this case, the following is true:

**Proposition 2.** *Suppose  $\theta_E^{(0)} > \frac{1}{2}$  and  $\theta_{I1}^{(0)} = \theta_{I2}^{(0)}$ . Then three agents optimizing Eq. (5) via its gradient flow do not undergo semantic drift. In fact, the eventual executor parameter  $\theta_E^{(t)}$  is independent of the initial speaker parameters  $\theta_{I1}^{(0)}$  and  $\theta_{I2}^{(0)}$ .*

Proof is again given in Appendix A. It is important to emphasize that these results concern the simplest possible policies for the signaling games considered here: agents with a single parameter which already “bake in” the assumption that different signals should trigger different behaviors. We leave generalization of this formal analysis to general signaling games with more complex agents and message spaces for future work, noting that—at least in this simple case—we have succeeded in constructing a concrete multitask objective that



reduces (indeed eliminates) the set of initial model parameters subject to semantic drift.

## 4 Multitask Communication in Practice: The MiniRTS Environment

We next verify whether this result extends to the complex LLP-learning tasks discussed in Section 2. Our focus in this section is the MINIRTS environment of Hu et al. (2019) (depicted in Fig. 1), in which agents must build and control an army of units like archers, spearmen, swordman, cavalry, and dragons, each with specialized abilities, with the goal of destroying the opponent’s town center. Using this game, Hu et al. (2019) crowdsourced a dataset of high-level instructions (like *attack with dragon* and *send idle peasant to mine*) paired with low-level action sequences (Fig. 1). They showed that an LLP trained on this supervised data via behavior cloning significantly outperformed a flat policy trained with imitation learning directly on low-level action sequences.

Here we investigate (1) whether these policy-cloned LLP can be further improved via reinforcement learning directly on a sparse win-loss signal from the game, (2) whether we can improve sample efficiency during reinforcement learning by jointly training executor models on multiple game variants simultaneously through multitask learning, and (3) whether semantic drift can be avoided during multi-task training. Below, Section 4.1, Section 4.2 and Section 4.3 provide more detail about the task, model, and training procedure. Section 4.4 reports experimental results.

### 4.1 Task and Training Data

MINIRTS is a partially-observable real-time strategy game environment, in which the actions of a large number of units must be coordinated on long time scales to defeat an opposing player. In a typical episode, a player must use its initial units to gather resources, use resources to build specialized structures for producing other units, and finally deploy these units to attack the opposing player’s base. This involves challenging problems in both low-level tactics (controlling the placement of individual units for resource-gathering and combat) and high-level strategy (deciding which unit types to build, and when to deploy them).

MINIRTS additionally features a dataset collected from pairs of humans playing collaboratively against rule-based opponents. One human, the *in-*

*structor*, designs high-level strategies and describes them in natural language. The other human, the *executor* observes the environment state as well as the natural language strategy descriptions from the instructor and selects appropriate low-level actions. The dataset consists of 5,392 games, with a total of 76,045 (instruction, execution) pairs.

### 4.2 Model

Hu et al. (2019) use the labeled data to train an LLP for the MINIRTS environment. Our experiments use the same model architecture (Fig. 3), which we briefly review here; see the original for details.

**Observation encoder** The *instructor* and *executor* models condition on a fixed-sized representation of the current game state which are constructed using different encoders for various aspects of the game state (Fig. 3):

- *Spatial input encoder*: The spatial information of the map is encoded using a convolutional neural network.
- *Non-spatial input encoder*: The non-spatial attributes and internal state of game objects are encoded using a simple MLP. These include attributes like the number of enemy units, the agent’s units, and resource locations.
- *Instruction encoder*: The current instruction is encoded with a recurrent neural network.
- *Auxiliary encoder*: Global variables, such as the total number of resources collected, are additionally encoded with an MLP.

**Instructor model** The *instructor* takes in the game state from the observation encoder and produces instructions. The 500 instructions appearing most frequently in the training set are encoded with an RNN into a fixed-sized vector. The score for each instruction is proportional to its dot product with the game state encoding. This instructor model achieved the best performance on several metrics in the original work (Hu et al., 2019). By restricting the instructor to the most frequent 500 well-formed natural language strings, we are able to focus our attention on **semantic** drift. A generative model free to generate arbitrary strings might also be subject to **syntactic** drift.

**Executor model** The *executor* predicts an action for every unit controlled by the agent based on of the current observation as encoded by the various encoders. The executor then predicts an action

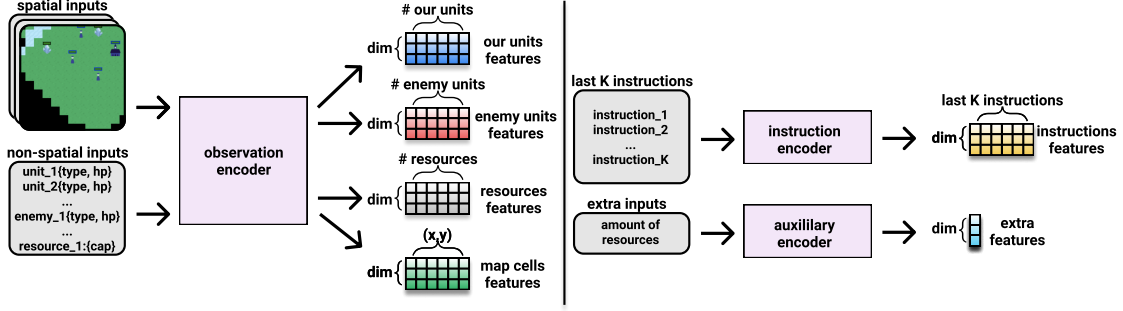


Figure 3: State representations for the MINIRTS environment. The model encodes the spatial observations, non-spatial observations, instructions, and auxiliary information about environment state at each timestep. These features are used by both the *instructor* and *executor* models. Reproduced with permission from Hu et al. (2019).

based on these features. In particular, for each unit, it predicts one of the 7 action types (IDLE, CONTINUE, GATHER, ATTACK, TRAIN UNIT, BUILD BUILDING, MOVE), an action target location (for the MOVE, ATTACK, GATHER and BUILD BUILDING actions) and a unit type (for the TRAIN UNIT and BUILD BUILDING actions). Taking the product of the factorized action and location arguments across all units, the action space for the executor can be enormous, with as many as  $10^{24}$  distinct actions available on a single turn.

