

# Single-Agent Planning in a Multi-Agent System: A Unified Framework for Type-Based Planners

Fengming Zhu

The Hong Kong University of Science and Technology  
Hong Kong SAR, China  
fzhu@connect.ust.hk

Fangzhen Lin

The Hong Kong University of Science and Technology  
Hong Kong SAR, China  
flin@cs.ust.hk

## ABSTRACT

We consider a general problem where an agent is in a multi-agent environment and must plan for herself without any prior information about her opponents. At each moment, this pivotal agent is faced with a trade-off between exploiting her currently accumulated information about the other agents and exploring further to improve future (re-)planning. We propose a theoretic framework that unifies a spectrum of planners for the pivotal agent to address this trade-off. The planner at one end of this spectrum aims to find exact solutions, while those towards the other end yield approximate solutions as the problem scales up. Beyond theoretical analysis, we also implement 13 planners and conduct experiments in a specific domain called *multi-agent route planning* with the number of agents **up to 50**, to compare their performances in various scenarios. One interesting observation comes from a class of planners that we call *safe-agents* and their enhanced variants by incorporating domain-specific knowledge, which is a simple special case under the proposed general framework, but performs sufficiently well in most cases. Our unified framework, as well as those induced planners, provides new insights on multi-agent decision-making, with potential applications to related areas such as mechanism design.

## KEYWORDS

Multi-Agent Planning; Opponent Modelling; Tree Search

### ACM Reference Format:

Fengming Zhu and Fangzhen Lin. 2025. Single-Agent Planning in a Multi-Agent System: A Unified Framework for Type-Based Planners. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025)*, Detroit, Michigan, USA, May 19 – 23, 2025, IFAAMAS, 10 pages.

## 1 INTRODUCTION

We consider a general problem where an agent is in a multi-agent system and is supposed to plan for herself without any prior information about her opponents. A motivating example of such a scenario is in autonomous driving [15], where a fully autonomous vehicle needs to navigate herself to the destination, but inevitably shares the roads with other cars that she may not know much about, ones either driven by human beings or possibly controlled by AI softwares from another company. Therefore, she must learn the behavioral patterns of the other cars in real-time and act accordingly.

There are three unique features that distinguish this setting from other seemingly related ones. **Firstly**, it is clearly different from a system where all the agents are under the control of a central planner, such as programming robot fleets to deliver packages in warehouses [30, 47, 55, 56], which is also known as *multi-agent path finding* (MAPF) problems. **Secondly**, the setting that we are investigating does not ensure pure competitiveness among participants, and therefore, cannot be modelled as zero-sum games (win or lose). In typical zero-sum games, e.g., Go/chess [5, 17, 42, 48] and poker [64], a player can assume the other one will do the most harm, and compute a minimax strategy, which coincides with the Nash equilibrium (NE) and later serves as a never-lose strategy regardless of what strategy the opponent eventually follows. **Thirdly**, an oracle that computes a sample NE does not *directly* offer much help, even if all agents are aligned with common interests [10, 16]. Chances are that there may be multiple NEs or the other agents are simply not playing the equilibrium strategies at all.

Conceptually, this general class of problems is usually addressed by conducting opponent modelling and opponent-aware replanning at the same time. Given no knowledge about the upcoming opponents, the controlled agent will start with an initial belief over possible opponent models (or types). With time going on, actions done by the other players are observed, which can be utilized to update the belief. Meanwhile, the revised belief can be employed for replanning. Essentially, to control the agent is to find a trade-off between exploiting current knowledge and acquiring more information about opponents to enhance future replanning [13, 14].

In principle, it can be formalized as a stochastic game [46, 52] with incomplete information [23], and the incompleteness is due to the lack of knowledge about opponents. There is a vast literature on how to devise an effective strategy for the controlled agent, mostly based on best responses [2, 10, 13, 14, 16, 18, 26, 37, 44, 45]. However, one can clearly observe two extremes along this line of work. At one extreme, researchers apply deliberate formulations on toy artificial problems such as repeated games with one state transiting to itself [2, 10, 13, 14, 16, 26], serving primarily as proof-of-concept examples. In fact, repeated games cannot fully reflect the complexity underneath as they are just special cases of stochastic games. At the other extreme, people jump to highly approximated ones, namely Monte Carlo methods [2, 18, 44, 45] or complex Neural Network (NN) pipelines [37], yet the largest known domain remains only  $10^{14}$  states with four agents [37, 44, 45]. We here ask a significant research question: **Is there any viable formulation in between, and even beyond?** More precisely, given such a multi-agent planning problem with  $10^6$  states ( $\ll 10^{14}$ ), can we identify a planner that is more accurate than Monte Carlo ones, and is there any planner that can scale to even larger state spaces ( $\gg 10^{14}$ )?



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

*Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025)*, Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

To this end, we first present a set of formulations that each of them *rediscovers and generalizes* an existing one from the literature. More specifically, to find an optimal plan, one has to compute an exact solution for the underlying infinite-horizon *partially observable Markov decision process* (POMDP) [25, 51, 53]. This is only feasible for tiny problem instances, as optimally solving infinite-horizon POMDPs is computationally expensive [34, 35]. As the environment scales up in terms of the number of agents, we have to appeal to approximated formulations. We then propose a general framework that unifies this entire set of formulations. This unified framework is designed to encompass a spectrum that interpolates the aforementioned formulations, enabling users to devise new ones for those problems of intermediate scales as well as much larger scales. **The theoretic framework is basically to perform layered lookahead search and backup from values of leaf nodes in each layer, wherein different formulations implement different lookahead search.** The induced formulation at one end of this spectrum aims at solving exact solutions, while the formulations towards the other extreme lead to approximated ones as the problem scales up. Within our proposed framework, approximations can be made by controlling the depth of tree search, altering node evaluation functions, adopting sampling-based backup, toggling heuristic action selection, etc. The framework effectively connects POMDP [25], *contextual-MDP/RL* [9, 22], *tree search* [12] and even *constraint satisfaction* [11].

It is highly non-trivial to compare those formulations purely in theory. Therefore, we implement a group of planners resulted from those formulations in a concrete domain, called *multi-agent route planning* (MARP). This benchmark is challenging in two aspects: 1) it is no longer repeated games, instead, involves complex state transitions; 2) it necessitates long-horizon planning, as the rewards are goal-conditioned and hence sparse. We select this domain on purpose to take the advantage that it can generate problem instances (parameterized by the map size and the number of agents) along a fine-grained axis of scales, from around 70 states with two agents to around  $4.62 \times 10^{145}$  states with 50 agents, matching our need of studying the capacities of the planners in the spectrum. As it is impossible to run each proposed planner against all possible hypothetical opponents and their combinations, we instead test our planners against three representative groups of opponents, including rational ones, malicious ones, and self-playing ones. Among all the planners, we also draw one's attention to a particular class, what we call *safe-agents* and their enhanced variants, which turns out to be a special case of our general framework but performs sufficiently well in most cases. It is also simple enough to save a great amount of computation.

The full paper is available at <https://arxiv.org/abs/2502.08950>.

## 2 RELATED WORK

The problem that we investigate pertains to a board range of theoretic areas, primarily *opponent modelling* and the control theory of POMDP, alongside practical tools, such as efficient *NE solvers*.

