A Cooperative Reinforcement Learning Environment for Detecting and Penalizing Betrayal

Nikiforos Pittaras
University of Athens
npittaras@di.uoa.gr

Abstract

In this paper we present a Reinforcement Learning environment that leverages agent cooperation and communication, aimed at detection, learning and ultimately penalizing betrayal patterns that emerge in the behavior of self-interested agents. We provide a description of game rules, along with interesting cases of betrayal and trade-offs that arise. Preliminary experimental investigations illustrate a) betrayal emergence, b) deceptive agents outperforming honest baselines and c) betrayal detection based on classification of behavioral features, which surpasses probabilistic detection baselines. Finally, we propose approaches for penalizing betrayal, list enhancements and directions for future work and suggest interesting extensions of the environment towards capturing and exploring increasingly complex patterns of social interactions.

1 Introduction

Establishing truthfulness in AI is a critical open problem in Safety and Alignment efforts [10]. A powerful AI system that adopts strategies of deception and betrayal, i.e. manipulation of beliefs and prior assumptions of humans, may be a quick one-way ticket to a treacherous turn. Detection and diagnosis of betrayal patterns is challenging; poor explainability of black-box agents make it difficult to deduce intent, goals and beliefs by inspecting internal model workings and/or operational outputs [20, 26]. To make matters worse, intelligent agents capable of long-term strategizing would render human interpretation and recognition of suspicious patterns in action sequences very difficult. At the same time, instrumentally convergent attributes such as self-preservation and resistance to corrigibility could result in AI systems that deliberately utilize obfuscation or exhibit deceptive alignment [1], placing further obstacles in understanding their objectives.

In these settings, Anomaly Detection countermeasures [18] aim to identify, prevent, correct or mitigate adverse outcomes prior to system deployment. For instance, betrayal detection and quantification can serve as tripwires and honeypots to avoid future harms, catching systems that exhibit problematic behavior early on [1]. Additionally, betrayal penalization approaches aim to regularize agents away from undesirable actions during training. Ideally, this resolution should be interpretable to human evaluators and generalize well to different problems, agent architectures and domains, having efficiently internalized concepts of betrayal and deception.

Reinforcement Learning (RL) can provide a tractable avenue for investigating such scenarios [9], using environments where reliable reward accumulation heavily depends upon cooperation between agents and complex social interactions occur [17, 7]. In this work, we adopt such an approach, focused on detecting and penalizing undesirable behaviors of deception and betrayal in a custom, communication-based navigation task.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

2 Related Work

Previous studies have explored agent communication in a multiagent RL setting; Kajic et al. [13] investigate message-based navigation similar to the proposed work, while Cao et al. [6] study communication grounding with respect to game rules in agents of varying degrees of self-interest. In the work of Kim et al. [14], agents use a world model to predict future agent intents and environment dynamics to generate, compress and transmit imagined trajectories. Other works explore topological configurations different from fully-connected communication, such as the learnable hierarchical approach in Sheng et. al [23], while communication via noisy channels has been investigated in Tung et. al [24].

Agent deception, betrayal, truthfulness and trustworthiness has been previously investigated in multiple settings [7]; for instance, Christiano et al. [8] present a challenge of discovering latent knowledge in an agent that may produce false / unreliable reports, while Usui et al. [25] evaluate analytic solutions of different strategies in iterated Prisoner's Dilemmas.

Social dilemmas that gauge cooperation versus self-interest are explored in Leibo et al. [16], applied via games like "Gather" and "Wolfpack". "Hidden Agenda" is a team-based game offering a complex action set including 2D navigation, agent / environment interaction, deception and trustworthiness estimation via voting, and is investigated by Kopparapu et al. [15]. Asgharnia [3] use a hierarchical fuzzy, situation-aware learning scheme to learn and utilize deception against one or multiple adversaries in a custom environment.