### 4.3 Training

As mentioned above, the original work of Hu et al. (2019) (and other work on learning LLPs) focused on behavior cloning or independent supervision of the *instructor* and *executor*. In the current paper, we are interested in the the dynamics of joint reinforcement learning of LLPs in both single- and multitask settings. Experiments in Section 4.4 make use of models trained with all three strategies.

**Rule-based opponent pool** Agents are trained against a similar pool of rule-based bots (see Hu et al., 2019) used to collect the human data. These bots follows a randomly selected, unit-specific strategy, building a fixed number of SWORDMEN, SPEARMEN, CAVALRY, ARCHERS or DRAGONS and attacking as soon as they are constructed.

**Behavior cloning** Behavior-cloned models are trained using the supervised MINIRTS dataset. Given a collection of game observations  $o$ , each annotated with a high-level action  $m$  and a low-level action  $a$ , we maximize:

$$\max_{\theta_I, \theta_E} \sum_{o, m, a} [\log I(m \mid o; \theta_I) + \log E(a \mid m, o; \theta_E)] . \quad (6)$$

During training, one frame is taken from every  $K$  frames to form the supervised learning dataset. To preserve unit level actions for the *executor* training, all actions that happen in  $[tK, (t+1)K)$  frames are stacked onto the  $tK$ th frame.

**Reinforcement learning** To train agents via reinforcement learning, we initialize them with the behavior cloning objective in Eq. (6) to provide an initial, human-meaningful grounding of message semantics (analogous to the initialization of the executor parameter  $\theta_E^{(0)}$  in Section 3). We then fine-tune them on game success, providing a sparse reward of 1 when agents win the game, -1 when they lose or draw the game.

As in Section 3, learned agents are trained on the game reward, using a **proximal policy optimization (PPO)** (Schulman et al., 2017) objective to optimize the expected reward:

$$\mathbb{E}_{(s,a) \sim (I,E)} R(s, a) . \quad (7)$$

**Multi-task RL** The original game of Hu et al. (2019) is defined by a set of **attack multipliers**: the aforementioned rock-paper-scissors dynamic arises because spearmen are especially effective against cavalry, archers against dragons, etc. To create alternative “tasks” in the MiniRTS environment, we create alternative versions of the game featuring different multipliers: e.g. making dragons invulnerable to archers or cavalry extra-effective against swordsmen. Table 1 shows these multipliers for the original rule, and a set of game **variants** with different multipliers are described in Appendix D. These variants are labeled B–J in the experiments that follow. Multiplier changes have significant effects on the optimal high-level strategy, affecting both which units are most effective overall, and how players should respond to opponents’ choices.

As in Section 3, we perform multitask LLP training in the MINIRTS environment by jointly optimizing expected reward across multiple game variants at once, assigning each variant its own set of *instructor* parameters  $\theta_I$  (initialized to the same value) but sharing a single set of *executor* parameters  $\theta_E$  across all contexts. The training pseudo-code can be found in Appendix C.

## 4.4 Experiments

Unlike in the signaling game considered in Section 3, MINIRTS is complex, and we cannot take for granted that reinforcement learning of LLPs (with either ordinary or multitask objectives) will converge to an improved good solution at all. We thus begin with an analysis of policy performance and sample efficiency, then conclude this section with an analysis of semantic drift. (Model training details can be found in Appendix B.)

### 4.4.1 Evaluating performance and sample efficiency

To evaluate policy quality and sample complexity, we compare the final win rate (against the fixed pool of rule-based agents) for the policy-cloned (BC), RL-tuned (RL<sup>joint</sup>), and multitask-RL-tuned (RL<sup>multi</sup>) agents described above. We perform this evaluation for multiple game configurations: **original**, with the same rules used by Hu et al. (2019) for human data collection and evaluation, and 3 alternative variants (**variant G**, **variant H**, **variant J**), in which the relative strengths of various units has been modified (see Appendix D). We train 4 separate (RL<sup>joint</sup>) agents corresponding to each of the environments and 2 (RL<sup>multi</sup>) agents. Following D’Eramo et al. (2019), we provide both training strategies with a fixed budget of training experience *across environments*: both RL<sup>joint</sup> and RL<sup>multi</sup> have been trained on the same number of game episodes per training environment. We also present the win rates for RL<sup>joint</sup> and RL<sup>multi</sup>, when trained on  $3\times$

Unit name	Attack multiplier				
	Swordman	Spearman	Cavalry	Archer	Dragon
SWORDMAN	1.0	1.5	0.5	1.0	0.0
SPEARMAN	0.5	1.0	1.5	1.0	0.0
CAVALRY	1.5	0.5	1.0	1.0	0.0
ARCHER	0.5	0.5	0.5	0.5	2.0
DRAGON	1.0	1.0	1.0	0.5	1.0

Table 1: Attack multipliers for the **original** game rules. For example, cavalry are extra-effective against swordsmen (1.5 in Swordsman col.); only archers and dragons can attack dragons (nonzero entries in Dragon col.). See Appendix D for other game variants’ multipliers.

more episodes per environment.

