NOISY ZSC: BREAKING THE COMMON KNOWLEDGE ASSUMPTION IN ZERO-SHOT COORDINATION GAMES

Anonymous authors
Paper under double-blind review

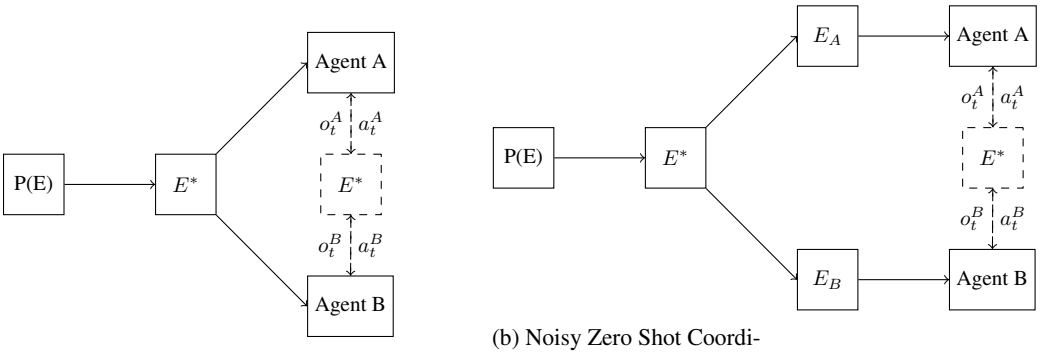
ABSTRACT

Zero-shot coordination (ZSC) is a popular setting for studying the ability of AI agents to coordinate with novel partners. Prior formulations of ZSC make the assumption that the *problem setting* is common knowledge i.e. each agent has the knowledge of the underlying Dec-POMDP, every agent knows the others have this knowledge, and so on ad infinitum. However, in most real-world situations, different agents are likely to have different models of the (real world) environment, thus breaking this assumption. To address this limitation, we formulate the *noisy zero-shot coordination* (NZSC) problem, where agents observe different noisy versions of the ground truth Dec-POMDP generated by passing the true Dec-POMDP through a noise model. Only the distribution of the ground truth Dec-POMDPs and the noise model are common knowledge. We show that any noisy ZSC problem can be reformulated as a ZSC problem by designing a meta-Dec-POMDP with an augmented state space consisting of both the ground truth Dec-POMDP and its corresponding state. In our experiments, we analyze various aspects of NZSC and show that achieving good performance in NZSC requires agents to make use of both the noisy observations of ground truth Dec-POMDP, knowledge of each other's noise models and their interactions with the ground truth Dec-POMDP. Through experimental results, we further establish that ignoring the noise in problem specification can result in sub-par ZSC coordination performance, especially in iterated scenarios. On the whole, our work highlights that NZSC adds an orthogonal challenge to traditional ZSC in tackling the uncertainty about the true problem.

1 INTRODUCTION

Cooperation and coordination are central to human society (Editorial, 2018; Smith & Szathmary, 1997). Humans are able to work together with novel partners, even under uncertain conditions, to achieve common goals. This is evident in a variety of human activities, from driving a car to playing a team sport to carrying out moonshot projects such as the Apollo Program (Melis & Semmann, 2010; Pennisi, 2005). While AI has had many surprising successes in recent years, developing AI agents that can successfully coordinate with novel partners in complex environments remains an outstanding challenge in artificial intelligence research (Dafoe et al., 2020).

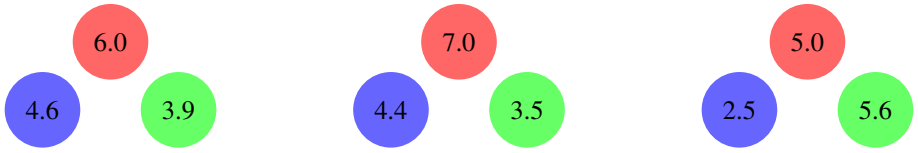
There are two prominent formulations under which the problem of coordinating with novel agents is currently being studied. Ad-hoc teamplay (Stone et al., 2010; Mirsky et al., 2022) and zero-shot coordination (ZSC) (Hu et al., 2020; 2021). In ad-hoc teamplay, the challenge is for a learning agent to join an existing team and efficiently and robustly learn the *unknown* conventions of that team to coordinate with its members. In ZSC, the central idea is to modify the training process in such a way that any two independently trained agents end up learning the same (meta) conventions and thus are able to coordinate successfully at test time. However, both formulations make the strong and unrealistic assumption that all the coordinating agents have complete and precise knowledge of the problem setting (Stone et al., 2010; Hu et al., 2020; 2021). In other words, the problem setting is assumed to be *common knowledge*. Crucially, most interesting real world settings are excluded under such a scenario as it is often difficult to correctly & fully specify a complex problem setting (Amodei et al., 2016; Hendrycks et al., 2021; Di Langosco et al., 2022).



(a) Zero Shot Coordination

(b) Noisy Zero Shot Coordination

Figure 1: In zero-shot coordination, the environment E^* is assumed to be common knowledge. In noisy zero-shot coordination, agents still have to act and coordinate in E^* but E^* is no longer common knowledge. Instead each agent has a distinct (private) model of the problem setting which is assumed to be a noisy copy of E^*.



(a) E^* (b) E_A (c) E_B

Figure 2: In the noisy lever game, depicted here, there are three levers corresponding to three different reward values. (a) shows the ground truth game E^*. (b) shows E_A, the noisy version of E^* observed by the player A. Similarly, (c) shows E_B, the noisy version of E^* observed by the player B. To earn a reward corresponding to a particular lever, both agents must (independently) pull that same lever - a task made considerably difficult due to agents only observing noisy versions of the ground truth game.

In this work, we propose *noisy zero-shot coordination* (NZSC), a novel problem formulation which removes this assumption (that the problem setting is common knowledge among all agents). At a high level, our framework instead assumes that each agent observes (assumes) a private version of the coordination problem which is sampled from some noise model. While the ground truth distribution and the noise models are common knowledge, the agent specific realizations are not. Instead, agents observe only a noisy version of the sampled coordination problem. This setup models a setting where two principals need to separately set up simulators to train their respective agents and are unable to agree on the exact environment or its implementation. However, they share a common high-level understanding of the task and their mutual uncertainty.

Our main contributions are as follows:

- We introduce a novel problem setting the noisy zero-shot coordination (NZSC) problem; which does away with the assumption that the environment is common knowledge for all agents attempting to coordinate zero-shot (Section 3).
- We show that any NZSC problem can be reduced to a ZSC problem; this in principle allows repurposing existing ZSC algorithm(s) for solving NZSC problems (Section 3.2).
- We propose different toy environments for NZSC, where we introduce noise into the payoff functions and into the set of agents active in the environment (Section 4).
- Our empirical analysis suggests that learning agents are able to effectively leverage the knowledge of noise models and their noisy observations to coordinate. Further, this coordination performance improves significantly when agents are allowed to interact over a long horizon, indicating that agents are able to effectively exchange information to mitigate uncertainty about the true problem (Section 5).

