Minimizing Negative Side Effects in Cooperative Multi-Agent Systems Using Distributed Coordination

Extended Abstract

Moumita Choudhury
University of Massachusetts Amherst
Amherst, United States
amchoudhury@cs.umass.edu

Sandhya Saisubramanian
Oregon State University
Corvallis, United States
sandhya.sai@oregonstate.edu

Hao Zhang
University of Massachusetts Amherst
Amherst, United States
hao.zhang@umass.edu

Shlomo Zilberstein
University of Massachusetts Amherst
Amherst, United States
shlomo@cs.umass.edu

ABSTRACT

Autonomous agents in real-world environments may encounter undesirable outcomes or *negative side effects* (NSEs) when working collaboratively alongside other agents. We frame the challenge of minimizing NSEs in a multi-agent setting as a *lexicographic decentralized Markov decision process* in which we assume independence of rewards and transitions with respect to the primary assigned tasks, but allowing negative side effects to create a form of dependence among the agents. We present a lexicographic Q-learning approach to mitigate the NSEs using human feedback models while maintaining near-optimality with respect to the assigned tasks—up to some given slack. Our empirical evaluation across two domains demonstrates that our collaborative approach effectively mitigates NSEs, outperforming non-collaborative methods.

KEYWORDS

AI Safety; Negative Side Effects; Cooperative Multi-agent Systems; Distributed Constraint Optimization Problems

ACM Reference Format:
Moumita Choudhury, Sandhya Saisubramanian, Hao Zhang, and Shlomo Zilberstein. 2024. Minimizing Negative Side Effects in Cooperative Multi-Agent Systems Using Distributed Coordination: Extended Abstract. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 INTRODUCTION

Autonomous agents operating in the real world frequently generate undesired outcomes [8, 10, 14, 18] that are challenging to rectify during their training phase. Prior works have identified several categories of side effects, such as misspecification of rewards in reinforcement learning (RL) or goals in symbolic planning [2, 12, 13], distributional shift in the deployed environment [11], and reward gaming [4]. Lately, there has been a growing focus on scenarios in which an agent's actions directly influence the performance of other entities within the environment [1, 10]. Given the numerous settings where cooperative agents coexist and collaborate [5, 9], it becomes essential to investigate the occurrence of side effects in such multi-agent environments.

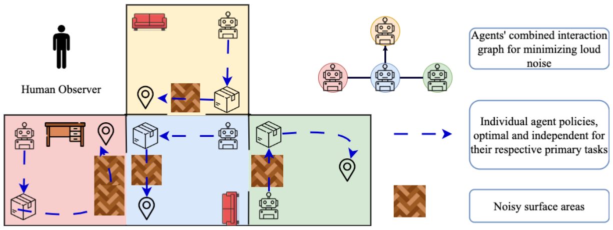
Human Observer
Agents' combined interaction graph for minimizing loud noise
Individual agent policies, optimal and independent for their respective primary tasks
Noisy surface areas

Figure 1: Illustration of multi-agent negative side effects (NSEs) in a boxpushing domain. Four robots collaborate by pushing boxes in a large space. Accomplishing the primary objective of each robot does not require coordination, but the surfaces in each room produce loud noises when adjacent agents push boxes simultaneously.

This work focuses on cooperative multi-agent settings, such as robot teams in warehouses or fleets of autonomous vehicles. While individual agents may excel in their primary tasks, their joint actions generate unforeseen side effects by leaving impacts on the agency of other agents or the environment itself, illustrated in Fig. 1. We address scenarios where isolated execution of agents' policies is side-effect-free, but their combined actions induce *negative side effects* (NSEs), initially unknown to the agents.

We adopt a lexicographic multi-objective approach by formulating the problem with a Lexicographic Markov Decision Process (LMDP) [16]. We present a combined approach integrating lexicographic multi-objective learning and coordinated Q-Learning to minimize the impacts of NSEs in a multi-agent environment. To the best of our knowledge, there is no existing solver for multi-agent lexicographic multi-objective problems. The occurrences of side effects are learned from human feedback, as the agents were initially unaware of the penalties associated with the NSEs.

Our primary contributions are fourfold: (1) formalizing the problem of multi-agent NSEs as a lexicographic Decentralized Markov Decision Process (DEC-MDP) [3] with local interaction, (2) defining a way to collect and generalize the joint penalty function from human feedback, (3) presenting a method for minimizing NSEs

This work is licensed under a Creative Commons Attribution International 4.0 License.

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

with a *coordinated lexicographic Q-learning* (C-LQL) solver, and (4) evaluating our approach and comparing it to non-coordinated and single-agent lexicographic Q-learning approaches.