Results are shown in Table 2. Both RL fine-tuning strategies allow the policy to significantly improve over the behavior-cloned initializer, showing that effective reinforcement learning of LLPs is possible in MINIRTS. In most environments, performance of the model fine-tuned with the multitask training is higher than ordinary joint RL training. When RL<sup>joint</sup> is provided extra training budget, it sometimes surpasses the performance of RL<sup>multi</sup> model with standard number of episodes. However, when RL<sup>multi</sup> is also given the extra training budget, it performs better in all but one environment. At a high level, these results indicate that multitask training of LLPs can be applied at small (and in some cases no) cost in accuracy and significantly less per-environment training cost.

### 4.4.2 Evaluating semantic drift

Next, we consider the second question from the introduction: outside of performance effects, does multitask training with populations of instructors reduce semantic drift in executors? We present two

Training strategy	Evaluation environment	Win rate	
		(standard)	( $3\times$ training)
BC	original	30.3	-
RL <sup>joint</sup> [orig.]		65.7	86.9
RL <sup>multi</sup> [orig., B, C]		<b>76.5</b>	90.6
BC	variant G	11.6	-
RL <sup>joint</sup> [G]		73.0	74.1
RL <sup>multi</sup> [G, H, J]		<b>75.7</b>	<b>77.6</b>
BC	variant H	26.2	-
RL <sup>joint</sup> [H]		82.2	<b>91.4</b>
RL <sup>multi</sup> [G, H, J]		79.4	83.5
BC	variant J	14.6	-
RL <sup>joint</sup> [J]		87.2	93.0
RL <sup>multi</sup> [G, H, J]		<b>91.2</b>	93.7

Table 2: Evaluation of policy quality in MINIRTS. Policies are evaluated against a rule-based opponent pool in four environments: MINIRTS with original rules, and three rule variants described in Appendix D. We compare the original behavior-cloned LLP of Hu et al. (2019) (BC) with one fine-tuned directly on the evaluation environment (RL<sup>joint</sup>[env]) and one with multitask tuning on the evaluation environment and two others (RL<sup>multi</sup>[env1, env2, env3]). Both RL fine-tuning strategies significantly outperform their behavior-cloned initializer. When using the same number of game episodes per training environment, RL<sup>multi</sup> is generally best but when RL<sup>joint</sup> is provided additional budget, it sometimes beats RL<sup>multi</sup>. Differences between models in environments original, G and J are significant ( $p < 0.05$  under a permutation test).

different quantitative evaluations that provide different perspectives on the answer to this question.

**Semantic drift** In MINIRTS, executor semantic drift occurs when the executor performs actions that are not consistent with the instruction produced by the instructor. (i.e., *create spearman* instruction produced by the instructor leads to the executor producing *swordman* instead). In particular, this occurs in  $RL^{joint}$  because the instructor and executor can co-adapt to new semantics during exploration as they are only trained to win the game.

**Agent interoperability** First, we evaluate the robustness of executor policies to alternative choices of instructors. Specifically, we pair each RL-trained executor with an instructor trained either via behavior cloning (and thus guaranteed to implement the human annotators’ semantics) or fine-tuning ( $RL^{instr}$ ) on a different game variant from the executor (and thus not co-adapted with it). Intuitively, to succeed at this task, executors must follow messages produced by the instructors trained in that domain. Executors that have undergone less semantic drift should perform better when paired with these different instructors. Results are shown in Table 3; here, it can be seen that multitask learning matches or exceeds the performance of single-task training on this evaluation of semantic drift in all rule variants studied, even when  $RL^{joint}$  is provided additional training budget. As evidence that performance comes from instructor–executor pairs,

Instructor	Executor	Eval. env.	Win rate	
			(standard)	( $3 \times$ training)
BC	$RL^{joint}[orig.]$	original	48.4	59.2
	$RL^{multi}[orig., B, C]$		<b>60.7</b>	<b>67.2</b>
$RL^{instr}[D]$	$RL^{joint}[orig.]$	variant D	74.8	85.9
	$RL^{multi}[orig., B, C]$		<b>88.3</b>	89.1
$RL^{instr}[E]$	$RL^{joint}[orig.]$	variant E	57.5	68.7
	$RL^{multi}[orig., B, C]$		<b>72.9</b>	<b>76.6</b>
$RL^{instr}[F]$	$RL^{joint}[orig.]$	variant F	73.2	83.8
	$RL^{multi}[orig., B, C]$		<b>87.3</b>	<b>92.3</b>

Table 3: Evaluation of semantic drift in MINIRTS. Here, reinforcement-learned executor models are paired with instructors different from those they are trained with: either the original behavior-cloned instructor, or a instructor fine-tuned in an entirely different environment. The multitask executor  $RL^{multi}$  performs better than  $RL^{joint}$  when paired with new instructors, even when  $RL^{joint}$  is given additional training budget. Differences in all environments are significant ( $p < 0.05$  under a permutation test).

rather than executors alone, using a random coach paired with  $RL^{multi}[orig., B, C]$  on variant D gives 33.2% accuracy. Additionally, when  $RL^{multi}[orig., B, C]$  is paired with a coach from a different variant, we get an accuracy of just 41% on variant D.