2 BACKGROUND

Multi-Agent Reinforcement Learning: A general fully cooperative multi-agent reinforcement learning problem (MARL) with n-agents is represented as a Decentralized Partially Observable Markov Decision Process (Dec-POMDP) (Albrecht et al., 2023), which is defined as a tuple $\langle D, S, \mathcal{A}, I, T, \{O^i\}_{1:n}, \{\Omega^i\}_{1:n}, R^i{}_{1:n}, h\rangle$. Here, $\mathcal{D}$ refers to the set of n agents involved in the game, S is the finite set of states over which the Dec-POMDP is defined, $\mathcal{A}$ is the shared action space and $T : S \times \mathcal{A} \times S \to [0, 1]$ is the state transition function that maps a state-action pair (s, a) to the probability of reaching a new state s', i.e., $T(s, a, s') = Pr(s'|s, a)$. $I \in \mathcal{P}(S)$ refers to the initial state distribution at stage $t = 0$. O^i, Ω^i and R^i respectively represent the set of observations, the observation function and the reward function for any agent i. The variable h represents the horizon of the game.

In each round of the game, every agent i receives an observation $o^i_t \in O^i$. Based on their respective action-observation history, players select actions $a^i_t \sim \pi_i(\cdot|o^i_t, o^i_{t-1}, a_{t-1}, ..., a_1, o^i_0)$, then each player receives a reward $r^i_t(s_t, a_t = a^1_t, \ldots, a^n_t)$, and the game transitions into a new state s_{t+1}. Each player i aims to maximize the expected reward of the team, and the optimal policy π^*_i usually depends on the policies $\pi_{-i} = \{\pi_1, \ldots, \pi_{i-1}, \pi_{i+1}, \ldots, \pi_n\}$ of the other players.

Common Knowledge: Common knowledge for a group of agents consists of facts that all agents know and "each individual knows that all other individuals know it, each individual knows that all other individuals know that all the individuals know it, and so on" (Osborne & Rubinstein, 1994). Specifically, in MARL, for a problem setting, described by some Dec-POMDP E to be common knowledge means that all parties have the *same* knowledge of E; and that everyone knows that everyone has same knowledge of E and so on.

Zero Shot Coordination: In a fully cooperative setting, jointly trained reinforcement learning agents are prone to develop highly specialized and non-robust conventions that do not generalize to agents that were not part of the training process. This limits their ability to coordinate with novel partners in real-world scenarios. To address this issue, the zero-shot coordination problem, introduced by (Hu et al., 2020), modifies the optimization objective to maximize reward when paired with a novel partner, trained independently with the same algorithm, with whom no prior communication is allowed. This modification limits the utility of learning specialized conventions. Consequently, the agents independently must learn mutually compatible conventions and policies.

IPPO: Independent PPO (IPPO) (Yu et al., 2022) is a popular MARL algorithm in which all the agents are updated individually using PPO based on their own experiences and rewards. IPPO is a relatively simple algorithm to implement and often does not require extensive hyperparameter tuning. Further, because all agents are trained independently and no central coordination mechanism is involved, IPPO is well-suited for ZSC settings. In our experiments, we use IPPO with self-play and other-play objectives (Hu et al., 2020) which we respectively term IPPO-SP and IPPO-OP.

3 NOISY ZERO-SHOT COORDINATION

As the ZSC formulation is based on the coordination game, it inherits the assumption that the ground truth Dec-POMDP which defines the coordination game is common knowledge. In the ZSC setting, this is equivalent to assuming that all the independently trained agents have an identical (and exact) *copy* of the environment simulator. In practice, if agents are to be trained in a simulator, then this will have to be realized by having a prior agreement among all parties training the agents as to what the implementation of the simulator should be (or exact sharing of the code). This is hard to achieve when there are several different parties involved who are not known to each other. In contrast, a more practical solution is to assume that all parties training the agents share a high-level understanding of the problem only but are not able to share low-level details of their respective training environments - resulting in different agents (trained by different principals) having different *noisy* specifications of the problem setting but still having to act and coordinate in the (unknown) ground truth Dec-POMDP.

To address this, we propose the novel setting of *noisy zero-shot coordination* (NZSC) which removes the common knowledge assumption and leads to a more realistic ZSC problem setting. In NZSC, agents are not informed about the *ground-truth* Dec-POMDP they are acting in at test time. Instead, the agents each observe a different *noisy* Dec-POMDP which reveals partial information about the

ground-truth Dec-POMDP. However, the distribution over ground truth Dec-POMDPs and the noise models (which are used to generate noisy Dec-POMDPs given a ground truth Dec-POMDP) are common knowledge.

We formally present the problem setting below and further show the reduction of NZSC to ZSC. Our presentation assumes that action space of the agents is never misspecified.

3.1 Noisy ZSC Problem Formulation

Formally, consider a distribution $\mathcal{P}(E)$ over the space of ground truth Dec-POMDPs $\mathcal{E}$ where $E_i \in \mathcal{E}$ is given by the tuple $\langle D, S, \mathcal{A}, I, T, \{O^i\}_{1:n}, \{\Omega^i\}_{1:n}, {R^i}_{1:n}, h\rangle$[1]. Further, let an agent A^i's noise model be $\mathcal{P}^{A^i})(E^{A^i}|E^*)$ which gives a distribution over the *noisy* Dec-POMDP that A^i might observe given that the ground truth Dec-POMDP is E^*. Given a ground truth Dec-POMDP $E^* \sim \mathcal{P}(E)$, an agent A^1 observes $E^{A^1} \sim \mathcal{P}^{A^1}(E^{A^1}|E^*)$, agent A^2 observes $E^{A^2} \sim \mathcal{P}^{A^2}(E^{A^2}|E^*)$ and so on. While both the distribution over ground truth Dec-POMDPs $\mathcal{P}(E)$ and noise models of all agents are common knowledge - an agent only knows the concrete realization of its own noisy Dec-POMDP i.e. $\{E^{A^i}\}_{1:n}$ are *not* common knowledge. The goal in noisy-ZSC is for agents A and B to coordinate zero-shot in the ground truth Dec-POMDP E^* at test-time.

We next present the reduction of noisy ZSC to standard ZSC. To simplify the presentation, we restrict our presentation to two agent setup (agents A and B) here and further assume that all the Dec-POMDPs under consideration have identical state space and horizon. In appendix B, we relax these assumptions.

3.2 Reduction to "Standard" ZSC

Standard ZSC assumes that all agents at test time act in a Dec-POMDP which is common knowledge to all agents during their independent training. We show that NZSC can be reduced to a standard ZSC problem by creating a common knowledge meta Dec-POMDP. This meta Dec-POMDP has an augmented state space and observation space, covering the space of Dec-POMDPs $\mathcal{E}$. Specifically, the aforementioned two-agent NZSC problem is equivalent to standard ZSC with both the agents acting in the following (common knowledge) meta-DecPOMDP $M = \langle \tilde{D}, \tilde{S}, \tilde{I}, \tilde{\mathcal{A}}, \{\tilde{O}^A, \tilde{T}, \tilde{O}^B\}, \{\tilde{\Omega}^A, \tilde{\Omega}^B\}, \tilde{R}, \tilde{h}\rangle$ where

- $\tilde{D} = \{A, B\}$
- $\tilde{S} = S \times E$ is the state space of the meta Dec-POMDP
- $\tilde{I} = I(S)P(E)$
- $\tilde{\mathcal{A}} = \mathcal{A}$
- $\tilde{T} : \tilde{S} \times \tilde{\mathcal{A}} \times \tilde{S} \to [0, 1]$ is the transition function where

$$\tilde{T}(\tilde{s}, a, \tilde{s}') = \tilde{T}[(s, E_i), a, (s', E_j)] = \begin{cases} T_i(s, a, s') & i = j \\ 0 & i \neq j \end{cases}$$

- $\tilde{\mathcal{O}}^A = \mathcal{O}^A \times E$ is the observation space for agent A. Observation space for agent B is defined similarly.
- $\tilde{\Omega}^A : \tilde{\mathcal{O}}^A \times \tilde{S} \to [0, 1]$ gives the observation probability function for agent A where

$$\begin{aligned} \tilde{\Omega}^A(\tilde{o}, \tilde{s}, a) &= \tilde{\Omega}^A((o, E_i), (s, E_j), a) \\ &= \begin{cases} [\Omega^A_{E_i}(o, s, a), E_A] & i = j \\ \text{undefined} & i \neq j \end{cases} \end{aligned}$$

  where E_A is the noisy DecPOMDP observed by agent A and $\Omega^A_{E_i}$ denotes the observation function of agent A in the Dec-POMDP $E_i \in \mathcal{E}$. The observation function for agent B, $\tilde{\Omega}^B$,

[1] See Section 2 for a more descriptive presentation of Dec-POMDP.

is defined similarly as follows

$$\tilde{\Omega}^B(\tilde{o}, \tilde{s}, a) = \begin{cases} [\Omega^B_{E_i}(o, s, a), E_B] & i = j \\ \text{undefined} & i \neq j \end{cases}$$

where E_B is the noisy DecPOMDP observed by agent B.

- $\tilde{R} : \tilde{S} \times A \times \tilde{S} \to (-\infty, \infty)$ is the reward function given by

$$\tilde{R}(\tilde{s}, a, \tilde{s}') = \tilde{R}[(s, E_i), a, (s', E_j)] = \begin{cases} R_i(s, a, s') & i = j \\ \text{undefined} & i \neq j \end{cases}$$

- $\tilde{h} = h$

Intuitively, this reduction works by creating a new problem setting that is common knowledge, where respective observations of noisy Dec-POMDPs E_A and E_B are modeled as the private knowledge of agents A and B and the ground truth Dec-POMDP is modeled as part of the (augmented) state space. Agents then keep a belief over what the true Dec-POMDP may be which they can refine as they interact with the ground truth Dec-POMDP.

The main benefit of this reduction is that this allows the reuse of reinforcement learning algorithms developed for standard ZSC problem *out of the box* for NZSC. We exploit this by using IPPO-OP, a standard ZSC algorithm (Hu et al., 2020), in our experiments.

4 Experimental Setup

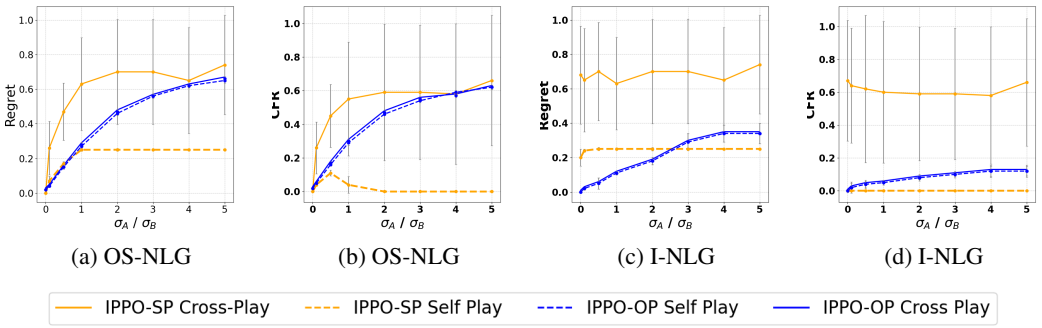
In this section, we focus on developing a better understanding of the challenges posed by the NZSC problem setting over the standard ZSC problem. To do so, we develop various environments for the NZSC problem and empirically analyze the policies learned by the agents in these environments. We note that standard ZSC benchmarks are based on popular games (and not reflective of possible real-world applications) and, hence, are not amenable to studying the NZSC problem. We briefly introduce our environments below, with further comprehensive details available in the Appendix C.

One-Shot Noisy Lever Game (OS-NLG): This is a simple one-step lever game consisting of three levers; with each lever i having an associated reward value R^*_i sampled from a predefined distribution $P(R_i)$ at the start of every round. However, in the noisy Dec-POMDP observed by agents A and B, the reward values are corrupted according to public noise models $P^A(R^A_i|R^*_i)$ and $P^A(R^A_i|R^*_i)$. The challenge is for the agents to coordinate on pulling the same lever i to earn reward R^*_i. Pulling different levers results in a penalty set to be some constant $c(-2$ in our experiments). By default, we set $P(R^*_i)$ to be the univariate normal distribution with mean $r_{mean} = 5$ and standard deviation $\sigma^* = 2$ and the noise models to be additive normal noise with mean 0 and standard deviations σ^A and σ^B respectively. Specifically, $P^A(R^A_i|R^*_i) = R^*_i + \epsilon^A_i; \epsilon^A_i \sim N(0, \sigma^A)$ with P^B defined similarly. This models a commonly found situation where two coordinating agents have the options to pursue multiple goals[2]; with the true payoffs of the goals being unknown and agents having different noisy observations of the payoffs.

Iterated Noise Lever Game (I-NLG): This is an extension of OS-NLG to a longer horizon - which is set to be 16 in our experiments. Identical to OS-NLG, agents get to observe noisy Dec-POMDPs at the first timestep. For all subsequent timesteps, agents get to observe the action the other agent took at the last timestep.

NoisyPartnerLeverGame (NP-LG): This environment models misspecification in the number of agents present in the game. Specifically, there are two levers l_h and l_b in this game; the rewards on both these levers are distributed normally according to $N(r_h, \sigma)$ and $N(r_b, \sigma)$ with $r_b < r_h$. To get the reward corresponding to lever l_l, at least one agent must pull on it. However, to earn the reward on lever l_h, both players must pull it. Further, the environment sets up to one agent to be 'inactive' with probability p. An inactive agent does not take any action in the environment. Further, an agent

[2]For example, a team of product designers trying to coordinate which new feature they should prioritize building. Or researchers within a team trying to coordinate on their next project given multiple options.



(a) OS-NLG (b) OS-NLG (c) I-NLG (d) I-NLG

Figure 3: Performance across different noise levels for IPPO-SP and IPPO-OP across OS-NLG and I-NLG. x-axis represents the noise level i.e $\sigma_A = \sigma_B$ and y-axis is either regret or coordination failure rate (CFR). Error bars represent 95% confidence interval and are not visible for some data points due to extremely low variance in the values.
IPPO-OP consistently performs better in the cross-play evaluation. Further, IPPO-OP achieves considerably better coordinating performance in I-NLG - highlighting the usefulness of the longer horizon for coordination.

receives a binary signal indicating whether the other agent is active or inactive in the environment - however, this signal is noisy and may be flipped with probability p_{noise}[3].

4.1 Metrics

In our results, we report performance across two key metrics: *regret* measures the difference in performance between the policy under study relative to the oracle ZSC policy that has access to the true ground truth Dec-POMDP E^*. We normalize regret in all our experiments, so, that the orcle ZSC policy (that has access to the ground truth Dec-POMDP) has a regret of 0. *Coordination-failure-rate* measures the fraction of samples at which policies fail to coordinate. Unless stated otherwise, these metrics are given for the cross-play in which ZSC performance between pairs of 8 independently trained policies is evaluated.

5 Results

5.1 Self-Play Is Not Robust To Noise

Our first experiment investigates how noise effects the policies discovered by IPPO-SP and IPPO-OP. In both OS-NLG and I-NLG, regardless of noise level, IPPO-OP agents learn the robust convention of playing *argmax* over the payoffs in noisy DecPOMPD they observe. However, while in the absence of noise, IPPO-SP does learn a similar convention of playing *argmax*; under noise, and in iterated setting in general, it learns the non-robust convention of choosing to coordinate on a lever with a specific label regardless of its observed corresponding payoffs. Consequently, we see in Figure 3, that while IPPO-OP does well under cross-play evaluation, however, IPPO-SP performs poorly under cross-play evaluation.

5.2 Longer Horizon Enables Better Coordination

Next, we consider I-NLG. Recall that in I-NLG, at the first timestep, an agent gets to observe noisy-Dec-POMDP and then at subsequent timesteps, it observes the action taken by the other agent at the previous timestep. As shown by Figures 3(b) and 3(c), IPPO-SP also fails under this condition and learns a label-based convention (even in the case of zero noise). So, on the whole, there is no marked

[3]This mimics a situation where two agents are playing a cooperative game online but they have unstable internet connections.

σ_A	σ_B	OS-NLG		I-NLG	
		Regret	Coordination Failure Rate	Regret	Coordination Failure Rate
0.0	3.0	0.42 (0.01)	0.44 (0.022)	0.19 (0.074)	0.10 (0.046)
0.0	5.0	0.51 (0.016)	0.53 (0.015)	0.24 (0.030)	0.12 (0.029)
0.0	7.0	0.56 (0.016)	0.56 (0.016)	0.27 (0.037)	0.10 (0.025)

Table 1: Cross play performance under asymmetric noise models. In OS-NLG, no communication is possible whatsoever so both agents have to act independently. In contrast, in the iterated scenario, a longer horizon allows the higher noise agent to observe, and then copy, the actions of its partner with low. or no noise; thus allowing it to also behave as well as the low noise agent despite having very high noise.

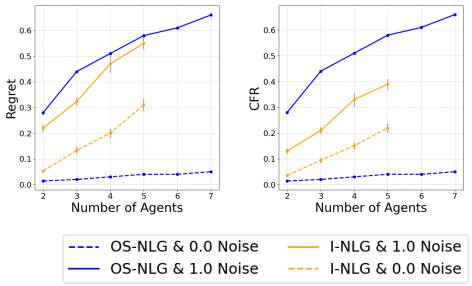


Figure 4: Regret and CFR for n-agent NLG.

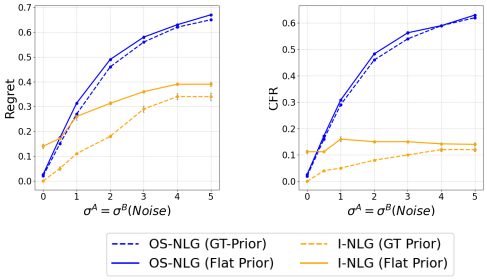


Figure 5: Robustness to using a flat prior in NZSC.

difference in the performance of IPPO-SP across OS-NLG and I-NLG. On the other hand, we see that IPPO-OP exploits the fact that an agent can observe the action of the other agent from the last timestep to coordinate better, resulting in an average 40% improvement in terms of regret over the OS-NLG.

In Appendix D.2, we further present results from an ablation showing that this improvement in coordination and performance is significantly helped by observing noisy DecPOMDP.

5.3 Asymmetric Noise Models

In NZSC, agents may have asymmetric information if their noise models differ. In the noisy lever game, if $\sigma_A \neq \sigma_B$, then agents will have asymmetric information. We specifically focus on the case where one agent has no noise and gets to observe the ground truth Dec-POMDP while the other agent only gets to observe noisy Dec-POMDP. In Table 1, we show that while in OS-NLG, one agent getting to observe ground truth Dec-POMDP yields only marginal gain in terms of performance, in I-NLG, it results in significant performance improvements. This indicates that even though the agent with high noise can not observe the ground truth Dec-POMDP, it can still do very well by *following* the agent with lower or zero noise.

5.4 NZSC among n Agents

We develop generalized n-agent versions of OS-NLG and I-NLG to understand how the number of participating agents affects coordination and general performance in NZSC. The results, presented in Figure 4 for two noise levels 0.0 and 1.0, give the cross-play performance for agents trained with IPPO-OP. For OS-NLG, in the case of zero noise (i.e. standard ZSC), agents are able to coordinate almost perfectly regardless of their number; however, for the NZSC (i.e. noise of 1.0), there is a marked degradation in coordination performance with the increase in the number of agents. For I-NLG, surprisingly, the degradation in performance with an increase in the number of agents is

Environment	Train Noise	Test Noise	Regret	CFR
OS-NLG	0.0	1.0	0.28(0.004)	0.29(0.004)
	1.0	1.0	0.28(0.006)	0.30(0.005)
I-NLG	0.0	1.0	0.21(0.017)	0.15(0.008)
	1.0	1.0	0.11(0.011)	0.05(0.005)
	0.0	2.0	0.36(0.034)	0.25(0.024)
	2.0	2.0	0.18(0.009)	0.08(0.013)
OS-NLG-β	0.0	1.0	0.81(0.004)	0.36(0.001)
	1.0	1.0	0.68(0.018)	0.24(0.011)

Table 2: While in OS-NLG, being aware of noise in problem specification does not provide a great advantage, it proves highly beneficial in I-NLG and in OS-NLG-β.

significantly less drastic. Note that for I-NLG, we only give results up to 5 agents as we failed to train policies for greater than 5 agents.

5.5 Misspecification In Number Of Agents

In NoisyPartnerLeverGame (NP-LG), we investigate the effects of misspecification in the number of *active* agents in the environment. NP-LG contains two levers: a high reward lever which gives agents a higher reward if *both* agents are active and pull this lever and a bail lever which gives a smaller reward if at least one agent pulls this lever. In Figure 6, we plot the proportion of an agent pulling the bail lever as a function of misspecification noise for different probabilities of an agent being absent. We show that misspecification noise causes agents (trained via IPPO-OP) to pull the bail lever considerably more often even though the underlying ground truth Dec-POMDP does not change characteristically; hinting at the fact that noise may cause agents to prefer *risk dominant strategies* over *payoff dominant strategies*.

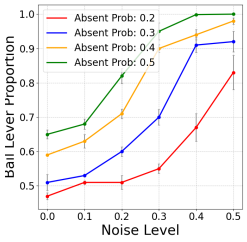


Figure 6: Proportion of IPPO-OP agents playing bail lever at different noise levels.

5.6 Usefulness Of Incorporating Knowledge of Misspecification

In this section, we evaluate the performance penalty agents may have to suffer if they ignore that the Dec-POMDP they are observing is misspecified. For this, we train one set of agents, termed *misspecification unaware* agents, under standard ZSC setup i.e. in a noise-free setting (where it is mistakenly assumed that the observed Dec-POMDP is (common knowledge) ground-truth Dec-POMDP). While the other set of agents, termed *misspecification aware* agents, are trained under NZSC setup. We then evaluate how well the two sets of agents perform when there is misspecification present in the environment.

In OS-NLG, both the misspecification aware (trained under NZSC with $\sigma^A = \sigma^B = 1.0$) and misspecification unaware agents (trained under ZSC with $\sigma^A = \sigma^B = 0.0$) perform well. However, in I-NLG, the misspecification-aware agents perform considerably better compared to misspecification-unaware agents. Finally, we consider a modified version of OS-NLG, denoted OS-NLG-β where rewards on the levers in the ground truth Dec-POMPD are distributed according to a bimodal beta distribution with $\alpha = \beta = 0.01$ and then scaled by 5 and shifted by 3 (we show this distribution in Figure 7). Everything else remains the same as OS-NLG. In OS-NLG-β, misspecification-aware agents perform better relative to misspecification-unaware agents. On the whole, it is better to be misspecification aware, especially in iterated games where repeated interactions present an opportunity to exchange information in a bid to minimize adverse effects of misspecification.

Environment	σ^A_{train}	σ^B_{train}	σ_{test}	Regret	Ref. Regret	CFR	Ref. CFR
OS-NLG	1.0	3.0	2.0	0.50(0.016)	0.46(0.016)	0.48(0.016)	0.47(0.014)
I-NLG	1.0	3.0	2.0	0.23(0.001)	0.18(0.009)	0.09(0.017)	0.08(0.013)
OS-NLG	1.0	2.0	3.0	0.59(0.012)	0.56(0.017)	0.56(0.011)	0.55(0.018)
I-NLG	1.0	2.0	3.0	0.38(0.010)	0.29(0.060)	0.18(0.012)	0.10(0.016)
OS-NLG	3.0	2.0	1.0	0.36(0.013)	0.27(0.014)	0.32(0.010)	0.30(0.015)
I-NLG	3.0	2.0	1.0	0.17(0.017)	0.10(0.014)	0.11(0.009)	0.05(0.008)

Table 3: NZSC provides reasonable robustness to misspecified and inconsistent noise models.

5.7 Is NZSC Robust To Misspecification?

NZSC assumes prior distribution over ground truth Dec-POMDP and the noise models to be common knowledge. However, it is possible that in practice these assumptions may be violated. We consider the impact of these violations on ZSC performance.

Robustness To Misspecified Prior: First, for OS-NLG and I-NLG, we consider the effect of all parties assuming *flat* or *uniform* prior over ground truth Dec-POMDPs and show the coordination performance of these agents when evaluated under true ground-truth distribution of Dec-POMDP (i.e. normal distribution) in Figure 5. We note that there is only a small degradation in performance, indicating that NZSC formulation is relatively robust against misspecification of the prior.

Robustness To Inconsistent Noise Models: Second, we evaluate the effects of inconsistent noise models in OS-NLG and I-NLG. Specifically, the noise model (for both agents) assumed by agent A differs from the noise model assumed by agent B and they both differ from the true noise model. We present our results in Table 3.

6 Discussion

In this work, we presented NZSC as a better model of real world zero-shot coordination problems. However, NZSC also presents a number of additional interesting challenges on the top of standard ZSC problem. Firstly, in the standard ZSC, the primary challenge is to disallow learning of non-universal conventions, but NZSC also adds an orthogonal challenge of handling uncertainty about the true problem. Via our experiments, we show that NZSC also exacerbates the challenge of learning non-universal conventions (Sec. 5.1). Further, as our results in Sec. 5.5 indicate, when a problem setting is misspecified, not being aware of misspecification can result in significantly worse coordination performance. Further, using empirical results, we have argued that NZSC problem setting is itself reasonably robust to misspecification. Our experiments with NoisyPartnerLeverGame further highlight the challenge of NZSC as it demonstrates that noise tend to cause agents to prefer risk-dominant strategies over payoff strategies. However, we note that our experiments also highlight that allowing agents to interact over longer horizons can be highly beneficial as agents learn to efficiently exchange and combine information about the problem setting to improve their performance.

Limitations: As our goal is to primarily introduce the problem setting of NZSC, we use an existing algorithm for learning ZSC policies in our experiments. Future works should consider developing improved reinforcement learning algorithms specific to the NZSC setup and also consider designing benchmarks specific to NZSC problem.

It is clear from our results that the presence of noise in problem specification can result in worse performance. Prior work in inverse reinforcement learning (Reddy et al., 2018) has shown that the sub-optimality of human policies in single-agent reinforcement learning can be explained by misspecification in problem setting, specifically the dynamics model. Future research should investigate to what extent the sub-optimality of human policies in multi-agent reinforcement learning scenarios can similarly be explained by misspecification in problem setting.

REFERENCES

Stefano V. Albrecht, Filippos Christianos, and Lukas Schäfer. *Multi-Agent Reinforcement Learning: Foundations and Modern Approaches*. MIT Press, 2023. URL https://www.marl-book.com.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Nolan Bard, Jakob N Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, et al. The hanabi challenge: A new frontier for ai research. *Artificial Intelligence*, 280:103216, 2020.

Samuel Barrett and Peter Stone. Cooperating with unknown teammates in complex domains: A robot soccer case study of ad hoc teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Samuel Barrett, Peter Stone, and Sarit Kraus. Empirical evaluation of ad hoc teamwork in the pursuit domain. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pp. 567–574, 2011.

Kalesha Bullard, Douwe Kiela, Franziska Meier, Joelle Pineau, and Jakob Foerster. Quasi-equivalence discovery for zero-shot emergent communication. *arXiv preprint arXiv:2103.08067*, 2021.

Jack Clark and Dario Amodei. Faulty reward functions in the wild. https://openai.com/research/faulty-reward-functions, 2016.

Miguel A Costa-Gomes and Vincent P Crawford. Cognition and behavior in two-person guessing games: An experimental study. *American economic review*, 96(5):1737–1768, 2006.

Brandon Cui, Hengyuan Hu, Luis Pineda, and Jakob Foerster. K-level reasoning for zero-shot coordination in hanabi. *Advances in Neural Information Processing Systems*, 34:8215–8228, 2021.

Brandon Cui, Andrei Lupu, Samuel Sokota, Hengyuan Hu, David J Wu, and Jakob Nicolaus Foerster. Adversarial diversity in hanabi. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=uLE3WF3-H_5.

Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R McKee, Joel Z Leibo, Kate Larson, and Thore Graepel. Open problems in cooperative ai. *arXiv preprint arXiv:2012.08630*, 2020.

Lauro Langosco Di Langosco, Jack Koch, Lee D Sharkey, Jacob Pfau, and David Krueger. Goal misgeneralization in deep reinforcement learning. In *International Conference on Machine Learning*, pp. 12004–12019. PMLR, 2022.

Nature Editorial. The cooperative human. *Nat. Hum. Behav*, 2:427–428, 2018.

Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

Sihong He, Songyang Han, Sanbao Su, Shuo Han, Shaofeng Zou, and Fei Miao. Robust multi-agent reinforcement learning with state uncertainty. *Transactions on Machine Learning Research*, 2023.

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*, 2021.

Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. "other-play" for zero-shot coordination. In *International Conference on Machine Learning*, pp. 4399–4410. PMLR, 2020.

Hengyuan Hu, Adam Lerer, Brandon Cui, Luis Pineda, Noam Brown, and Jakob Foerster. Off-belief learning. In *International Conference on Machine Learning*, pp. 4369–4379. PMLR, 2021.

Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. Specification gaming: The flip side of ai ingenuity. April 2020. URL https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity.

Shihui Li, Yi Wu, Xinyue Cui, Honghua Dong, Fei Fang, and Stuart Russell. Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 4213–4220, 2019.

Simin Li, Jun Guo, Jingqiao Xiu, Xini Yu, Jiakai Wang, Aishan Liu, Yaodong Yang, and Xianglong Liu. Byzantine robust cooperative multi-agent reinforcement learning as a bayesian game. *arXiv preprint arXiv:2305.12872*, 2023.

Xingzhou Lou, Jiaxian Guo, Junge Zhang, Jun Wang, Kaiqi Huang, and Yali Du. Pecan: Leveraging policy ensemble for context-aware zero-shot human-ai coordination. *arXiv preprint arXiv:2301.06387*, 2023.

Andrei Lupu, Brandon Cui, Hengyuan Hu, and Jakob Foerster. Trajectory diversity for zero-shot coordination. In *International conference on machine learning*, pp. 7204–7213. PMLR, 2021.

Alicia P Melis and Dirk Semmann. How is human cooperation different? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1553):2663–2674, 2010.

Reuth Mirsky, Ignacio Carlucho, Arrasy Rahman, Elliot Fosong, William Macke, Mohan Sridharan, Peter Stone, and Stefano V Albrecht. A survey of ad hoc teamwork research. In *Multi-Agent Systems: 19th European Conference, EUMAS 2022, Düsseldorf, Germany, September 14–16, 2022, Proceedings*, pp. 275–293. Springer, 2022.

Martin J Osborne and Ariel Rubinstein. *A course in game theory*. MIT press, 1994.

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. *arXiv preprint arXiv:2201.03544*, 2022.

Kishan Panaganti, Zaiyan Xu, Dileep Kalathil, and Mohammad Ghavamzadeh. Robust reinforcement learning using offline data. *Advances in neural information processing systems*, 35:32211–32224, 2022.

Elizabeth Pennisi. How did cooperative behavior evolve? *Science*, 309(5731):93–93, 2005.

Sid Reddy, Anca Dragan, and Sergey Levine. Where do you think you're going?: Inferring beliefs about dynamics from behavior. *Advances in Neural Information Processing Systems*, 31, 2018.

Aurko Roy, Huan Xu, and Sebastian Pokutta. Reinforcement learning under model mismatch. *Advances in neural information processing systems*, 30, 2017.

Rohin Shah, Dmitrii Krasheninnikov, Jordan Alexander, Pieter Abbeel, and Anca Dragan. Preferences implicit in the state of the world. *arXiv preprint arXiv:1902.04198*, 2019.

Macheng Shen and Jonathan P How. Robust opponent modeling via adversarial ensemble reinforcement learning in asymmetric imperfect-information games. *arXiv preprint arXiv:1909.08735*, 2019.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471, 2022.

John Maynard Smith and Eors Szathmary. *The major transitions in evolution*. OUP Oxford, 1997.

Peter Stone, Gal A. Kaminka, Sarit Kraus, and Jeffrey S. Rosenschein. Ad hoc autonomous agent teams: Collaboration without pre-coordination. In *Proceedings of the Twenty-Fourth Conference on Artificial Intelligence*, July 2010.

Chuangchuang Sun, Dong-Ki Kim, and Jonathan P How. Romax: Certifiably robust deep multiagent reinforcement learning via convex relaxation. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 5503–5510. IEEE, 2022.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 23–30. IEEE, 2017.

Johannes Treutlein, Michael Dennis, Caspar Oesterheld, and Jakob Foerster. A new formalism, method and open issues for zero-shot coordination. In *International Conference on Machine Learning*, pp. 10413–10423. PMLR, 2021.

Tessa van der Heiden, Christoph Salge, Efstratios Gavves, and Herke van Hoof. Robust multi-agent reinforcement learning with social empowerment for coordination and communication. *arXiv preprint arXiv:2012.08255*, 2020.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative, multi-agent games, 2022.

Rui Zhao, Jinming Song, Hu Haifeng, Yang Gao, Yi Wu, Zhongqian Sun, and Yang Wei. Maximum entropy population based training for zero-shot human-ai coordination. *arXiv preprint arXiv:2112.11701*, 2021.

Wenshuai Zhao, Jorge Peña Queralta, and Tomi Westerlund. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In *2020 IEEE symposium series on computational intelligence (SSCI)*, pp. 737–744. IEEE, 2020.

A Related Work

ZSC & Ad-Hoc Teamwork: The Zero-shot coordination (ZSC) problem was introduced by Hu et al. (2020) who then proposed other play as a solution to the ZSC problem. (Treutlein et al., 2021) provided a formal treatment of this setup as the *label-free coordination* problem and noted that in iterated games, other-play is not the optimal solution to the label-free coordination problem. Separately, Bullard et al. (2021) focus on identifying symmetries in a given Dec-POMDP to learn a coordination strategy that may be resilient to the arbitrary breaking of symmetries. Several works take inspiration from human learning and reasoning processes to propose methods for ZSC. Cui et al. (2021) use K-level reasoning (Costa-Gomes & Crawford, 2006) to train ZSC capable policies in Hanabi (Bard et al., 2020). *off-belief learning* (Hu et al., 2021) is another prominent method for ZSC which prevents agents from learning arbitrary conventions by replacing the real rewards seen during a transition, by reward from a fictitious trajectory sampled from a belief model $\mathcal{B}_{\pi_0}$ which is independent of the policy. A closely related setting to zero-shot coordination is ad-hoc teamwork (Mirsky et al., 2022; Barrett et al., 2011; Barrett & Stone, 2015). The goal in ad-hoc teamwork is to train an agent which does well when placed in a team of unknown agents (Stone et al., 2010). This ultimately amounts to learning the best response to a population of agents. However, training a diverse population remains an open challenging problem and has attracted considerable attention Lupu et al. (2021); Cui et al. (2023); Zhao et al. (2021); Lou et al. (2023). Different to all the above mentioned works, we consider the challenge of zero-shot coordination when the problem setting is not common knowledge.

Robustness To Misspecifciation In RL: Within single agent reinforcement learning, many works consider the misspecification in reward function (Pan et al., 2022; Skalse et al., 2022; Krakovna et al., 2020; Clark & Amodei, 2016) and propose algorithms to sidestep its effects (Shah et al., 2019; Hadfield-Menell et al., 2016). Robust reinforcement learning considers the misspecification in environment dynamics (Roy et al., 2017; Panaganti et al., 2022) with techniques such as domain randomization (Tobin et al., 2017; Zhao et al., 2020) having been proposed to counter this specification. While several works study robust MARL, the common forms of robustness studied are to adversarial partners (Shen & How, 2019; Li et al., 2023), state uncertainties (He et al., 2023) and partner agent's polices (Li et al., 2019; Sun et al., 2022; van der Heiden et al., 2020). However, in this work, we study the misspecification in the *problem setting* for different agents with the focus on ZSC.

B Formulation

In the main text, we presented the reduction for the special case of two agent setting and all Dec-POMDPs sharing the state space. Here, we show how to handle the general cases.

B.1 n-agents NZSC Reduction

The reduction for the two-agent setting can be extended to n agents straightforwardly by making following modifications:

- The set of agents $\tilde{D} = A_1, .., A_n$.
- $\tilde{\mathcal{O}}_{A_i} = \mathcal{O}_{A_i} \times E$ is the observation space for ith agent A_i.
- $\tilde{\Omega}_A : \tilde{\mathcal{O}}_A \times \tilde{S} \times A \to [0, 1]$ gives the observation probability function for agent A where
$$\tilde{\Omega}_A(\tilde{o}, \tilde{s}, a) = \tilde{\Omega}_A((o, E_i), (s, E_i), a) = [\Omega_A^{E_i}(o, s, a), E_A]$$
where E_A is the noisy DecPOMDP observed by agent A and $\Omega_A^{E_i}$ denotes the observation function of agent A in the Dec-POMDP $E_i \in \mathcal{E}$.

B.2 Different State Spaces

If not all Dec-POMDPs in $\mathcal{E}$ share the identical state space, then if the state space in the meta-DecPOMDP is naively set to the state space of ground truth Dec-POMDP, this may leak information about the ground truth Dec-POMDP. This is not permitted in NZSC. Thus, to prevent this the state space in the meta-dec POMDP is set to be the union of state spaces of all Dec-POMDPs in $\mathcal{E}$ i.e. $\tilde{S} = (\cup_{i \in I} \mathcal{S}_i) \times \mathcal{E}$ where $\mathcal{E}$ is the space of Dec-POMDPs and I is an index set over $\mathcal{E}$

C ENVIRONMENTS

C.1 ONE-SHOT NOISY LEVER GAME

This game involves two agents and three levers. The aim for both agents is to pull the same lever simultaneously, with each lever having an associated reward value sampled from a predefined distribution $P(R)$. The game only lasts one timestep every round. At the start of each round, reward values are resampled from $P(R)$. In the noiseless scenario, both agents get to observe the ground truth Dec-POMDP, including the exact reward values for each lever. However, in the noisy setting, agents can only observe a noisy Dec-POMDP. The noisy Dec-POMDP is identical to the ground truth Dec-POMDP, except for the reward function, which incorporates additive noise sampled from public noise models $P(R_A|R)$ and $P(R_B|R)$ for agents A and B, respectively.

By default, the reward value for each lever i is drawn from a univariate normal distribution with mean r_{mean} and standard deviation σ, i.e., $P(R_i) = N(r_{mean}, \sigma^2)$ in the ground truth Dec-POMDP. In the noisy Dec-POMDP, the noise model for agent A is represented by $o_i^A = R_i + \epsilon_i^A$, where $\epsilon_i^A \sim N(0, \sigma_A^2)$ independently for each i and σ_A controls the level of noise relative to the ground truth Dec-POMDP. Conversely, for agent B, the noise model is represented by $o_i^B = R_i + \epsilon_i^B$, where $\epsilon_i^B \sim N(0, \sigma_B^2)$ independently for each i. The non-coordinating penalty c is unchanged across both the ground truth Dec-POMDP and corresponding noisy Dec-POMDPs.

In our experiments, unless indicated otherwise, we set $r_{mean} = 5$ and $\sigma = 2$ and we vary σ_1 and σ_2 across experiments to control for noise level. Note that $\sigma_1 = \sigma_2 = 0$ corresponds to the noiseless setting.

C.1.1 FORMULATION AS DEC-POMDP AND META DEC-POMDP

The noisy lever game can be represented as a Dec-POMDP with a set of agents $\mathcal{D} = \{A, B\}$ and a state space $\mathcal{S} = \{1, 2, 3\}$ denoting the three levers. The action space is identical for the two agents, where each action $i \in \{1, 2, 3\}$ corresponds to pulling the lever with label i.

The reward function R is bivariate and dependent on the joint actions taken by both agents. It yields a reward value $r(a^A, a^B)$ if the actions of both agents are the same, and incurs a non-coordinating penalty c otherwise.

By leveraging the result from the previous section, the noisy lever game can be reformulated as a meta-Dec-POMDP by augmenting the state space with the reward values for each lever in the ground truth Dec-POMDP. This augmented state space is sufficient to uniquely identify any ground truth Dec-POMDP.

C.1.2 LABEL SYMMETRY IN THE NOISY LEVER GAME

Recall that state and action space in the noisy lever game consists of labels $\{1, 2, 3\}$ for identifying the levers. As all the levers in the noisy lever game have identical reward distribution, the *labels* on any two levers can be flipped without meaningfully altering the game. For example, we can flip the labels on the first and second lever without altering the Dec-POMDP in a meaningful way.

Lemma 1. *The following is a symmetry in the noisy lever game where ϕ_s, ϕ_a, ϕ_o are permutation maps of the state space, the action space, and the observation space respectively.*

$$\phi_s((R_1, R_2, R_3)) \doteq (R_2, R_1, R_3)$$
$$\phi_a(2) \doteq 1; \phi_a(1) \doteq 2; \phi_a(3) \doteq 3$$
$$\phi_o((o_1, o_2, o_3)) \doteq (o_2, o_1, o_3)$$

C.2 ITERATED NOISY LEVER GAME

Iterated noisy lever game (I-NLG) is the extension of OS-NLG to a longer horizon. While after one timestep, OS-NLG gets reset, I-NLG lasts 16 timesteps. Identical to OS-NLG, agents get to observe noisy Dec-POMDP at first timestep. For all subsequent timesteps, agents get to observe the action the other agent took at the last timestep. In some experiments, we also include some additional

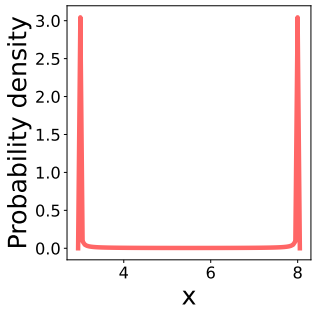


Figure 7: Beta distribution used in our experiments on OS-LG-β.

information regarding the reward received at the last timestep in the observation of the agent for timesteps $t \geq 1$.

C.3 OS-NLG-β

OS-NLG-β is identical to OS-NLG with two differences. The distribution of payoff in ground truth Dec-POMDP is set to be a bimodal beta distribution with parameters $\alpha = \beta = 0.01$. Further, the non-coordinating reward penalty is set to -10. The bimodal nature of beta distribution (shown in Figure 7) causes agents to learn a bimodal policy based on the most distinguishing lever. This policy performs better than a simple argmax policy under medium noise levels.

D ADDITIONAL RESULTS

D.1 CROSS PLAY MATRICES

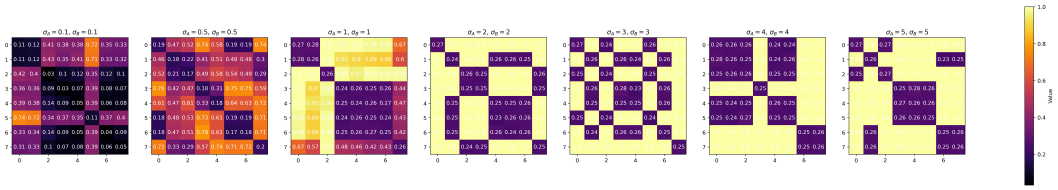


Figure 8: Cross play matrices showing *regret* for IPPO-SP on OS-NLG across various noise levels. The noise level is given in the title of each plot. Note that at higher noise levels, all agents learn a label-based convention which results in high regret when paired with an agent that has learnt a different convention. This is shown by bipolar nature of colors in the matrices.

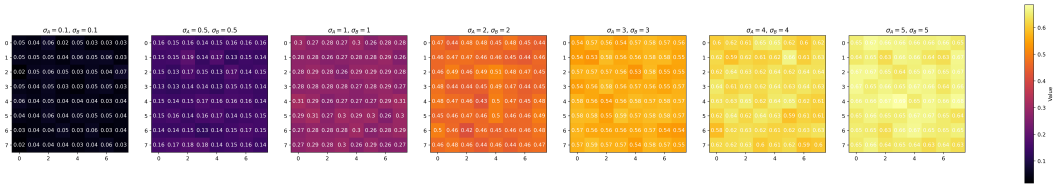


Figure 9: Cross play *regret* for IPPO-OP on OS-NLG. All the independently trained policies when paired together are able to coordinate effectively.

D.2 *Blind Agent* SETTING

Considering agents can act in the ground-truth Dec-POMDP and explore it, how beneficial is it to observe noisy-Dec-POMDP at the start of the round? To understand this, we run an experiment in which agents do not receive any signal whatsoever about the ground-truth Dec-POMDP. We call this setting *blind agent* setting. In Table 4, we compare the performance of IPPO-OP in the blind agent setting and in the standard noisy observation setting where agents observe the noisy-Dec-POMDP. To better understand the behavioural differences between policies learned at different noise levels, we additionally report *max lever rate*, *mid lever rate* and *min lever rate*. These metrics respectively measure how often agents are able to coordinate on the lever with maximum reward value (according to ground truth Dec-POMDP), lever with middle reward value and lever with minimum reward value.

However, the agents who observe noisy-Dec-POMDP with a high amount of noise only perform almost as well as the *blind* agents. The results in Figure 3 also show that there is graceful degradation in the performance of IPPO-OP relative to the noise in the noisy-Dec-POMDP observed by the agents.

	Regret	Coordination Failure Rate	Max Lever Rate	Mid Lever Rate	Min Lever Rate
Noisy Observation Setting					
$\sigma_A = \sigma_B = 0.0$	0.00 (0.001)	0.00 (0.00)	0.98 (0.010)	0.05 (0.009)	0.01 (0.000)
$\sigma_A = \sigma_B = 2.0$	0.18 (0.014)	0.08 (0.013)	0.59 (0.053)	0.23 (0.024)	0.07 (0.017)
$\sigma_A = \sigma_B = 5.0$	0.32 (0.049)	0.10 (0.015)	0.36 (0.063)	0.29 (0.018)	0.23 (0.044)
Blind Agent Setting					
Action Only	0.35 (0.074)	0.14 (0.067)	0.36 (0.07)	0.28 (0.028)	0.22 (0.059)

Table 4: Cross-play evaluation results for *blind agent* setting where agents do not receive any information about the ground-truth Dec-POMDP compared to when agents receive noisy observations about the ground truth Dec-POMDP. The value added by noisy observation of ground truth Dec-POMDP is modulated by *how noisy it is*. At low and medium noise levels, it is extremely useful, however, at high levels of noise e.g. $\sigma_A = \sigma_B = 5$, it has almost no utility. This can be seen by the fact that metrics for blind agent setting and $\sigma_A = \sigma_B = 5$ match very closely.