2 FRAMEWORK FOR MINIMIZING NSE

Consider a cooperative multi-agent setting with n agents operating independently to complete their respective assigned tasks, which are their primary objectives, $O_1 = \{o_1, \ldots, o_n\}$. The agents operate based on a transition and reward independent DEC-MDP [3], M' that contains all the necessary information to complete their assigned tasks. However, the agents' models do not fully capture all the objectives in the complex real-world environment in which the agents operate. In this case, there is an additional secondary objective, O_2, that the agents need to minimize NSEs. The two objectives in M' are: primary assigned tasks (O_1) and mitigating side effects (O_2), where $O_1 \succ O_2$. Although the agents are transition and reward independent w.r.t. O_1, NSEs occur primarily because of their joint interaction.

We make the following assumptions: (1) executing the primary policy of each agent in isolation produces no negative side effects, but their joint policy $\pi' = \{\pi_1, \ldots \pi_n\}$ produces NSEs, unknown to the agents apriori, (2) the subset of agents interacting with each other to produce NSEs is much smaller than the total number of agents, (3) NSEs are undesirable but not catastrophic, and (4) NSEs result immediately from the joint execution in a state. Building on this, we define *multi-agent negative side effects* (MANSE), in which the occurrence and penalty for NSEs, denoted by R_N, depends on what actions agents perform jointly in a state. We assume a given interaction graph to facilitate the coordination between the agents.

Definition 2.1 Let $G = (X, E)$ be an interaction graph where each node $x_i \in X$ represents an agent i and each hyperlink $l \in E$ connects a subset of agents to form the reward component R_l. $F = \{f_1, \ldots, f_l\}$ denotes set the cost functions where f_l represents the cost function associated with each hyperlink l. Moreover, we define $\mathcal{F}_i$ to be the set of functions denoting which function nodes are connected to variable x_i, representing agent i. This hypergraph is formed to facilitate the interaction between agents to optimize the joint penalty where each hyperlink represents a subgroup of agents creating NSEs.

Definition 2.2 The joint penalty function, $R_N : S \times A \rightarrow \mathbf{R}$ for MANSE is divisible among subgroups of agents and can be expressed as $R_N(s, a) = \sum_l R_l(s_l, a_l)$ where $l = \{i_1, \ldots, i_k\}$ denotes a subgroup of size k. Moreover, $s_l = \langle s_{l_1}, \ldots, s_{l_k} \rangle$ denotes the state of group l and $a_l = \langle a_{l_1}, \ldots, a_{l_k} \rangle$ denotes the action of group l.

Definition 2.3 The augmented MDP for a given MANSE problem is a lexicographic DEC-MDP (LDEC-MDP), which is a multi-agent extension of LMDP [16], denoted as $\tilde{M} = \langle \tilde{S}, \tilde{A}, \tilde{P}, \tilde{\mathbf{R}}, o, \tilde{\mathbf{\Delta}} \rangle$. $\tilde{M}$ is a DEC-MDP with two reward functions $\tilde{R} = \{R_1, R_N\}$ where R_1 is the independent reward associated with the primary objective and R_N is the joint reward associated with NSE of joint actions. R_N follows the decomposition described in Defn. 2.2. Moreover, $O = \{O_1, O_2\}$ is the ordering of the objectives where $O_1 = \{o_1, \ldots, o_n\}$ is the primary objectives associated with the agents' independent assigned tasks described by reward R_1. Here, o_i represents the primary objective for agent i. O_2 denotes the objective to minimize NSEs and $O_1 \succ O_2$. $\tilde{\mathbf{\Delta}}$ refers to the collection of $\mathbf{\Delta}$ for each agent.

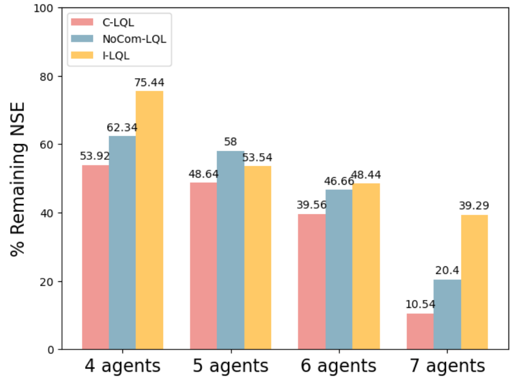

Figure 2: Minimizing Negative Side Effects across different problem sizes in the boxpushing domain

In order to reduce negative side effects in a multi-agent setting we have to to solve the corresponding LDEC-MDP. Our complete solution framework for minimizing NSEs involves the following two steps: (1) gathering information about NSEs using human feedback and generalizing them to unseen situations; (2) using a coordinated learning approach to solve the augmented LDEC-MDP. In the first step, the oracle, typically representing human feedback, provides signals about undesirable actions, which is later generalized by the simulator. In the second step, the agents learn to minimize NSEs jointly with the penalty they receive from the simulator using Coordinated Lexicographic Q-learning (C-LQL). We use a combination of two approaches: (1) a lexicographic Q-learning solver for LMDP [15], and (2) a Coordinated Q-learning (CQL) approach that uses a Distributed Constraint Optimization Problem (DCOP) [6, 7] solver to acquire joint Q-values for NSE minimization [17].

3 RESULTS AND CONCLUSION

We compare three baselines: (1) Independent Lexicographic Q-learning *(I-LQL)* approach [13] which is a model-free modification of the model-based LMDP solver (2) No Communication Lexicographic Q-learning *(NoCom-LQL)* approach where the agents learn individual Q functions by dividing the joint penalty equally among each member of group, and (3) *Prior* denotes the amount of side effects before minimizing NSEs. We explore the scalability of our approach by varying the problem size in number of agents and density. Figure 2 shows the performance in different problem sizes in a modified boxpushing problem [13]. In all the problem setups, *C-LQL* performs better than the other two approaches with the best result of 90% reduction in NSEs in the 7 agent setup.

We analyze the performance of using lexicographic Q-learning with and without coordination. Our analysis shows that *C-LQL* minimizes NSEs in different problem settings better than the uncoordinated version, without using much slack. In future work, we aim to extend our approach to fit a more general class of multi-agent problems where the side effects are generated by dynamic interactions among agents.

ACKNOWLEDGMENTS

This work was supported in part by NSF grants number 1954782 and 2326054, and by ONR grant number N00014-23-1-2171.

REFERENCES

[1] Parand Alizadeh Alamdari, Toryn Q Klassen, Rodrigo Toro Icarte, and Sheila A McIlraith. 2021. Avoiding negative side effects by considering Others. In *Safe and Robust Control of Uncertain Systems Workshop at NeurIPS*.

[2] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565* (2016).

[3] Raphen Becker, Shlomo Zilberstein, Victor Lesser, and Claudia V Goldman. 2003. Transition-independent decentralized Markov decision processes. In *Proceedings of the Second International Joint Conference on Autonomous agents and Multiagent systems*. 41–48.

[4] Jack Clark and Dario Amodei. 2016. Faulty reward functions in the wild. *Internet: https://blog. openai. com/faulty-reward-functions* (2016).

[5] Raffaello D'Andrea. 2012. A revolution in the warehouse: A retrospective on kiva systems and the grand challenges ahead. *IEEE Transactions on Automation Science and Engineering* 9, 4 (2012), 638–639.

[6] Alessandro Farinelli, Alex Rogers, and Nick R Jennings. 2014. Agent-based decentralised coordination for sensor networks using the max-sum algorithm. *Autonomous Agents and Multi-agent Systems* 28 (2014), 337–380.

[7] Ferdinando Fioretto, Enrico Pontelli, and William Yeoh. 2018. Distributed constraint optimization problems and applications: A survey. *Journal of Artificial Intelligence Research* 61 (2018), 623–698.

[8] Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart J Russell, and Anca Dragan. 2017. Inverse reward design. *Advances in neural information processing systems* 30 (2017).

[9] Ryan Hoque, Lawrence Yunliang Chen, Satvik Sharma, Karthik Dharmarajan, Brijen Thananjeyan, Pieter Abbeel, and Ken Goldberg. 2023. Fleet-dagger: Interactive robot fleet learning with scalable human supervision. In *Conference on Robot Learning*. PMLR, 368–380.

[10] Victoria Krakovna, Laurent Orseau, Ramana Kumar, Miljan Martic, and Shane Legg. 2019. Penalizing side effects using stepwise relative reachability. In *IJCAI AI Safety Workshop*.

[11] Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. 2008. *Dataset shift in machine learning*. Mit Press.

[12] Ramya Ramakrishnan, Ece Kamar, Besmira Nushi, Debadeepta Dey, Julie Shah, and Eric Horvitz. 2019. Overcoming blind spots in the real world: Leveraging complementary abilities for joint execution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 6137–6145.

[13] Sandhya Saisubramanian, Ece Kamar, and Shlomo Zilberstein. 2020. A multi-objective approach to mitigate negative side effects. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, Yokohama, Japan, 354–361.

[14] Sandhya Saisubramanian, Shlomo Zilberstein, and Ece Kamar. 2022. Avoiding Negative Side Effects due to Incomplete Knowledge of AI Systems. *AI Magazine* 42, 4 (2022), 62–71.

[15] Joar Skalse, Lewis Hammond, Charlie Griffin, and Alessandro Abate. 2022. Lexicographic multi-objective reinforcement learning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)*.

[16] Kyle Wray, Shlomo Zilberstein, and Abdel-Illah Mouaddib. 2015. Multi-objective MDPs with conditional lexicographic reward preferences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 29.

[17] Chongjie Zhang and Victor Lesser. 2011. Coordinated multi-agent reinforcement learning in networked distributed POMDPs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 25. 764–770.

[18] Shun Zhang, Edmund H Durfee, and Satinder Singh. 2018. Minimax-regret querying on side effects for safe optimality in factored markov decision processes.. In *Proceedings of the Twenty-sixth International Joint Conferences on Artificial Intelligence*. 4867–4873.