**Low-level action semantics** As an alternative means of gaining insight into learned behaviors, we can directly inspect the correspondence between instructor messages and executor actions. We do this by uniformly sampling messages from a random instructor, then feeding them to the  $RL^{multi}$  and  $RL^{joint}$  executors and observing their choice of low-level actions  $a$ . We then restrict these  $(m, a)$  pairs to those in which (1) the text of  $m$  includes one of the words *create*, *build*, *train* or *make* and the name of a unit (*peasant*, *spearman*, etc.) and (2)  $a$  is a TRAIN UNIT action for *any* unit. We then compute the empirical probability  $P(\text{unit}_1 \in a \mid \text{unit}_2 \in m)$  as shown in Fig. 4. If there is semantic drift, we expect to observe non-zero probability on the off-diagonal entries (the executor is building units different from those it is instructed to build).  $RL^{multi}$  places less probability mass on the off-diagonal entries compared to  $RL^{joint}$ , consistent with less semantic drift. In Fig. 4, one can also note that some word meanings change more than others. We hypothesize that, this is because like in natural languages, environmental pressures cause the meanings of some words to change at a greater rate than others. In this case, the dynamics of the game makes the *spearman* unit slightly stronger than the *swordman* unit overall. This results in unexpectedly good performance for players who accidentally misinterpret *swordman* as *spearman*. Therefore, this creates pressure for the conventional meaning of *swordman* to shift more than other units.

Taken together, these two evaluation results show that, when fine-tuning a policy initialized via imitation learning on the same objective, ordinary RL can be quite effective: the resulting executor model performs well even when paired with other instructors. But as in Section 3, multitask training is even more helpful, especially by reducing semantic drift in both familiar and new environments.

## 5 Conclusions

We have presented a theoretical and empirical analysis of semantic drift and sample efficiency in multitask reinforcement learning of latent language policies (LLPs). In a Lewis signaling game, we



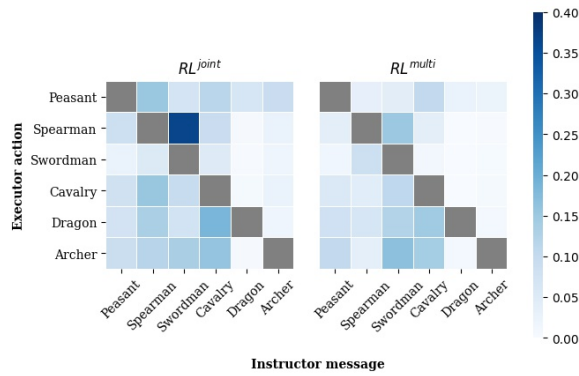


Figure 4: Message-action drift in MINIRTS. Unit  $X$  on the x-axis indicates that a message  $m$  of the form *build unit  $X$*  was sampled. Unit  $Y$  on the y-axis shows the low-level TRAIN UNIT  $Y$  action  $a$  sampled by the executor. Matrix entries show the empirical probability  $P(a | m)$  computed using the relative frequencies of the sampled instructor messages and the corresponding executor actions. The total sum of off-diagonal entries is 2.98 for  $RL^{joint}[orig.]$  and 1.98 for  $RL^{multi}$ , indicating less semantic drift for  $RL^{multi}[orig., B, C]$ .

proved that multitask training can completely eliminate semantic drift. In a two-player real-time strategy game, we showed that multitask training is effective at mitigating semantic drift, improves the quality of learned policies and is sample efficient. Future work might integrate these results with other forms of population-based training (like those proposed by Gupta et al. (2019) for reference games) and explore other environmental factors affecting dynamics of language change in populations of learned agents.

## Acknowledgments

We thank Hengyuan Hu for assistance in reproducing the original work and MIT Supercloud for compute resources. We also thank Eric Chu, Alana Marzoev, Ekin Akyürek and Debby Clymer for feedback and proof-reading.

## References

- Akshat Agarwal, Swaminathan Gurumurthy, Vasu Sharma, Mike Lewis, and Katia Sycara. 2019. Community regularization of visually-grounded dialog. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1042–1050.
- Jacob Andreas, Dan Klein, and Sergey Levine. 2017. Modular multitask reinforcement learning with policy sketches. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 166–175.
- Jacob Andreas, Dan Klein, and Sergey Levine. 2018. Learning with latent language.
- Pierre-Luc Bacon, Jean Harb, and Doina Precup. 2017. The option-critic architecture. In *Thirty-First AAAI Conference on Artificial Intelligence*.
- Andreas Blank. 1999. Why do new meanings occur? A cognitive typology of the motivations for lexical semantic change. *Historical semantics and cognition*, 13(6).
- Lindell Bromham, Xia Hua, Thomas G Fitzpatrick, and Simon J Greenhill. 2015. Rate of language evolution is affected by population size. *Proceedings of the National Academy of Sciences*, 112(7):2097–2102.
- Emma Brunskill and Lihong Li. 2013. Sample complexity of multi-task reinforcement learning. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 122–131.
- Kevin Clark, Minh-Thang Luong, Urvashi Khandelwal, Christopher D Manning, and Quoc Le. 2019. Bam! Born-again multi-task networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5931–5937.
- Peter Dayan and Geoffrey E. Hinton. 1993. Feudal reinforcement learning. *Advances in Neural Information Processing Systems*.
- Carlo D’Eramo, Davide Tateo, Andrea Bonarini, Marcello Restelli, and Jan Peters. 2019. Sharing knowledge in multi-task deep reinforcement learning. In *International Conference on Learning Representations*.
- Thomas G Dietterich. 2000. Hierarchical reinforcement learning with the maxq value function decomposition. *Journal of artificial intelligence research*, 13:227–303.
- Michael C Frank and Noah D Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.
- Edward Gibson, Richard Futrell, Julian Jara-Ettinger, Kyle Mahowald, Leon Bergen, Sivalogeswaran Ratnasingam, Mitchell Gibson, Steven T Piantadosi, and Bevil R Conway. 2017. Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences*, 114(40):10785–10790.
- Shixiang Gu, Ethan Holly, Timothy P. Lillicrap, and Sergey Levine. 2016. Deep reinforcement learning for robotic manipulation. *CoRR*, abs/1610.00633.
- Abhinav Gupta, Ryan Lowe, Jakob Foerster, Douwe Kiela, and Joelle Pineau. 2019. Seeded self-play for language learning. In *Proceedings of the Beyond Vision and LANGUAGE: inTEgrating Real-world kNOWLEDge (LANTERN)*, pages 62–66.

- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don’t stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.
- Hengyuan Hu, Denis Yarats, Qucheng Gong, Yuan-dong Tian, and Mike Lewis. 2019. Hierarchical decision making by generating and following natural language instructions. In *Advances in neural information processing systems*, pages 10025–10034.
- Simon M Huttegger, Brian Skyrms, Rory Smead, and Kevin JS Zollman. 2010. Evolutionary dynamics of Lewis signaling games: signaling systems vs. partial pooling. *Synthese*, 172(1):177.
- Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki, Tom Schaul, Joel Z Leibo, David Silver, and Koray Kavukcuoglu. 2016. Reinforcement learning with unsupervised auxiliary tasks. *arXiv preprint arXiv:1611.05397*.
- Yiding Jiang, Shixiang Shane Gu, Kevin P Murphy, and Chelsea Finn. 2019. Language as an abstraction for hierarchical deep reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 9419–9431.
- Sham Machandranath Kakade et al. 2003. *On the sample complexity of reinforcement learning*. Ph.D. thesis, University of London London, England.
- Andres Karjus, Richard A Blythe, Simon Kirby, and Kenny Smith. 2020. Communicative need modulates competition in language change. *arXiv preprint arXiv:2006.09277*.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Satwik Kottur, José Moura, Stefan Lee, and Dhruv Batra. 2017. Natural language does not emerge ‘naturally’ in multi-agent dialog. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2962–2967.
- Angeliki Lazaridou, Anna Potapenko, and Olivier Tieleman. 2020. Multi-agent communication meets natural language: Synergies between functional and structural language learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7663–7674.
- Jason Lee, Kyunghyun Cho, and Douwe Kiela. 2019. Countering language drift via visual grounding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4376–4386.
- David Lewis. 1969. *Convention: A philosophical study*. John Wiley & Sons.
- Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? End-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2443–2453.
- Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.
- Yuchen Lu, Soumye Singhal, Florian Strub, Aaron Courville, and Olivier Pietquin. 2020. Countering language drift with seeded iterated learning. In *International Conference on Machine Learning*, pages 6437–6447. PMLR.
- April MS McMahon and McMahon April. 1994. *Understanding language change*. Cambridge University Press.
- V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. 2013a. Playing Atari with deep reinforcement learning. *CoRR*.
- Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013b. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.
- Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc Bellemare. 2016. Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 1054–1062.
- Junhyuk Oh, Satinder Singh, Honglak Lee, and Pushmeet Kohli. 2017. Zero-shot task generalization with multi-task deep reinforcement learning. *ICML*.
- Mark Pagel, Quentin D Atkinson, and Andrew Meade. 2007. Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature*, 449(7163):717–720.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037.
- Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, John Schulman, Emanuel Todorov, and Sergey Levine. 2017. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. *CoRR*, abs/1709.10087.

- Marc’Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*.
- Sutton Richard, Precup Doina, and Singh Satinder. 1999. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artif. Intell.*, 112(1-2).
- John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. 2015. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Damien Scieur, Vincent Roulet, Francis Bach, and Alexandre d’Aspremont. 2017. Integration methods and optimization algorithms. In *Advances in Neural Information Processing Systems*, pages 1109–1118.
- Evan Shelhamer, Parsa Mahmoudieh, Max Argus, and Trevor Darrell. 2016. Loss is its own reward: Self-supervision for reinforcement learning. *arXiv preprint arXiv:1612.07307*.
- Tianmin Shu, Caiming Xiong, and Richard Socher. 2018. Hierarchical and interpretable skill acquisition in multi-task reinforcement learning. In *International Conference on Learning Representations*.
- D. Silver, A. Huang, C. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, and M. Lanctot. 2016. Mastering the game of go with deep neural networks and tree search. *Nature* 529.
- Elizabeth S Spelke. 2017. Core knowledge, language, and number. *Language Learning and Development*, 13(2):147–170.
- Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33.
- Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.
- Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

## A Lewis signaling games: details

This appendix provides details of the formal analysis of the signaling game discussed in Section 3.

**Single-task learning: Proof of Proposition 1** In the single-task case, we wish to train the policy given in Eq. (2) (with parameters  $(\theta_I, \theta_E)$ ). For the reward function  $R$  in Fig. 2, this policy has expected reward:

$$\begin{aligned} J(\theta_E, \theta_I) &= \sum_{\substack{o \in O \\ m \in M \\ a \in A}} p(o) I(m \mid o; \theta_I) E(a \mid m; \theta_E) R(o, a) . \\ &= \frac{1}{2} \theta_I \theta_E + \frac{1}{2} (1 - \theta_I) (1 - \theta_E) \end{aligned} \quad (8)$$

As noted in Eq. (4), the gradient of this expected reward with respect to agent parameters is

$$\frac{\partial J}{\partial \theta_I} = \theta_E - \frac{1}{2} \quad (9)$$

$$\frac{\partial J}{\partial \theta_E} = \theta_I - \frac{1}{2} \quad (10)$$

Performing gradient ascent will thus give a series of parameters updates with

$$\theta_I^{(t)} = \theta_I^{(t-1)} + \alpha^{(t)} \left( \theta_E^{(t-1)} - \frac{1}{2} \right) \quad (11)$$

$$\theta_E^{(t)} = \theta_E^{(t-1)} + \alpha^{(t)} \left( \theta_I^{(t-1)} - \frac{1}{2} \right) \quad (12)$$

The exact sequence of iterates will depend on the choice of optimization algorithm, step size, and other hyperparameters. In order to provide the most general characterization of learning in this signaling game, we consider optimization of  $\theta_I$  and  $\theta_E$  in *continuous time*. (This can be viewed as the limiting case of ordinary SGD as the step size goes to zero; for more discussion of relationships between gradient descent and continuous gradient flows see [Scieur et al., 2017](#).) Taking  $\theta_I^{(t)}$  and  $\theta_E^{(t)}$  to now be functions of a real-valued variable  $t$ , optimization corresponds to the system of ordinary differential equations:

$$\frac{d\theta_I^{(t)}}{dt} = \theta_E - \frac{1}{2} \quad (13)$$

$$\frac{d\theta_E^{(t)}}{dt} = \theta_I - \frac{1}{2} \quad (14)$$

It can be verified that solutions to this system of equations have the following general form:

$$\theta_I^{(t)} = (c_1 + c_2)e^t + (c_1 - c_2)e^{-t} + \frac{1}{2} \quad (15)$$

$$\theta_E^{(t)} = (c_1 + c_2)e^t + (c_2 - c_1)e^{-t} + \frac{1}{2} \quad (16)$$

They are visualized in Fig. 5 (left). By setting  $t = 0$ , we can solve for  $c_1 = \frac{1}{2}\theta_I^{(0)} - \frac{1}{4}$  and  $c_2 = \frac{1}{2}\theta_E^{(0)} - \frac{1}{4}$ . Thus, if (and only if)  $\theta_E^{(0)} + \theta_I^{(0)} > 1$ ,  $\theta_I$  and  $\theta_E$  will both tend towards 1, and any  $\theta_I^{(0)} < \frac{1}{2}$  is susceptible to semantic drift (Fig. 5, right).

One minor complication is that probabilities must be between 0 and 1; these equations only govern  $\theta_E, \theta_I \in [0, 1]$ . If we clip these values by defining:

$$\frac{d\theta_E^{(t)}}{dt} = \begin{cases} 0 & \text{if } \theta_E = 0 \\ 0 & \text{if } \theta_E = 1 \\ \theta_I - \frac{1}{2} & \text{otherwise} \end{cases} \quad (17)$$



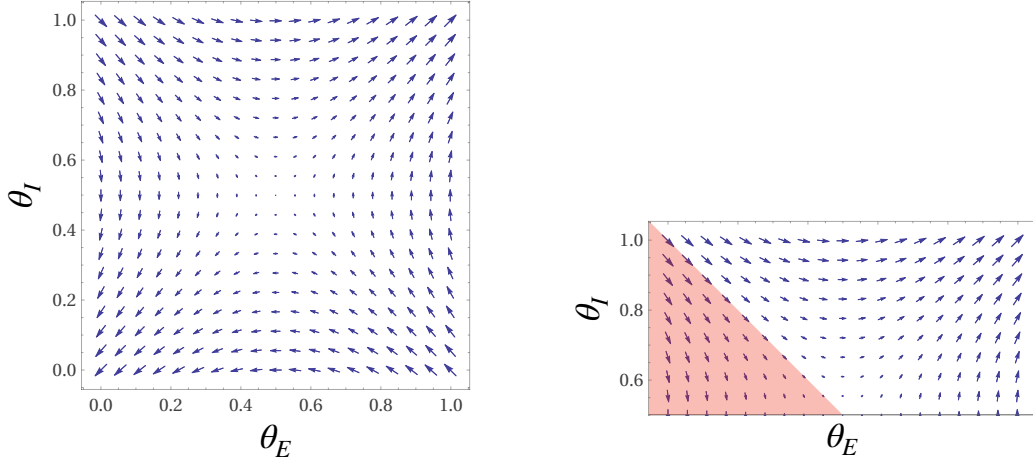


Figure 5: Dynamics of learning in the signaling game. Left: flow field for a single-task pair. Right: parameter initializations that are susceptible to semantic drift (shaded).

defining  $\theta_I$  analogously, and assuming that  $\theta_E^{(0)}, \theta_I^{(0)} \in (0, 1)$ , a small amount of additional work suffices to prove that convergence behavior is the same as the unconstrained case presented above. First observe that once one parameter has reached a value of 1 or 0, the other parameter will converge to the same value: e.g. with  $0 < \theta_I < 1$ ,

$$\frac{d\theta_I}{dt} = \begin{cases} -\frac{1}{2} & \text{if } \theta_E = 0 \\ \frac{1}{2} & \text{if } \theta_E = 1 \end{cases} \quad (18)$$

Parameters will evolve as in the unconstrained case until either  $\theta_E$  or  $\theta_I$  reaches the boundary of the feasible set. This can happen in one of four ways:

- Case 1:**  $\theta_E^{(t)} = 0, 0 < \theta_I^{(1)} \leq 1$ .      **Case 2:**  $\theta_E^{(t)} = 1, 0 < \theta_I^{(1)} \leq 1$ .  
**Case 3:**  $\theta_I^{(t)} = 0, 0 \leq \theta_E^{(1)} < 1$ .      **Case 4:**  $\theta_I^{(t)} = 1, 0 \leq \theta_E^{(1)} < 1$ .

Semantic drift occurs in **Case 1** and **Case 3** and is avoided in **Case 2** and **Case 4**. By setting Eq. (15) and Eq. (16) to 1 and solving for  $t$ , it can be verified that solutions to  $\theta^{(t)} = 1$  exist for positive  $t$  only when  $c_1 + c_2 = \frac{1}{2}\theta_E^{(0)} + \frac{1}{2}\theta_I^{(0)} - \frac{1}{2} > 0$ . Thus **Case 2** and **Case 4** occur (and semantic drift is avoided) only when  $\theta_E^{(0)} + \theta_I^{(0)} > 1$ . ■

**Multitask learning: Proof of Proposition 2** Now suppose we train models for both  $R$  and  $R'$  simultaneously, with a shared executor and reward-specific instructors with parameters  $\theta_{I1}$  and  $\theta_{I2}$ . The expected reward is now:

$$J(\theta_{I1}, \theta_{I2}, \theta_E) = \frac{1}{4} \left( \theta_{I1}\theta_E + (1 - \theta_{I1})(1 - \theta_E) + \theta_{I2}(1 - \theta_E) + (1 - \theta_{I2})\theta_E \right) \quad (19)$$

Then,

$$\frac{d\theta_{I1}^{(t)}}{dt} = \frac{1}{2}\theta_E^{(t)} - \frac{1}{4} \quad (20)$$

$$\frac{d\theta_{I2}^{(t)}}{dt} = -\frac{1}{2}\theta_E^{(t)} + \frac{1}{4} \quad (21)$$

$$\frac{d\theta_E^{(t)}}{dt} = \frac{1}{2}\theta_{I1}^{(t)} - \frac{1}{2}\theta_{I2}^{(t)} \quad (22)$$

Restricting solutions to those satisfying the initial conditions  $\theta_{I1}^{(0)} = \theta_{I2}^{(0)}$ , it can again be verified that

$$\theta_{I1}^{(t)} = \frac{1}{4}e^{-t/\sqrt{2}}(4c_1e^{t/\sqrt{2}} + \sqrt{2}c_2e^{\sqrt{2}t} - \sqrt{2}c_2) \quad (23)$$

$$\theta_{I2}^{(t)} = -\frac{1}{4}e^{-t/\sqrt{2}}(-4c_1e^{t/\sqrt{2}} + \sqrt{2}c_2e^{\sqrt{2}t} - \sqrt{2}c_2) \quad (24)$$

$$\theta_E^{(t)} = \frac{1}{2}\left(\theta_E^{(0)} - \frac{1}{2}\right)\left(e^{(\sqrt{2}-\frac{1}{\sqrt{2}})t} + e^{-\frac{1}{\sqrt{2}}t}\right) + \frac{1}{2} \quad (25)$$

As noted in the body of the paper, the question of whether  $\theta_E \rightarrow 1$  (no semantic drift) is *independent* of  $S^{(0)}$  and  $S'^{(0)}$ , and happens whenever  $\theta_E^{(0)} > \frac{1}{2}$ .

Now consider the clipped version of this objective described in Eq. (17). With the initial conditions  $\theta_{I1}^{(0)} = \theta_{I2}^{(0)}$  and  $\theta_E^{(0)} > \frac{1}{2}$ ,  $\theta_{I1}$  must increase monotonically,  $\theta_{I2}$  must decrease monotonically, and  $\theta_E$  must increase monotonically within the interior of the unit cube until one of the following conditions holds:

**Case 1:**  $\theta_{I1} = 1$ . Thereafter,  $\theta_E > \frac{1}{2}$ ,  $0 \geq \theta_{I2} < 1$ , so  $\frac{d\theta_E}{dt} > 0$ .

**Case 2:**  $\theta_{I2} = 0$ . Thereafter,  $\theta_E > \frac{1}{2}$ ,  $0 < \theta_{I1} \leq 1$ , so  $\frac{d\theta_E}{dt} > 0$ .

**Case 3:**  $\theta_E = 1$  and will remain fixed by definition.

Thus, semantic drift is avoided globally. ■

## B Implementation and hyperparameter details

We use the same executor and instructor architecture and model hyperparameters as used in (Hu et al., 2019). As described in section 4.2, we use a PPO objective to train  $RL^{\text{joint}}$ ,  $RL^{\text{multi}}$  and  $RL^{\text{instr}}$  agents. For all the models, we set the PPO batch size to 32 and the PPO update epochs to 4. We do not use an entropy term as it led to instability during training. We used the Adam optimizer (Kingma and Ba, 2014) and used the learning rate of 6e-6 for  $RL^{\text{joint}}$ ,  $RL^{\text{multi}}$  and 3e-6 for  $RL^{\text{instr}}$ . We swept through the following learning rates:  $[1e-7, 1e-6, 2e-6, 3e-6, 4e-6, 5e-6, 6e-6, 7e-6, 8e-6, 9e-6, 1e-5, 1e-4]$  to pick the best learning rates (during evaluations) for each of the models.

Each of the agents in section 4.4 were trained with several random seeds and played 15000 game episodes per training environment. And the model checkpoints used in the experiments were picked by evaluating those agents on 100 games. 2000 game episodes were used to compute the win rates in tables 2, 3.

$RL^{\text{joint}}$  and  $RL^{\text{multi}}$  models takes approximately 70 hours to train on Intel Xeon Gold 6248 and Nvidia Volta V100, while,  $RL^{\text{instr}}$  takes 35 hours to train on 45000 game episodes. Each instructor in the models presented in the paper have 2.8M parameter, while the executors have 2.4M parameters. Models were implemented in Pytorch(Paszke et al., 2019).

## C Multitask RL LLP training pseudo-code

The training pseudo-code for Multitask RL LLP described in Section 4.3 is presented in Algorithm 1.

---

### Algorithm 1: Multitask LLP RL Training

---

```

N Environments  $M_1, M_2, \dots, M_N$ ;
N Instructors  $\pi_{I1}, \pi_{I2}, \dots, \pi_{IN}$ ;
5 Opponents  $O_1, O_2, \dots, O_5$ ;
Shared Executor  $\pi_E$ ;
for epoch  $i = 1, 2, \dots$  do
  for iteration  $j = 1, \dots, N$  do
    Sample environment  $M_k$ ;
    Select policy  $\pi^{(ij)} = f(\pi_{I_k}, \pi_E)$ 
    Begin  $T$  games  $g_1, \dots, g_T$  where game  $g_\ell$  uses opponent  $O_{\ell \% 5}$ 
    Reset Buffer  $B$ 
    while  $g_1, \dots, g_T$  are not terminated do
      Simulate gameplay between  $\pi^{(ij)}$  and  $O_{\ell \% 5}$ 
      Add each game state, value and actions:  $(s_{g_j(t)}, v_{g_j(t)}, a_{g_j(t)})$  to buffer  $B^{ij}$ .
    end
    Compute rewards  $R_1, \dots, R_T$  for  $g_1, \dots, g_T$ .
    Optimize the PPO objective w.r.t.  $\theta_{I_k}$  and  $\theta_E$  using buffer  $B$  for  $K$  epochs.
  end
end

```

---

## D Game variants

Tables 4–11 below enumerate the attack multipliers of the units under the various multi-task rule sets.

	Attack multiplier				
Unit name	Swordman	Spearman	Cavalry	Archer	Dragon
SWORDMAN	1.0	1.0	1.0	0.5	1.0
SPEARMAN	1.5	0.5	1.0	1.0	0.0
CAVALRY	0.5	1.0	1.5	1.0	0.0
ARCHER	0.5	0.5	0.5	0.5	2.0
DRAGON	1.0	1.5	0.5	1.0	0.0

Table 4: Rule B attack modifier

	Attack multiplier				
Unit name	Swordman	Spearman	Cavalry	Archer	Dragon
SWORDMAN	0.5	1.0	1.5	1.0	0.0
SPEARMAN	1.0	1.5	0.5	1.0	0.0
CAVALRY	1.5	0.5	1.0	1.0	0.0
ARCHER	1.0	1.0	1.0	0.5	1.0
DRAGON	0.5	0.5	0.5	0.5	2.0

Table 5: Rule C attack multipliers

	Attack multiplier				
Unit name	Swordman	Spearman	Cavalry	Archer	Dragon
SWORDMAN	0.5	0.5	0.5	0.5	2.0
SPEARMAN	1.0	1.0	1.0	0.5	1.0
CAVALRY	0.5	1.0	1.5	1.0	0.0
ARCHER	1.5	0.5	1.0	1.0	0.0
DRAGON	1.0	1.5	0.5	1.0	0.0

Table 6: Rule D attack multipliers

	Attack multiplier				
Unit name	Swordman	Spearman	Cavalry	Archer	Dragon
SWORDMAN	1.5	0.5	1.0	1.0	0.0
SPEARMAN	0.5	1.0	1.5	1.0	0.0
CAVALRY	1.0	1.5	0.5	1.0	0.0
ARCHER	1.0	1.0	1.0	0.5	1.0
DRAGON	0.5	0.5	0.5	0.5	2.0

Table 7: Rule E attack multipliers

	Attack multiplier				
Unit name	Swordman	Spearman	Cavalry	Archer	Dragon
SWORDMAN	1.0	1.5	0.5	1.0	0.0
SPEARMAN	1.0	1.0	1.0	0.5	1.0
CAVALRY	1.5	0.5	1.0	1.0	0.0
ARCHER	0.5	0.5	0.5	0.5	2.0
DRAGON	0.5	1.0	1.5	1.0	0.0

Table 8: Rule F attack multipliers

	Attack multiplier				
Unit name	Swordman	Spearman	Cavalry	Archer	Dragon
SWORDMAN	0.5	1.0	1.5	1.0	0.0
SPEARMAN	1.5	0.5	1.0	1.0	0.0
CAVALRY	0.5	0.5	0.5	0.5	2.0
ARCHER	1.0	1.0	1.0	0.5	1.0
DRAGON	1.0	1.5	0.5	1.0	0.0

Table 9: Rule G attack multipliers

	Attack multiplier				
Unit name	Swordman	Spearman	Cavalry	Archer	Dragon
SWORDMAN	0.5	1.0	1.5	1.0	0.0
SPEARMAN	1.5	0.5	1.0	1.0	0.0
CAVALRY	1.0	1.5	0.5	1.0	0.0
ARCHER	0.5	0.5	0.5	0.5	2.0
DRAGON	1.0	1.0	1.0	0.5	1.0

Table 10: Rule H attack multipliers

	Attack multiplier				
Unit name	Swordman	Spearman	Cavalry	Archer	Dragon
SWORDMAN	0.5	1.0	1.5	1.0	0.0
SPEARMAN	1.0	1.0	1.0	0.5	1.0
CAVALRY	1.0	1.5	0.5	1.0	0.0
ARCHER	0.5	0.5	0.5	0.5	2.0
DRAGON	1.5	0.5	1.0	1.0	0.0

Table 11: Rule J attack multipliers