**Opponent Modelling.** Albrecht and Stone [4] surveyed a number of methodologies under this topic, where our work falls into the category of type-based reasoning. Types usually refer to pre-defined opponent models. For instance, Carmel and Markovitch

[13, 14] investigate a similar problem by modelling opponent strategies as deterministic finite automata, while other researchers have approached this by using reactive policy oracles, either given by human experts [2] or trained by separated pipelines [44, 45]. An critical follow-up issue is how to integrate opponent modelling with opponent-aware planning. For repeated games, much of the literature emphasizes planners that best respond to the current belief over hypothetical opponent models, e.g.,  $\epsilon$ -BR [26], IL/JAL [10, 16], CJAL [6], etc. For a more general setting, HBA [2] is proposed as a belief-dependent planning operator that takes long-term returns into account. The authors have also examined some convergence results [3], and the conditions under which the belief correctly reveals the truth [1]. Recent studies have also explored MCTS methods [44, 45], and neural network approximations [37].

**POMDP.** We later formalize the underlying problem as a *Contextual MDP* (CMDP) [22], where the set of contexts corresponds to the set of all possible opponent models. However, CMDPs reduce to POMDPs [25, 51, 53] when those contexts are not revealed. One can therefore resort to existing techniques in the literature of the POMDP theory to alleviate the computational burden [31, 61]. Moreover, some meta/contextual learning techniques developed by the RL community may also be related [7, 9].

**NE Solvers.** Lastly, we want to clarify that our work may *not directly* align with the interests of researchers who are investigating the problem of *Multi-Agent Reinforcement Learning* (MARL) [19, 20, 32, 41, 58, 62] or *Multi-Agent Path Finding* (MAPF) [30, 47, 55, 56, 60]. Both lines of work focus on solving a sample NE in a centralized manner, which does not strictly fall into the scope of our discussion as they force all agents to execute the NE plan that is found upfront. Nevertheless, as we will show that with the help of an oracle that computes an NE very fast in real-time, our proposed framework can effectively repair it while replanning. We emphasize that this routing domain resembling MAPF is selected as our benchmark solely for the purpose of evaluating and illustrating the capabilities of our proposed planners.

## 3 MODEL AND SOLUTION CONCEPTS

### 3.1 Stochastic Game

The whole system where the agents interact is modelled as a *stochastic game* (SG, also known as Markov games) [46, 52], which can be seen as an extension of both *normal-form games* (to dynamic situations with stochastic transitions) and *Markov decision processes* (to strategic situations with multiple agents). A stochastic game is a 5-tuple  $\langle \mathcal{N}, \mathcal{S}, \mathcal{A}, T, R \rangle$  given as follows,

- (1)  $\mathcal{N}$  is a finite set of agents.
- (2)  $\mathcal{S}$  is a finite set of normal form games. We call each possible  $S \in \mathcal{S}$  a stage game. In the proceeding discussion, each  $S$  is also called an environment state, or just a state.
- (3)  $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$  is a set of joint actions, where  $\mathcal{A}_i$  is the action set of agent  $i$ . We write  $a_i$  as the action of agent  $i$  and  $a = (a_i, a_{-i})$  (without any subscript) as the joint action.
- (4)  $T : \mathcal{S} \times \mathcal{A}_1 \times \cdots \times \mathcal{A}_n \mapsto \Delta(\mathcal{S})$  defines a stochastic transition among stage games.
- (5)  $R_i : \mathcal{S} \times \mathcal{A}_1 \times \cdots \times \mathcal{A}_n \mapsto \mathbb{R}$  denote the utility function of agent  $i$  in each stage game.

Each agent  $i$  is only aware of her own utilities  $R_i$ . In fact, knowing other agents' utility functions does not help, as she has no knowledge about the opponents until the game begins and will be possibly exposed to different opponents in each match. This adds extra difficulty of evaluation as we will cover that later in Section 5. Following the terminology in [4], we also call agent  $i$  the modelling agent under our control, and agent  $-i$  the modelled agents.

### 3.2 The Modelling Agent's Problem

The modelling agent under such an environment acts in a manner of maximizing her own accumulated discounted utility, i.e.,  $\mathbb{E}_{\tau \sim SG}[\sum_{t=0}^{T(\tau)} \gamma^t R_{i,t}]$ , where  $\tau$  denotes the trajectory of one single match of the SG,  $\gamma < 1$  denotes the discount factor, and  $T(\tau)$  denotes the length of the trajectory. For the sake of simplicity, we assume perfect recall, i.e., the modelling agent can remember the entire history from the beginning of the match up to the current moment, including every past states and joint actions. Formally, a history segment is given as  $h \in \mathcal{H} \triangleq (\mathcal{S} \times \mathcal{A})^*$ , i.e. a sequence consisting of alternating states and joint-actions.

Therefore, a candidate solution concept for the modelling agent is a history-dependent stochastic policy  $\pi_i : \mathcal{H} \times \mathcal{S} \mapsto \Delta(\mathcal{A}_i)$  that maps from all possible history segments and current states to (randomized) available actions. From agent  $i$ 's perspective, she also assumes every other opponent  $j$ 's potential strategy is drawn from a set of basis policies  $\Pi_j$ , where each  $\pi_j^k \in \Pi_j$  is a stationary Markov policy that maps states to actions, i.e.  $\pi_j^k : \mathcal{S} \mapsto \Delta(\mathcal{A}_j)$ . This is indeed a strong assumption that suggests those opponents do not care about histories, and they are not learning and evolving during the game play.

We then show, from the modelling agent's perspective, conditioned on the strategies of opponents, one can model the rest of the problem as a *Contextual Markov Decision Process* (CMDP) [22]. The resulted formulation is given as a 3-tuple  $\langle \mathcal{C}, \mathcal{A}, \mathcal{M} \rangle$ ,

- (1)  $\mathcal{C}$  is a set of contexts. In our setting,  $\mathcal{C} \triangleq \Pi_{-i} = \prod_{j \neq i} \Pi_j$ , i.e., the set of contexts are exactly the set of all possible combinations of opponent models.
- (2)  $\mathcal{A}$  is a set of actions. In our setting,  $\mathcal{A} \triangleq \mathcal{A}_i$ , as we are approaching the problem from agent  $i$ 's perspective.
- (3)  $\mathcal{M}$  is a mapping from contexts to corresponding Markov decision processes. An interpretation is, if the modelling agent knows for sure how her opponents would act, then she is basically faced with a single-agent MDP with opponent behaviors subsumed into transition/utility functions as noises. The resulted MDP from the perspective of agent  $i$ , denoted as  $\mathcal{M}(\pi_{-i})$ , is induced as a 5-tuple  $\langle \mathcal{S}, \mathcal{A}_i, T^{\pi_{-i}}, R^{\pi_{-i}}, \gamma \rangle$ :
  - $\mathcal{S}, \mathcal{A}_i$  and  $\gamma$  inherit from the previous setup,
  - $T^{\pi_{-i}}(S'|S, a) \triangleq \sum_{a_{-i} \in \mathcal{A}_{-i}} T(S'|S, a)\pi_{-i}(a_{-i}|S)$ ,
  - $R^{\pi_{-i}}(S, a) \triangleq \sum_{a_{-i} \in \mathcal{A}_{-i}} R_i(S, a)\pi_{-i}(a_{-i}|S)$ ,

If the context is revealed, it is equivalent to augmenting the MDP state by an additional dimension representing the context, and such a CMDP can therefore be factorized into  $|\mathcal{C}|$  MDPs and solved one by one via any off-the-shelf solver accordingly. However, the issue is that in our situation the context is unobservable as it represents the unknown strategies of opponents. It then reduces to solving the corresponding POMDP [25, 51, 53], where a state in the POMDP

consists of the environment state and the underlying context (opponent model), and the observation function takes as input a POMDP state, and returns only the environment state exposed to everyone yet keeps the context (opponent model) hidden.

We also assume that the pivotal agent model every other opponent *independently*, and each independent belief is a distribution over potential basis policies. Mathematically, for any belief  $b \in \mathcal{B} \triangleq \prod_{j \neq i} \Delta(\Pi_j)$ , we have  $\pi_{-i}(a_{-i}|s) = \prod_{j \neq i} \pi_j(a_j|s)$  and  $b(\pi_{-i}) = \prod_{j \neq i} b(\pi_j)$ . Depending on the set of presumed opponent policies, any belief  $b_t$  at time  $t$  can be updated to a successor belief  $b_{t+1}$  inductively by a revision operator, denoted as  $\xi : \mathcal{B} \times \mathcal{S} \times \mathcal{A} \mapsto \mathcal{B}$ . Equivalently speaking, given a prior belief and a sequence consisting of past states, and the actions taken under each corresponding state, a posterior belief can be inferred. Here we formulate it in a Bayesian way,  $b_{t+1} = \xi(b_t, S, a)$  is given as,

$$\begin{aligned} b_{t+1}(\pi_j^k) &\triangleq \mathbb{P}[\pi_j^k | S, a_j] = \frac{(\mathbb{P}[a_j | S, \pi_j^k] \cdot \mathbb{P}[\pi_j^k | S])^{1/\beta}}{\sum_{\pi_j^l \in \Pi_j} (\mathbb{P}[a_j | S, \pi_j^l] \cdot \mathbb{P}[\pi_j^l | S])^{1/\beta}} \\ &= \frac{(\pi_j^k(a_j|S) \cdot b_t(\pi_j^k))^{1/\beta}}{\sum_{\pi_j^l \in \Pi_j} (\pi_j^l(a_j|S) \cdot b_t(\pi_j^l))^{1/\beta}} \end{aligned} \quad (1)$$

where  $\beta$  is a tunable temperature parameter. If  $\beta = 1$  is always the case, it will be exactly the same as how beliefs are computed in the corresponding POMDP. However, altering the value of  $\beta$  will open up a broader set of choices for implementations, as we will illustrate in Table 1.

### 3.3 Potential Solution Formulations

**3.3.1 Exact Dynamic Programming.** As we mentioned, the modelling agent is faced with a CMDP with unobservable contexts which subsequently reduces to a POMDP. Following from the control theory of POMDP [51], one can actually find out that at each moment a compound *information state* consisting of the environment state and the belief serves as a sufficient statistic, summarizing all the past history. Hence, the transitions among those information states from the modelling agent's perspective, denoted as  $\mathcal{T} : \mathcal{S} \times \mathcal{B} \times \mathcal{A}_i \mapsto \Delta(\mathcal{S}) \times \mathcal{B}$ , are resulted from a joint effect of the environment transition and a sampling process of opponents' actions according to the belief. Mathematically,

$$\mathcal{T}((S', b') | (S, b), a_i) \triangleq \sum_{\pi_{-i} \in \mathcal{B}} b(\pi_{-i}) \sum_{a_{-i} \in \mathcal{A}_{-i}} T(S'|S, a)\pi_{-i}(a_{-i}|S)$$

when  $T(S'|S, a) > 0$  and  $b' = \xi(b, S, a)$ , and all the other transitions are invalid and assigned zero probability. The expected reward from the modelling agent's perspective is induced analogously,

$$\mathcal{R}((S, b), a_i) \triangleq \sum_{\pi_{-i} \in \mathcal{B}} b(\pi_{-i}) \sum_{a_{-i} \in \mathcal{A}_{-i}} R_i(S, a)\pi_{-i}(a_{-i}|S)$$

Then, the Bellman optimality equation can be established over a continuous space,

$$\begin{aligned} &V_i(S, b) \\ &= \max_{a_i \in \mathcal{A}_i} \left\{ \mathcal{R}((S, b), a_i) + \gamma \sum_{S', b'} \mathcal{T}((S', b') | (S, b), a_i) \cdot V_i(S', b') \right\} \\ &= \max_{a_i \in \mathcal{A}_i} \left\{ \mathbb{E}_{\substack{\pi_{-i} \sim b \\ a_{-i} \sim \pi_{-i}}} [R_i(S, a) + \gamma \sum_{S'} T(S'|S, a)V_i(S', b')] \right\} \end{aligned} \quad (2)$$

when  $b' = \xi(b, S, a)$ , and  $V_i(\cdot, \cdot)$  is the desired optimal value function for the modelling agent. Again, if  $\xi$  is implemented with  $\beta = 1$ ,

then solving Equation (2) is equivalent to optimally solving the corresponding infinite-horizon POMDP. Despite that it is computationally costly in theory as revealed by [34, 35], we can resort to off-the-shelf POMDP solvers like *pomdp-solve*<sup>1</sup> to solve small problem instances in practice, as we will show in Appendix D.

As one may notice, this formulation resembles the HBA operator proposed by [2], where the authors thereof call it a best-response rule based on Bellman control. We here point out that it is not just any random rule, but also the exact characterization of what the modelling agent is faced with, and the optimal control strategy can be therefore *derived* and computed upfront.

In fact, Equation (2) also leads to potential (contextual-)RL solutions [9]. Subsequently, the most challenging part lies in how to effectively train such a policy that converges to the desired optimum. As we will show in Appendix C, there is usually an intermediate plateau phase before convergence, which we conjecture is caused due to the gap between feasible strategies and optimal strategies.

**3.3.2 Belief-Induced MDP.** As an alternative, one can amortize the burden of computing an optimal strategy completely in advance over repeated online replanning during the game play. We first extend the aforementioned context-to-MDP mapping  $\mathcal{M}(\pi_{-i})$  to  $\mathcal{M}(b)$ , allowing inputs of distributions of contexts,

- $T^b(S'|S, a_i) \triangleq \sum_{\pi_{-i} \in b} b(\pi_{-i}) \sum_{a_{-i} \in \mathcal{A}_{-i}} T(S'|S, a) \pi_{-i}(a_{-i}|S)$
- $R^b(S, a_i) \triangleq \sum_{\pi_{-i} \in b} b(\pi_{-i}) \sum_{a_{-i} \in \mathcal{A}_{-i}} R_i(S, a) \pi_{-i}(a_{-i}|S)$

Therefore, optimally solving the belief-induced MDP  $\mathcal{M}(b_t)$  at each time  $t$  is equivalent to solving a surrogate target,

$$V_i(S, b_t) = \max_{a_i \in \mathcal{A}_i} \{ \mathbb{E}_{\substack{\pi_{-i} \sim b_t \\ a_{-i} \sim \pi_{-i}}} [R_i(S, a) + \gamma \sum_{S'} T(S'|S, a) V_i(S', b_t)] \} \quad (3)$$

where  $V_i(\cdot, b_t)$  is the optimal value function characterizing the optimal control of the belief-induced MDP  $\mathcal{M}(b_t)$ . Note that the belief on the LHS is the same as that on the RHS, i.e.,  $b_t$  is fixed as a hyperparameter. That is to say, the modelling agent is assuming the others will play the stationary mixed strategies throughout the rest of the game according to the current belief.

One may notice that solving such an online surrogate target involves inherent inconsistency issues. That is, the effect on revising the belief by observing future plays of the opponents is ignored while inducing such MDPs, which is exactly the cause that leads to sub-optimal strategies for the modelling agent.

Understandably, one can then come up with two alternatives. If solving a newly induced MDP at each move is computationally feasible, then repeatedly updating the belief and solving the MDP induced by the updated belief will be a better choice. In contrast, if compiling and solving an MDP is costly or the belief itself is not worthy of iterative update, one might as well only compute the optimal policy of the MDP induced by a cautiously selected initial belief and commit to it until the end of the game.

By Equation (3) we generalize what is called *independent learners* (ILs) [16] in repeated games, which evaluate the accumulated return of the modelling agents with the opponents subsumed in the environment as stationary noises and best respond accordingly, to the ones that act under the same principle in stochastic games.

<sup>1</sup><https://pomdp.org/code/>

**3.3.3 Belief-mixed MDP (QMDP).** One may want to figure out the best response against each opponent model in the first place and then mix them afterwards according to the real-time belief, i.e., the action choice at each time step  $t$  is hence given as

$$\begin{aligned} a_t^* &\in \arg \max_{a_i \in \mathcal{A}_i} \sum_{\pi_{-i} \in \Pi_{-i}} b_t(\pi_{-i}) \cdot Q_{\mathcal{M}(\pi_{-i})}(S, a_i) \\ &= \arg \max_{a_i \in \mathcal{A}_i} \sum_{(S, \pi_{-i}) \in S \times \Pi_{-i}} \mathbb{P}[(S, \pi_{-i})] \cdot Q((S, \pi_{-i}), a_i) \end{aligned} \quad (4)$$

where  $Q_{\mathcal{M}(\pi_{-i})}$  is the optimal Q-function of  $\mathcal{M}(\pi_{-i})$  that can be solved upfront, but there are in total  $|\Pi_{-i}|$  MDPs to solve, which grows exponentially in terms of the number of agents. In fact, the second line is exactly the QMDP formulation from the POMDP community [31], which pretends that the modelling agent can observe the underlying opponent model, and the Q-function  $Q((S, \pi_{-i}), a_i)$  is obtained by optimally solving the underlying hypothetical MDP. The equality holds because the opponents are assumed not evolving their strategies, and therefore, those factorized MDPs, namely all such  $\mathcal{M}(\pi_{-i}), \forall \pi_{-i} \in \Pi_{-i}$ , are independent of each other.

Moreover, Equation (4) extends *joint action learners* (JALs) [16] in repeated games to the ones that act under the same principle in stochastic games. A JAL learns Q-values with respect to all possible opponent actions, and then assesses the best expected return by mixing the Q-values over the observed action distribution of the opponent. Clearly, ours generalizes opponent actions (in single-state scenarios) to presumed opponent policies (in multi-state scenarios), on which a corresponding Q-function is derived.

We postpone the pseudocode for these three planners to Appendix B, and some case studies to Appendix D.

## 4 UNIFICATION

In this section, we propose a unified planning framework conceptually based on tree search. The general framework has all the aforementioned formulations embedded and even sheds lights on new ones. Given the state  $S_t$  and the belief  $b_t$  at each step  $t$ , we do *three layers of lookahead search* rooted at the node  $(S_t, b_t)$ , denoted as  $(S_t^0, b_t^0)$  inside the tree (with  $t$  omitted below),

- (1) Belief-updated lookahead for the first  $n$  levels, i.e., for  $l \in [0, n)$ :

$$V_i(S^l, b^l) \leftarrow \max_{a_i \in \mathcal{A}_i} \sum_{\substack{\pi_{-i} \in b^l \\ a_{-i} \in \pi_{-i}}} [R_i(S^l, a) + \gamma \sum_{S^{l+1}} T(S^{l+1}|S^l, a) V_i(S^{l+1}, b^{l+1})],$$

where  $b^{l+1} = \xi(b^l, S^l, a)$ ,

- (2) Belief-fixed lookahead for the next  $m$  levels, for  $l \in [n, n+m)$ :

$$V_i(S^l, b^n) \leftarrow \max_{a_i \in \mathcal{A}_i} \sum_{\substack{\pi_{-i} \in b^n \\ a_{-i} \in \pi_{-i}}} [R_i(S^l, a) + \gamma \sum_{S^{l+1}} T(S^{l+1}|S^l, a) V_i(S^{l+1}, b^n)],$$

- (3) Heuristic evaluation for the leaf nodes in level  $(n+m)$ :

$$V_i(S^{n+m}, b^n) \leftarrow \text{EVAL}_i(S^{n+m}, b^n),$$

where  $\text{EVAL}_i(\cdot)$  can be any heuristic value function that estimates a reasonable future return for the modelling agent  $i$ .

Note that, for long-horizon planning problems, especially goal-conditioned ones, the planner is supposed to lookahead in depth to ensure non-trivial backup from leaf nodes, which implies that either  $(n+m)$  should be large enough or  $\text{EVAL}_i$  should be sufficiently

informative. In fact, the possession of such a heuristic enables one to evaluate any given situation by directly enquiring  $\text{EVAL}_i(S_t, b_t)$ , without doing any lookahead search. *Nevertheless, the lookahead search by the first two layers is highly necessary as it serves as an online policy improvement operator.* We postpone the proof of this rather intuitive statement to Appendix A.

By various options of  $(n, m, \text{EVAL}_i)$ , we here reproduce the aforementioned three formulations, and introduce a new fourth one,

- F1. If  $n = \infty$ , then it is equivalent to exactly solving the infinite-horizon POMDP as **Equation (2)**, and there is no need to go to layer (2) and (3), i.e.,  $m = 0$  and  $\text{EVAL}_i(\cdot) = \text{any}$ . No replanning will be needed.
- F2. If  $n = 0, m = \infty, \text{EVAL}_i(\cdot) = \text{any}$ , then it is equivalent to solving the belief-induced MDP as **Equation (3)**. One should note that  $\mathcal{S}$  is a finite set. Therefore, layer (2) search can instead be implemented as dynamic programming to handle value backup on repeated states.
- F3. If  $n = 0, m = 0$ , and

$$\text{EVAL}_i(S, b) = \max_{a_i \in \mathcal{A}_i} \sum_{\pi_{-i}} b(\pi_{-i}) \cdot Q_{M(\pi_{-i})}(S, a_i),$$

then it is equivalent to the QMDP approach as **Equation (4)**.

- F4. If  $0 < n < \infty, m = \infty$  and  $\text{EVAL}_i(\cdot) = \text{any}$ , then it is equivalent to solving a (finite)  $n$ -horizon POMDP with terminal states evaluated by respective MDPs. Considerably, this will perform better than F2, but replanning will be needed.

Note that once  $n = \infty$  or  $m = \infty$ , it does not matter which  $\text{EVAL}_i$  is adopted, as the backup operators in the first two layers are both  $\gamma$ -contractions which eventually lead to unique fixed points. The proof is postponed to Appendix A. Although one can resort to off-the-shelf MDP solvers, formulations involving optimally solving MDPs will soon become computationally infeasible as the number of agents grows. The framework then guides one to additional scalable implementations,

- F5.  $0 < n < \infty$  and  $0 < m < \infty$ , then it is equivalent to solving a finite-horizon POMDP with terminal states evaluated as finite-horizon MDPs with terminal states evaluated by a heuristic. Chances are that belief update itself is not that costly. Then one may reallocate all the computational budget of  $m$  to  $n$ , i.e., set  $(n, m) \leftarrow (n + m, 0)$ , since the complexity of doing lookahead search is the same for the first two layers.

For both cases in F5, the choice of  $\text{EVAL}_i$  is no longer trivial. It has to serve as an oracle that returns a heuristic value that is informative enough in a flash, *which is why we draw one's attention to contextual-RL policies in Section 3.3 as well as efficient NE solvers later in Section 5.* Once such an oracle is ready, F5 is all about improving (or sometimes, repairing) the oracle in real-time by online lookahead search. In fact, F5 leads to a full-width EXPECTIMAX tree search paradigm as Algorithm 6 in Appendix B, while in the literature there is indeed some alternative framework like  $\text{MAX}^n$  tree search [40, 54]. However, the latter one aims at converging to NEs, which imposes much stronger assumptions on the opponents.

One should notice that in order to perform exact backup and compute  $\max\{\cdot\}$ , the agent has to exhaustively enumerate all joint actions over belief distributions, which is of complexity  $(|\mathcal{A}_i| \times \prod_{j \neq i} |\mathcal{P}_j| \times \prod_{j \neq i} |\mathcal{A}_j|)^{n+m}$  and hence expensive when the problem

further scales up. The framework is then extended to alleviate the heavy computation, but unavoidably compromises the accuracy, by the following two implementations.

- F6. Sample actions over belief distributions and use sample mean to approximate the exact backup.
- F7. Use bandit-based lookahead search to approximate  $\max\{\cdot\}$ .

For bandit-based lookahead search to work, the planner has to utilize a node selection function that well balances exploration and exploitation, to eventually minimize accumulated regret [12]. Besides the most widely used UCT formula [27], it is also proposed to use a certain prior policy to guide the selection, resulting the pUCT formula [39, 42, 44, 45, 49]. Note a the prior policy is usually not easy to acquire, which leads to the following discussion on how to make use of NE strategies, to extract such priors as well as value estimates. An ultimate version integrating F6 and F7 leads to an opponent-aware MCTS planner as Algorithm 7 in Appendix B. It resembles the POMCP implementation [50], while we avoid using random Monte-Carlo rollouts as value estimates, since they usually cannot backpropagate useful information when the problem is goal-conditioned or the rewards are sparse. *With the help of efficient NE solvers, we scale the implementation up to suit 50-agent long-horizon planning problems, which to the best of our knowledge is the first opponent-modelling planner of this capability.*

To conclude, this unified framework draws a spectrum of planners. On top of it, one can predict the performance of a planner devised within this framework, as the deeper (and more deliberately) it searches, the better it performs. For example, with the same choice of  $\text{EVAL}_i$ , one can easily predict a performance ranking among them:  $F1 \geq F4 \geq F2 \sim F5 \geq F6 \sim F7$ . One can later see such a trend as expected in the experimental results in Table 3.

## 5 EXPERIMENTS

### 5.1 Benchmark

**5.1.1 Multi-Agent Route Planning.** In the domain of *multi-agent route planning* (MARF), a group of agents are placed in a grid world with possible obstacles. Each of them is designated a goal. We here name the benchmark as MAPP in order to drag one's attention away from the particular research area called *multi-agent path finding* (MAPF) [55, 56]. In MAPF, this fleet of robots are supposed to find a set of paths to reach their goals from the initial positions without any collision among them. Usually, the set of paths is computed upfront by an efficient centralized planner [30, 47] and then robots are forced to strictly execute them. In other words, all robots are under the control of a central planner. By contrast, in our MARF setting, only one agent is under our control, without knowing others' goals and strategies. The agent will be positively rewarded when she arrives at her goal, and will be penalized if she hits others or gets hit. The consequences of actions are deterministic and commonly known. The environment state consists of all agents' locations and is completely observable to everyone. We borrow this domain for the reason that it provides us with problem instances of customizable scales. Our experiments are conducted on configurations up to 32x32 maps with 50 agents. Compared with contemporary work [2, 13, 18], This benchmark is challenging for two reasons: (1) it is no longer repeated games, instead, involves a large number of states and joint-actions, as shown in Table 2;

**Table 1: Detailed description of the implemented planners under the unified framework.**

Planners	n	m	$\beta$	Collision penalty	EVAL <sub>i</sub>	Backup	Lookahead	Replanning
A*-Agent		1		$\infty$	single-agent A*	exact	full-width	✓
Safe-Agent		1	$1 \rightarrow 0$	$\infty$	single-agent A*	exact	full-width	✓
EnhancedSafe-Agent		1	$1 \rightarrow 0$	$\infty$	single-agent A*	exact	full-width	✓
MDPAgentFixedBelief	0	$\infty$		$< \infty$		exact		
MDPAgentUpdateBelief	0	$\infty$		$< \infty$		exact		✓
RLAgentFixedBelief	$\infty$	0		$< \infty$				
RLAgentUpdateBelief	$\infty$	0		$< \infty$				
UniformTSAgentRL	(0, $\infty$ )			$< \infty$	contextual-RL	exact	full-width	✓
CBSAgentFixedBelief	0	0		$< \infty$	NE by EECBS			
CBSAgentUpdateBelief	0	0		$< \infty$	NE by EECBS			✓
UniformTSAgentCBS	(0, $\infty$ )			$< \infty$	NE by EECBS	{exact, sampling}	full-width	✓
MCTSAgentCBSuct		0		$< \infty$	NE by EECBS	sampling	bandit-based	✓
MCTSAgentCBSpuct		0		$< \infty$	NE by EECBS	sampling	bandit-based	✓

**Table 2: Statistics in each scenario. The numbers in the “lower bounds” row are means (standard deviations).**

	Small2a (8x8 map)		Square2a (12x12 map)		Square4a (12x12 map)		Medium20a (18x18 map)		Large50a (32x32 map)	
	Rational	Self	Rational	Self	Rational	Self	Rational	Self	Rational	Self
#(states)	930		7310		$5.47 \times 10^7$		$6.44 \times 10^{46}$		$4.62 \times 10^{145}$	
#(joint-actions)	25		25		625		$5^{20} \sim 9.54 \times 10^{13}$		$5^{50} \sim 8.88 \times 10^{34}$	
#(op-types each)	30		85		85		218		818	
Lower bounds	4.16(2.14)	4.26(1.54)	6.78(3.34)	6.79(2.54)	6.61(3.35)	6.85(1.77)	11.12(5.56)	11.32(1.19)	23.2(11.25)	23.33(1.54)
Upper bounds		32		48		48		144		256

(2) deliberate long-horizon planning is definitely needed, as the rewards are goal-conditioned and hence sparse.

**5.1.2 Nash Equilibrium.** An NE typically reveals a predictable future outcome. In this routing domain, given a set of initial positions and goals, if a set of collision-free paths is found optimal up to certain metrics, e.g., minimizing the sum of lengths, it can serve as an NE as no agent will unitarily deviate in the sense of finding a shorter path without colliding to any others. In general, it is an NP-hard problem [59]. However, efficient planners can solve it very fast in practice, by making good use of constraint propagation, e.g., conflict-based search [47], its bounded sub-optimal variant [30], and even some universal planner [63].

Despite our argument that NE as a solution concept does not directly offer much help, an agent can still compute one by sampling hypothetical opponents from her belief, and make a smart use of it: (1) either directly execute the NE strategy expecting that the others will do the same, or (2) transform it into the EVAL<sub>i</sub> heuristic as we mentioned, and therefore, can plug it into our tree search framework to repair it in an online manner. A heuristic extracted from NEs is considerably more informative than random Monte Carlo rollouts. Detailed procedures will be revealed in Appendix C.

For other domains, one can always appeal to MARL methods for NEs, e.g., MADDPG [32], QMIX [38], IPPO/MAPPO [58], etc.

**5.1.3 A Special Case: Safe Agents.** Given this domain, we now present one simple-yet-powerful agent which is a myopic special case of the unified tree search framework, called *safe-agents*. An action  $a_i \in \mathcal{A}_i$  in state  $S$  is unsafe if there exists another agent  $j$  and an action  $a_j \in \mathcal{A}_j$  that will drive agent  $i$  and  $j$  into collision in the successor state. For each step, a *safe-agent* first rules out all unsafe actions and takes the best one among the rest. By “the best action”

here, we mean that the agent simply finds the action through which the shortest distance (ignoring other agents) to her goal can be minimized. In fact, one can easily construct a counter-example to get a *safe-agent* stuck at an intermediate position forever. For example, if one of the opponents stops right in the way of a *safe-agent*’s only shortest path to her goal, then this *safe-agent* would rather stop forever than move to a nearby safe position since that leads to a longer path. We can therefore make *safe-agents* slightly cleverer by incorporating domain-specific knowledge: no matter how random an opponent is, she will no longer move once she reaches her goal. This leads to the *enhanced safe-agents*, who will simply treat a nearby agent that has stayed still for a sufficiently long time as an obstacle, and replan a shortest path to get around.

In the language of our unified framework, *safe-agents* are induced by ignoring belief update, doing one-depth full-width lookahead search, using constraints to rule out unsafe actions, and evaluating the leaf nodes by values of the onward shortest paths. The *enhanced* version is further obtained by descending the belief temperature  $\beta$  from 1 to 0 when the aforementioned situations are detected, making the belief distribution a hard-max one. In this sense, these two *safe-agent* variants are both online improved versions of A\* agents (ignoring opponents) by one depth of tree search.

**5.1.4 Setup.** We now instantiate each theoretic module to the corresponding domain-specific implementation.

**Belief Initialization and Update.** Since the specific goal of each opponent is not revealed to the modelling agent, the modelling agent therefore associates each of the others a *uniform* initial belief over all empty cells. She also assumes everyone else is a goal-directed agent up to some randomness. More precisely, for each opponent  $j$ , for each empty cell  $g_k$  in the map layout, a plan will be hypothesized as a shortest path tree from all possible positions to  $g_k$ , denoted as

$\pi_j^k$  and  $\mathbb{P}[\pi_j^k] = 1/\#(\text{empty cells})$ . The modelling agent  $i$  assumes agent  $j$  will play  $\pi_j^k$  with probability  $(1 - \epsilon) \cdot \mathbb{P}[\pi_j^k]$ , and go to a random adjacent position with probability  $\epsilon$ . As for belief update, we simply use Equation (1) with  $\beta = 1$  except for the *enhanced safe-agents*. Intuitively, when  $\epsilon \rightarrow 0$ , the term  $\pi_j^k(a_j|s) \cdot b_t(\pi_j^k)$  will be zero if agent  $j$  does not follow the course of shortest-path actions towards  $g_k$ . Therefore, the belief update, to some extent, actually makes soft inference about opponents’ true goals.

**Solvers.** We use *mdptoolbox*<sup>2</sup> as the MDP solver, EECBS<sup>3</sup> [30] as the constraint-based approximate NE solver, which returns joint-plans whose lengths are no more than a user-specified factor away from the optimum, and PPO [43] implemented by *stable-baseline*<sup>4</sup> [36] as the contextual-RL algorithm.

**Hardware.** The RL experiments are conducted on a Linux machine with one NVIDIA 2080Ti GPU. The rest are done on a Mac mini (M1 CPU, 16GB memory) with multiprocessing.

## 5.2 Evaluation

In this section, we conduct a comprehensive study over 13 planners resulted from our unified framework. Table 1 shows how each implemented planner is induced by the unified framework. For each class of “FixedBelief” and its opposite “UpdateBelief”, we mean whether the planner will update the belief after observing opponents’ actions. For example, “MDPAgentFixedBelief” means the planner only solves the MDP induced by the initial belief and does no replanning, while “MDPAgentFixedBelief” means the planner will update the belief at each step and solve a new MDP. “RL” agents solve the underlying POMDP and thus no replanning is needed.

As we are investigating the setting where no prior information about opponents’ goals or strategies are revealed to the modelling agents, the golden standard is to run each proposed planner against all possible hypothetical opponents and their combinations, which is clearly impossible. To get around, we prepare three representative sets of opponents against which every planner will be tested for an average performance.

- (1) “Rational types”: a mixture of opponents that are, to some extent, goal-directed, mainly including three types: *ShortestPathAgent*, *RandomAgent(p)* and *SafeAgent*. A *ShortestPathAgent* is an agent that goes towards her own goal ignoring other agents. A *RandomAgent(p)* is a modified version who does random actions with probability  $p$ , while replans a shortest path to her goal with probability  $(1 - p)$ .
- (2) “Malicious types”: a mixture of opponents that sometimes intend to do harm to the modelling agent, what we call *ChasingAgent(p)*. A *ChasingAgent(p)* is an agent that plans a shortest path to the modelling agent with probability  $p$ , while replans a shortest path to her own goal otherwise.
- (3) “Self-play”: all agents run the same planner as every one is an autonomous agent modelling others simultaneously.

Experiments against “rational” opponents aim at simulating the situations where belief modelling more or less aligns with the underlying truth, while experiments against “malicious” ones examine whether belief-dependent planning will be broken down as belief

modelling is significantly attacked by the chasing behavior. Empirical results under “self-play” scenarios try to figure out the gap between the outcome and the potential (but unreachable) ex-post NE (of the SG realized by specific goals).

Table 2 presents some statistics for each testing scenario. For example, the one named **Medium20a** refers to a configuration with 20 agents initially randomly spawned on maps of size 18x18 with a few obstacles. Each opponent will be associated with 218 hypothetical policies by the modelling agent. The scenario contains around  $6.44 \times 10^{46}$  states and  $9.54 \times 10^{13}$  joint-actions under each state. The lower bounds for “rational/malicious” cases are computed as the average lengths of single-agent shortest paths (ignoring all collisions) for the modelling agent, while the lower bounds for “self-play” cases are computed as the average lengths per agent of the multi-agent optimal joint paths (avoiding any collision). By definition, these lower bounds are impossible to reach and hence only for reference. The upper bounds are the maximum number of steps allowed for route planning. Also, once any collision happens, the length of the route will be set to this upper bound.

We attach the overall comparison among those 13 planners in Table 3. The numbers are the average path lengths penalized by collisions. Additional details about raw path lengths and collision ratios are attached in Appendix C. We intend to present Table 3 as a practical handbook for one to select a suitable planner, or devise their own ones in a similar way, given any problem in a certain scale, and therefore, answer the research question in Section 1.

One can imagine Table 3 as an upper-triangular matrix, where each column corresponds to a particular planner, and each row shows the performances of all planners in a particular testing scenario with the best numbers in bold fonts. *One should compare the numbers in each row, while their absolute values may not be meaningful as once a collision happens the penalty will be huge.*

**As an overview**, with the testing scenario scaling up, one can notice that the most capable planners gradually become computationally infeasible as they do deliberate computation, starting from “MDPUpdate”, then “MDPFixed”, finally “UnifTSCBS”. Understandably, when exhaustive enumeration of joint-actions becomes infeasible, the only options left are MCTS planners and *safe-agent* variants. Despite that MCTS is in principle an anytime algorithm, if the increase of computational budget is not able to catch up with the inherent exponential complexity, it will definitely lead to compromised performance.

**Look closer**, Table 3 indeed echoes the potential of tree search that serves as an online policy improvement operator, even while modelling a large number of opponents, e.g., by comparing vanilla RL/CBS agents against their improved tree search versions. Additionally, MCTS planners with pUCT formula guided by NE policy priors beat those with vanilla UCT formula, but the gap is not significantly large, mainly because policy priors out of NE strategies presume others will do the same, which indeed causes some inconsistency. One should also notice that CBS agents are never the best ones, which supports our argument that NE offers little help if one aims at computing effective strategies against priorly unknown opponents. Contextual-RL here rather serves as a proof-of-concept example. Although online tree search further improves the learned contextual-policy, it is still challenging to invent a learning diagram that can be immediately applied to any configuration once trained.

<sup>2</sup><https://pymdptoolbox.readthedocs.io/en/latest/api/mdptoolbox.html>

<sup>3</sup><https://github.com/Jiaoyang-Li/EECBS>

<sup>4</sup><https://stable-baselines3.readthedocs.io/en/master/>



**Table 3: Detailed comparison across all planners. The numbers are means (standard deviations). Those numbers in bold fonts are the best ones in each scenario. Slots filled with “/” means the corresponding planners are computationally infeasible at that problem scale. We also copy the leftmost three columns of “Small2a” and “Square2a” to the lower table for better comparison.**

		Astar	Safe	EnhancedSafe	MDPFixed	MDPUpdate	RLFixed	RLUpdate	UnifTSRL
Small2a	Rational	7.25(9.18)	7.33(8.90)	4.95(3.67)	6.62(7.85)	4.76(2.91)	5.92(6.91)	5.74(6.56)	5.06(4.51)
	Malicious	12.33(13.11)	5.18(3.04)	5.18(3.04)	5.00(2.84)	<b>4.96(2.79)</b>	9.59(11.14)	9.47(11.05)	6.45(6.91)
	Self	9.19(10.92)	9.96(10.89)	5.98(5.95)	9.35(10.62)	<b>4.96(2.94)</b>	6.44(7.54)	6.18(7.11)	5.56(5.38)
Square2a	Rational	9.42(10.82)	9.60(10.85)	7.14(3.87)	8.90(9.49)	/	9.44(10.82)	9.31(10.55)	8.19(7.84)
	Malicious	17.40(18.64)	7.75(4.41)	7.75(4.41)	<b>7.28(3.58)</b>	/	17.11(18.45)	17.10(18.45)	10.10(11.14)
	Self	11.18(13.02)	11.60(13.12)	7.70(5.83)	10.91(12.38)	/	12.08(14.13)	11.78(13.76)	8.92(9.08)
		Astar	Safe	EnhancedSafe	CBSFixed	CBSUpdate	UnifTSCBS	MCTSCBSuct	MCTSCBSpuct
Small2a	Rational	7.25(9.18)	7.33(8.90)	4.95(3.67)	6.85(8.62)	6.09(7.33)	<b>4.74(3.19)</b>	5.48(5.07)	5.47(5.24)
	Malicious	12.33(13.11)	5.18(3.04)	5.18(3.04)	12.30(13.08)	11.81(12.79)	5.32(4.08)	7.15(7.8)	6.93(7.53)
	Self	9.19(10.92)	9.96(10.89)	5.98(5.95)	8.46(10.22)	6.77(8.09)	5.20(4.30)	5.48(4.21)	5.32(3.86)
Square2a	Rational	9.42(10.82)	9.60(10.85)	7.14(3.87)	9.48(10.89)	8.74(9.41)	<b>7.13(4.14)</b>	8.41(7.00)	8.27(6.95)
	Malicious	17.40(18.64)	7.75(4.41)	7.75(4.41)	17.28(18.57)	16.77(18.22)	8.41(7.63)	12.77(13.89)	11.92(13.00)
	Self	11.18(13.02)	11.60(13.12)	7.70(5.83)	10.94(12.71)	9.10(9.77)	<b>7.05(3.16)</b>	7.98(5.17)	7.61(3.96)
Square4a	Rational	13.40(15.92)	13.53(15.41)	8.44(6.60)	13.04(15.56)	11.50(13.72)	<b>8.26(7.42)</b>	12.51(12.00)	11.71(11.47)
	Malicious	20.89(20.43)	<b>11.43(10.68)</b>	11.76(11.33)	20.06(20.03)	19.71(19.87)	13.65(15.29)	19.58(17.88)	18.54(17.71)
	Self	27.42(20.78)	27.45(19.30)	11.59(10.50)	25.57(20.69)	19.14(18.92)	<b>9.56(9.55)</b>	15.93(13.99)	14.04(12.80)
Medium20a	Rational	88.45(66.72)	73.92(62.80)	<b>35.52(40.65)</b>	86.14(67.10)	71.56(66.62)	/	59.15(52.10)	56.04(52.59)
	Malicious	96.27(64.83)	<b>40.26(48.63)</b>	40.62(49.03)	91.89(65.88)	90.82(65.98)	/	64.97(56.53)	64.02(57.07)
	Self	144.00(0.00)	<b>49.27(28.73)</b>	67.38(51.92)	144.00(0.00)	144.00(0.00)	/	/	/
Large50a	Rational	182.01(111.03)	111.80(95.58)	<b>74.60(78.27)</b>	187.32(108.28)	163.06(114.24)	/	120.48(91.51)	119.26(96.18)
	Malicious	193.73(105.03)	<b>79.42(92.44)</b>	79.84(92.48)	182.22(109.31)	188.22(106.83)	/	145.88(106.18)	132.87(104.97)
	Self	256.00(0.00)	<b>76.65(5.30)</b>	209.15(82.39)	/	/	/	/	/

**Unexpectedly**, *safe-agents* and its *enhanced* versions perform surprisingly well, especially in cases of larger scales and against malicious opponents. For large-scale instances, due to the limited computational budget, tree search agents are not capable of doing deliberate lookahead search to significantly outperform *safe-agents*, while the latter ones easily beat all the rest in terms of fast replanning due to their simple algorithmic structure. For situations against malicious opponents where belief modelling does not match the underlying truth, *safe-agents* conservatively rule out all unsafe action and therefore guarantee a lower chance of collision. Alternatively, one highly interpretable angle is to see *safe-agents* as MINIMAX tree search that offers a fairly good worst-case performance, especially when your opponents turn out to be harmful ones.

**For case-by-case recommendations**, Table 3 roughly shows the computational limit of each planner. One should always identify the scale of the problem at first and find the most suitable planner accordingly. For example, given a problem of  $10^6$  states with three agents, we would definitely vote for two-depth full-width tree search with an NE oracle against MCTS. However, in a huge scale where value estimate oracles like RL/NE become out of reach, the only choice left is to devise rule-based planners such as *safe-agents*, which at least offer certain conservative guarantees.

For other planners that may not be suitable for a multi-run evaluation, we attach a few case studies about them in Appendix D.

## 6 CONCLUSION

To conclude, we formalize the problem of controlling one single agent against multiple opponents that are priorly unknown. A spectrum of formulations are drawn, all of which are further unified under a tree search perspective. We underpin the investigation by

offering a challenging benchmark, namely multi-agent route planning. We also show by this general framework how to customize domain-specific planners such as *safe-agents*. To offer a practical handbook of proper selection of the induced planners, we have empirically tested each of them against three representative groups of opponents. One interesting observation from our experiments is that those conservative and myopic *safe-agents* perform sufficiently well in most cases, especially when belief modelling does not match the underlying truth or deliberate replanning is computationally infeasible. As for **future work**, we point out a few valuable directions:

- (1) How to design a formal language to describe domain knowledge that can be embedded into those general planners [21]?
- (2) How to model dependencies among agents which may potentially decompose the complexity of modelling joint-transitions and joint-utilities [33], and what if such dependencies are dynamic, i.e., agents may come and go?
- (3) What if there are infinitely many opponents, is it justifiable to model the problem as a mean-field game [29]?
- (4) Other applicable domains like mechanism design [57] and even negotiation [8, 24, 28], which also turn out to involve strategic behaviors that explore and exploit opponents during repeated interaction.

## ACKNOWLEDGMENTS

This work was supported in part by a generous research grant from Xiaoi Robot Technology Limited. We thank Yuxin Pan, Chenglin Wang, Haozhe Wang, and Yangfan Wu for their valuable feedback during the revision of this paper. We also appreciate the anonymous reviewers for their insightful comments.



## REFERENCES

- [1] Stefano V Albrecht, Jacob W Crandall, and Subramanian Ramamoorthy. 2016. Belief and truth in hypothesised behaviours. *Artificial Intelligence* 235 (2016), 63–94.
- [2] Stefano V Albrecht and Subramanian Ramamoorthy. 2015. A game-theoretic model and best-response learning method for ad hoc coordination in multiagent systems. *arXiv preprint arXiv:1506.01170* (2015).
- [3] Stefano V Albrecht and Subramanian Ramamoorthy. 2019. On convergence and optimality of best-response learning with policy types in multiagent systems. *arXiv preprint arXiv:1907.06995* (2019).
- [4] Stefano V Albrecht and Peter Stone. 2018. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence* 258 (2018), 66–95.
- [5] Ioannis Antonoglou, Julian Schrittwieser, Sherjil Ozair, Thomas K Hubert, and David Silver. 2022. Planning in Stochastic Environments with a Learned Model. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=X6D9bAHhBQ1>
- [6] Dipayan Banerjee and Sandip Sen. 2007. Reaching pareto-optimality in prisoner’s dilemma using conditional joint action learning. *Autonomous Agents and Multi-Agent Systems* 15 (2007), 91–108.
- [7] Jacob Beck, Risto Vuorio, Evan Zheran Liu, Zheng Xiong, Luisa Zintgraf, Chelsea Finn, and Shimon Whiteson. 2023. A survey of meta-reinforcement learning. *arXiv preprint arXiv:2301.08028* (2023).
- [8] Martin Beer, Mark d’Inverno, Michael Luck, Nick Jennings, Chris Preist, and Michael Schroeder. 1999. Negotiation in multi-agent systems. *The Knowledge Engineering Review* 14, 3 (1999), 285–289.
- [9] Carolin Benjamins, Theresa Eimer, Frederik Schubert, André Biedenkapp, Bodo Rosenhahn, Frank Hutter, and Marius Lindauer. 2021. Carl: A benchmark for contextual and adaptive reinforcement learning. *arXiv preprint arXiv:2110.02102* (2021).
- [10] Craig Boutilier. 1996. Planning, learning and coordination in multiagent decision processes. In *TARK*, Vol. 96. Citeseer, 195–210.
- [11] Sally C Brailsford, Chris N Potts, and Barbara M Smith. 1999. Constraint satisfaction problems: Algorithms and applications. *European journal of operational research* 119, 3 (1999), 557–581.
- [12] Cameron B Browne, Edward Powley, Daniel Whitehouse, Simon M Lucas, Peter I Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. 2012. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in games* 4, 1 (2012), 1–43.
- [13] David Carmel and Shaul Markovitch. 1998. How to explore your opponent’s strategy (almost) optimally. In *Proceedings International Conference on Multi Agent Systems (Cat. No. 98EX160)*. IEEE, 64–71.
- [14] David Carmel and Shaul Markovitch. 1999. Exploration strategies for model-based learning in multi-agent systems: Exploration strategies. *Autonomous Agents and Multi-agent systems* 2 (1999), 141–172.
- [15] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. 2024. End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).
- [16] Caroline Claus and Craig Boutilier. 1998. The dynamics of reinforcement learning in cooperative multiagent systems. *AAAI/IAAI 1998*, 746–752 (1998), 2.
- [17] Ivo Danihelka, Arthur Guez, Julian Schrittwieser, and David Silver. 2022. Policy improvement by planning with Gumbel. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=bERaNdOegNO>
- [18] Adam Eck, Maulik Shah, Prashant Doshi, and Leen-Kiat Soh. 2020. Scalable decision-theoretic planning in open and typed multiagent systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 7127–7134.
- [19] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. 2018. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.
- [20] Wei Fu, Chao Yu, Zelai Xu, Jiaqi Yang, and Yi Wu. 2022. Revisiting Some Common Practices in Cooperative Multi-Agent Reinforcement Learning. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 6863–6877. <https://proceedings.mlr.press/v162/fu22d.html>
- [21] Zihang Gao, Fangzhen Lin, Yi Zhou, Hao Zhang, