**ID number) Experiment descriptions**

**RLLib**

1. The manager observes the goal and communicates a goal to the worker (in a set of possible landmarks), then receives a reward equal to the inverse distance of the listener from the goal. The worker computes his reward locally.
2. Equal to (1), but the worker receives as reward the sum of the locally computed (inverse distance from the communicated goal) reward and the environmental reward.
3. No manager. The worker receives a sparse reward whenever it reaches a goal.
4. The manager observes the goal and communicates the worker reward directly to the worker, then receives a reward equal to the inverse distance of the listener from the goal. The worker receives as reward the one communicated by the manager. There is pretraining and extended communication heuristic.
5. Equal to (4) but there is neither pretraining or extended communication heuristic.
6. Equal to (5) but the manager not only observes the goal, but also the inverse distance of the worker from the goal
7. Equal to (5) but the manager not only observes the goal, but also the inverse distance of the worker from each landmark
8. Meta-learning: the worker is trained from scratch at each meta-training iteration and receives as reward the reward communicated by the manager. The manager is meta-trained with random search. It observes the goal landmark and gets as reward the inverse distance from the goal target. The benchmark used for the meta-training is the sum (across all episodes of one experiment) of the rewards obtained by the manager.
9. Meta-learning: the worker is trained from scratch at each meta-training iteration and receives as reward the reward communicated by the manager. The manager is meta-trained with a genetic algorithm. It observes the goal landmark and gets as reward the inverse distance from the goal target. The objective function to minimize in the meta-training is the sum (across all episodes of one experiment) of the rewards obtained by the manager.

**Keras**

1. Algorithm is reinforce. The manager observes the goal and communicates a goal to the worker (in a set of possible landmarks), then receives a reward equal to the inverse distance of the listener from the goal. The worker computes his reward locally.
2. .
3. .
4. .
5. Algorithm is reinforce. Equal to (4) but there is neither pretraining or extended communication heuristic.