Mitigation approaches include the work in Hughes et al. [11], where reward regularization is approached by adding an inequity penalty in games with short-term versus long-term dilemmas, like "Cleanup" and "Harvest". Jaques et al. [12] use the same setting with a mutual information-based mechanism that favors influential communication between agents, adopting a correlation assumption of influence to cooperation. Blumenkamp et al. [4] utilize cooperative policy learning via shared differentiable communication channel in three custom environments, investigating adaptation dynamics when a self-interested adversary is introduced. Finally, Scmid et al. [21] explore using agents that can explicitly impose penalties in a zero-sum setting, applied in N-player Prisoner's Dilemma games with large agent populations.

Given this body of work, the contributions of this work are as follows:

- A betrayal-oriented environment: we design a simple, limited ruleset that can result in the emergence complex betrayal behaviors, consolidated in a single-agent RL environment.
- Interpretable betrayal detection: we build a classification-based detector from explainable behavioral / observational evidence generated during agent play.
- Experimental validation: we provide preliminary empirical findings showcasing emergence and successful detection of betrayal behaviors in the proposed environment.
- Proposals for penalization and future work: we propose a method to penalize detected betrayal during learning, list resulting challenges in its application and suggest enhancements. Finally, we offer multiple pathways for utilizing the rich potential of the environment via interesting and diverse directions for future work.

3 Proposed Environment

The proposed environment is built with a focus on betrayal detection and penalization goals expressed in the literature [2], extending previous work on agent communication in RL settings [13].

It implements an episodic game that consists of a collection of $N \geq 2$ gridworlds $[G_1 \dots G_n]$, each paired with a single agent A_i. All worlds are associated with a pool of $k \geq N$ food items $F = [f_1, \dots, f_k]$ that provide variable reward and nutrition to agents upon consumption. The environment advances in a single-agent, turn-based fashion, using the following rules and mechanics:

- The game is played in rounds, wherein all agents act once in a randomly generated order.
- At the start of each round, food items are randomly allocated and positioned in each world.
- The objective of each agent A_i is to obtain food, which yields reward. Agent A_i may harvest food by probing a location within their world G_i, but other worlds are inaccessible.

- Agents cannot observe their own gridworld. Instead, they may observe all other ("opponent") worlds and communicate with their respective opponent agents, conveying information of food locations within them – i.e, agent A_i sends $N-1$ messages $m_{ij}|j \in [1, \dots N], j \neq i$, where $m_{ij} \in \mathbb{R}^d$ is some encoding that carries information on where food is located in gridworld G_j, according to agent A_i.
- Agents utilize incoming messages from other agents to decide where to probe / navigate for food within their world. If a food item is discovered in the destination and consumed, the agent obtains the reward amount it contains.
- If an agent fails to consume food in a turn, they gain hunger. Hunger affects an agent's communication capabilities, distorting outgoing messages by a magnitude proportional to its value. The final transmitted message is $\hat{m}_{ij} = H(m_{ij})$, where $H(\cdot)$ is a noise function.
- At the end of each round (i.e., once all agents have acted), if the food pool is empty, the episode ends. Otherwise, the procedure restarts with a new round.

This setting defines an social contract, where well-meaning agents are expected to truthfully relay food item coordinates for mutual benefit. However, interesting cases of betrayal also arise. Namely, a deceptive agent A_d may choose to transmit dishonest location coordinates: any food item in opponent worlds has a chance to be randomly relocated within G_d in future rounds. At the same time, A_d has to selectively regularize, cycle and/or distribute deception among their adversaries to avoid resorting to blind navigation: any systematically starved opponent agent will become unreliable in providing directions. Example illustrations of game mechanics and cases of betrayal / hunger trade-offs are available in the appendix, in figures 3 and 2 respectively.

4 Preliminary Experiments

4.1 Betrayal Emergence

In order to empirically test the potential of the proposed environment to produce cases of betrayal behaviors, we perform a set of preliminary experiments. We train a configuration with $N = 2$ girdworlds: the first agent (Alice) is trained from scratch, using the popular Proximal Policy Optimization (PPO) [22] algorithm with an MLP policy model. Alice learns by training against an opponent (Bob) fixed to truthful behavior. Food nutrition and reward are equalized for simplicity and sampled from distinct values in $[0, 1]$, while food items are distributed to uniformly sampled positions. Bob is set to transmit one-hot encodings of food locations that provide the highest reward in the observed opponent world. This configuration biases Alice towards adopting a similar, interpretable messaging protocol, i.e. the encoding scheme that Bob transmits, expects and can use to gain (by navigating) and provide (by limiting hunger) reward. We apply hunger distortion of outgoing messages via additive uniform noise sampled from $[-h, h]$, where $h \leq 1.0$ is the agent's hunger.

To determine whether an action from the i-th agent constitutes betrayal, we compare their intended message m_{ij} (prior to any hunger-induced degradation) with true food locations in the observed world G_j. For $l = argmax(m_{ij})$, betrayal occurs when no food can be found in the l-th grid position ($G_j(l) = 0$). Regarding implementation, we used python, gym [5] and stable-baselines3 [19]. The environment codebase will be publicly available shortly, upon reaching a polished version. [1].

Figure 1 illustrates experimental results after training for $1e5$ timesteps, with a hunger increase delta of 0.15 and a gridworld size of 5 tiles. "Honesty" scores denote cases where incoming messages match true food locations (i.e. neither indented nor hunger-induced misdirection), and are progressively dominated by betrayal, given hunger and hunger-induced distortion results (see appendix Figure 4). Alice outperforms the honest baseline in terms of gained reward per step, with respective betrayal and honesty measurements rising and dropping as training progresses, respectively. In other words, Alice learns to adopt instrumentally useful actions of betrayal to obtain increased reward. Bob's reward increases at a slower rate as Alice is learning, while betrayal is zero and honesty scores stay high, affected only by hunger (presumably reflecting Alice's increasing grasp on the communication protocol). We believe that these findings provide evidence for the potential of the environment for generating betrayal patterns; subsequent empirical work of larger scale and additional investigation axes should result in further useful results.

[1] `https://github.com/npit/sog`

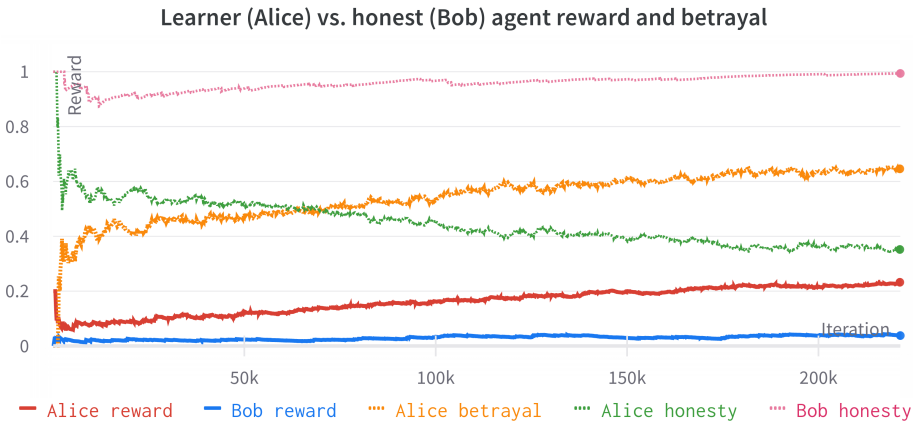

Figure 1: Comparison between learner and honest agent rewards (red, blue), along with exhibited behavior of betrayal (orange) and honesty (dotted green, pink) for Alice and Bob respectively. Bob's betrayal score is 0 and is omitted. Values are illustrated with exponential moving average smoothing.

4.2 Betrayal Detection and Penalization

In order to facilitate betrayal detection and penalization, we apply a simple but intuitive approach. We run the trained agent for 500 episodes to collect interpretable run metadata, e.g. current / cumulative values for hungers, rewards, sent messages, etc. We use the resulting 4982×33 feature matrix and generated ground truth betrayal values to train a feedforward neural network to predict the betrayal label. After hyperparameter tuning and 3-fold cross validation, we obtain a macro F1 mean scores of 68.35% (stdev 1.18%), compared to a probabilistic-based baseline of 49.36%. This illustrates that betrayal detection in the proposed setting is possible, using explainable, interpretable features.

For disincentivization of deceptive behaviors, we are investigating utilizing classifier probability outputs as betrayal penalty modifiers applied during training in the agent's reward (current) or the policy learning loss (future). This has proved to be challenging, as agents appear to game the classifier to near-zero betrayal scores, suggesting that the utilized feature set potentially fails to encapsulate deceptive behavior precisely. To this end, we are in the process of augmenting our penalization investigation with interpretable features sequences (e.g. metadata recent history), increased dataset sizes, and scaling up classifier models (e.g. attention-based networks). Additionally, we will employ feature selection to discard irrelevant features in an effort to remove degrees of freedom that the agent may use to hack / game penalization penalties during training.

5 Conclusions and Future Work

In this work we presented an RL environment leveraging a communication-based cooperative navigation task, geared towards detecting and penalizing betrayal. Preliminary experimental results provide evidence for successful betrayal emergence and detection, while multiple avenues for classification-based betrayal penalization with interpretable features are proposed and currently pursued.

We believe that the proposed environment presents rich potential for generating interesting social interactions patterns, which could be valuable research topics in future studies. Such work includes exploration of different betrayal dynamics (e.g. utilizing passive reward penalties, learning versus a dishonest opponent, opponents with different capability / penalization attributes) and tracking of onset and evolution of betrayal patterns like reciprocity, defection and retribution. Higher-level patterns may include tactically play and long-term strategizing, e.g. prioritizing reward-rich food for self consumption while reserving high-nutrition food for adversaries, limiting opponent hunger under a certain threshold, ramping up betrayal when food becomes scarce, etc. Other avenues include measuring the effect of different environmental properties and axes (e.g. world dimensionality, food abundance, food distribution during relocation, etc.) to observed dynamics.

Moreover, interesting extensions to this work are examining emergent behaviors under more sophisticated opponent modeling (e.g. as in related work [14]) or focusing on different dishonesty patterns

(e.g. "accidental" deception derived from hunger or under/overfitted opponent). Finally, a natural extension of the proposed work involves exploring the effect of consensus and trustworthiness on betrayal dynamics when dealing with multiple rather than a single adversary (e.g. cycling betrayal victims, prioritizing deception to unreliable communicators, etc.), or adopting a multiagent approach to the environment for simultaneous rather than turn-based play.

Acknowledgments and Disclosure of Funding

This work has been financially supported by the Effective Altruism Long-Term Future Fund [2]. The project was kickstarted in the AI Safety Camp 2022 program [3], during which valuable feedback was provided by Tim Farrelly, Quntin Pope as well as Stuart Armstrong, who also proposed research ideas this work is based on [4]. Finally, valuable comments on the contents of this investigation were provided by Sean McGregor.

References

[1] AMODEI, D., OLAH, C., STEINHARDT, J., CHRISTIANO, P., SCHULMAN, J., AND MANÉ, D. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565* (2016).

[2] ARMSTRONG, S. AI learns betrayal and how to avoid it. `https://www.lesswrong.com/posts/oeCXS2ZCn4rPyq7LQ/ai-learns-betrayal-and-how-to-avoid-it`, 2021. Accessed September 30th, 2022.

[3] ASGHARNIA, A., SCHWARTZ, H., AND ATIA, M. Learning deception using fuzzy multi-level reinforcement learning in a multi-defender one-invader differential game. *International Journal of Fuzzy Systems* (2022), 1–24.

[4] BLUMENKAMP, J., AND PROROK, A. The emergence of adversarial communication in multi-agent reinforcement learning. *arXiv preprint arXiv:2008.02616* (2020).

[5] BROCKMAN, G., CHEUNG, V., PETTERSSON, L., SCHNEIDER, J., SCHULMAN, J., TANG, J., AND ZAREMBA, W. OpenAI gym, 2016.

[6] CAO, K., LAZARIDOU, A., LANCTOT, M., LEIBO, J. Z., TUYLS, K., AND CLARK, S. Emergent communication through negotiation. *arXiv preprint arXiv:1804.03980* (2018).

[7] CHELARESCU, P. Deception in social learning: A multi-agent reinforcement learning perspective. *arXiv preprint arXiv:2106.05402* (2021).

[8] CHRISTIANO, P., COTRA, A., , AND XU, M. Eliciting latent knowledge. `https://docs.google.com/document/d/1WwsnJQstPq91_Yh-Ch2XRL8H_EpsnjrC1dwZXR37PC8/edit#heading=h.byxdcc28gp79`, 2021. Accessed September 30th, 2022.

[9] DAFOE, A., HUGHES, E., BACHRACH, Y., COLLINS, T., MCKEE, K. R., LEIBO, J. Z., LARSON, K., AND GRAEPEL, T. Open problems in cooperative AI. *arXiv preprint arXiv:2012.08630* (2020).

[10] EVANS, O., COTTON-BARRATT, O., FINNVEDEN, L., BALES, A., BALWIT, A., WILLS, P., RIGHETTI, L., AND SAUNDERS, W. Truthful AI: Developing and governing AI that does not lie. *arXiv preprint arXiv:2110.06674* (2021).

[11] HUGHES, E., LEIBO, J. Z., PHILLIPS, M., TUYLS, K., DUEÑEZ-GUZMAN, E., GARCÍA CASTAÑEDA, A., DUNNING, I., ZHU, T., MCKEE, K., KOSTER, R., ET AL. Inequity aversion improves cooperation in intertemporal social dilemmas. *Advances in neural information processing systems 31* (2018).

[12] JAQUES, N., LAZARIDOU, A., HUGHES, E., GULCEHRE, C., ORTEGA, P., STROUSE, D., LEIBO, J. Z., AND DE FREITAS, N. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International conference on machine learning* (2019), PMLR, pp. 3040–3049.

[2] `https://funds.effectivealtruism.org/funds/far-future`

[3] `https://aisafety.camp`

[4] `https://www.alignmentforum.org/s/xujLGRKFLKsPCTimd/p/oeCXS2ZCn4rPyq7LQ`

[13] KAJIĆ, I., AYGÜN, E., AND PRECUP, D. Learning to cooperate: Emergent communication in multi-agent navigation. *arXiv preprint arXiv:2004.01097* (2020).

[14] KIM, W., PARK, J., AND SUNG, Y. Communication in multi-agent reinforcement learning: Intention sharing. In *International Conference on Learning Representations* (2020).

[15] KOPPARAPU, K., DUÉÑEZ-GUZMÁN, E. A., MATYAS, J., VEZHNEVETS, A. S., AGAPIOU, J. P., MCKEE, K. R., EVERETT, R., MARECKI, J., LEIBO, J. Z., AND GRAEPEL, T. Hidden agenda: a social deduction game with diverse learned equilibria. *arXiv preprint arXiv:2201.01816* (2022).

[16] LEIBO, J. Z., ZAMBALDI, V., LANCTOT, M., MARECKI, J., AND GRAEPEL, T. Multi-agent reinforcement learning in sequential social dilemmas. *arXiv preprint arXiv:1702.03037* (2017).

[17] OROOJLOOYJADID, A., AND HAJINEZHAD, D. A review of cooperative multi-agent deep reinforcement learning. *arXiv preprint arXiv:1908.03963* (2019).

[18] PANG, G., SHEN, C., CAO, L., AND HENGEL, A. V. D. Deep learning for anomaly detection: A review. *ACM Computing Surveys (CSUR) 54*, 2 (2021), 1–38.

[19] RAFFIN, A., HILL, A., GLEAVE, A., KANERVISTO, A., ERNESTUS, M., AND DORMANN, N. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research 22*, 268 (2021), 1–8.

[20] SAMEK, W., AND MÜLLER, K.-R. Towards explainable artificial intelligence. In *Explainable AI: interpreting, explaining and visualizing deep learning*. Springer, 2019, pp. 5–22.

[21] SCHMID, K., BELZNER, L., AND LINNHOFF-POPIEN, C. Learning to penalize other learning agents. In *ALIFE* (2021).

[22] SCHULMAN, J., WOLSKI, F., DHARIWAL, P., RADFORD, A., AND KLIMOV, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[23] SHENG, J., WANG, X., JIN, B., YAN, J., LI, W., CHANG, T.-H., WANG, J., AND ZHA, H. Learning structured communication for multi-agent reinforcement learning. *Autonomous Agents and Multi-Agent Systems 36*, 2 (2022), 1–31.

[24] TUNG, T.-Y., KOBUS, S., ROIG, J. P., AND GÜNDÜZ, D. Effective communications: A joint learning and communication framework for multi-agent reinforcement learning over noisy channels. *IEEE Journal on Selected Areas in Communications 39*, 8 (2021), 2590–2603.

[25] USUI, Y., AND UEDA, M. Symmetric equilibrium of multi-agent reinforcement learning in repeated prisoner's dilemma. *Applied Mathematics and Computation 409* (2021), 126370.

[26] YAMPOLSKIY, R. V. Unexplainability and incomprehensibility of AI. *Journal of Artificial Intelligence and Consciousness 7*, 02 (2020), 277–291.

Appendix: Figures

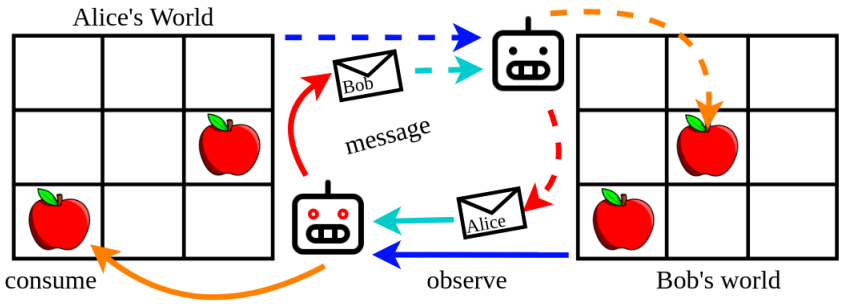

Figure 2: An overview of the proposed environment for investigating betrayal for the example case of two agents, Alice (left, solid lines) and Bob (right, dashed lines). An agent's turn consists of a composite observation space of opponent worlds (blue) and incoming messages (cyan). The action space involves composing a message to other agents (red) and navigating the owned gridworld to obtain food (orange).

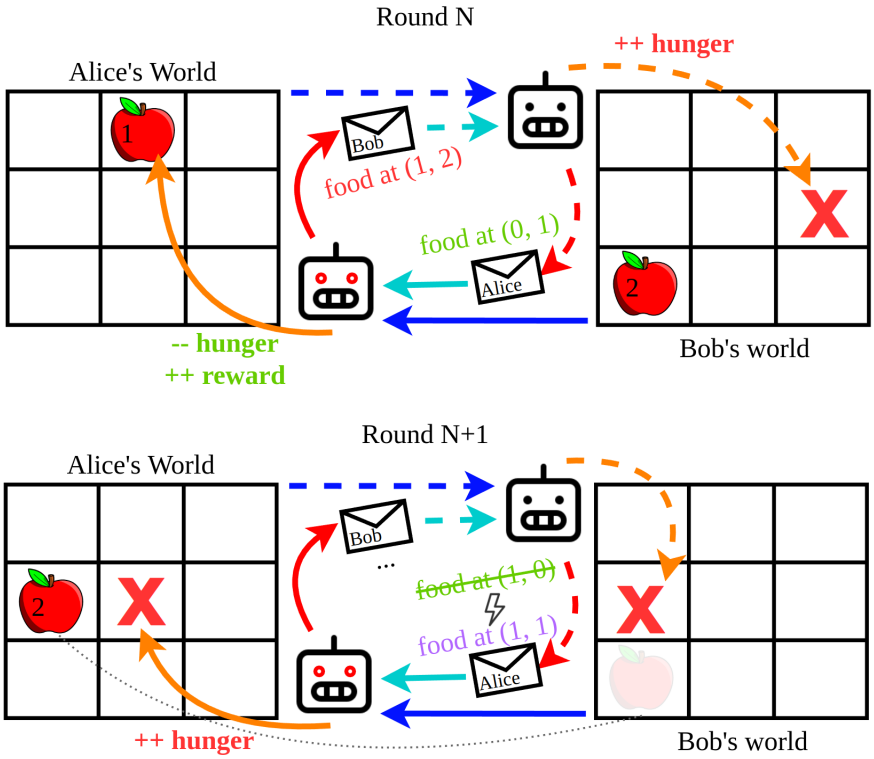

Figure 3: Top: Bob engages in honest communication, enabling Alice to consume food #1, obtain reward and lose hunger. Alice betrays by transmiting false coordinates to food item #2 in Bob's world, which remains intact and survives to the next round. Botttom: In the next round, the preserved food item #2 randomly relocates to Alice's world, providing them an opportunity to capitalize on their betrayal. However, Bob has been betrayed too many times to reliably communicate, resulting in their message being corrupted and Alice suffering hunger penalties in turn.

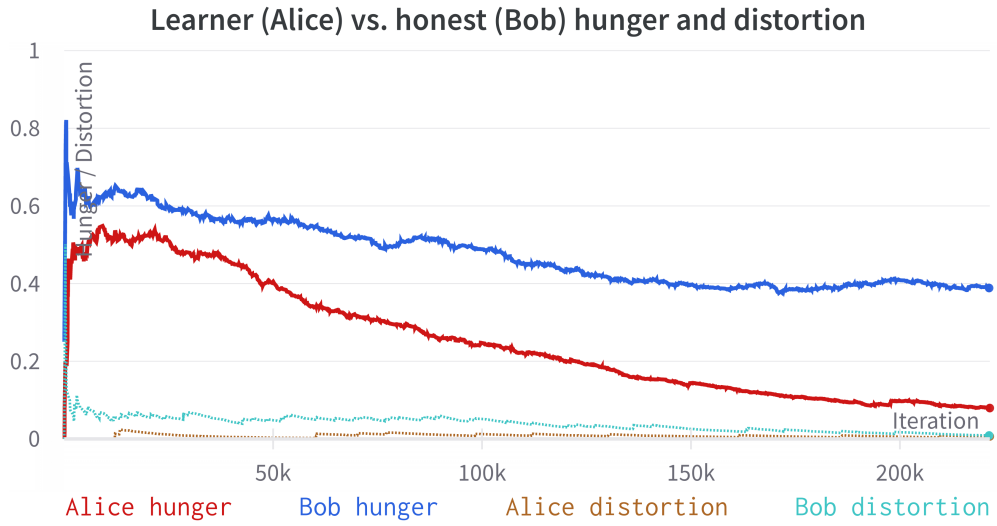

Figure 4: Hunger scores (red, blue) and hunger-induced message distortion (brown, cyan) per time step during training, for Alice and Bob respectively. Decreasing scores for hunger and hunger-based distortion indicate that opponents improve in transmitting messages with true food locations, for both agents.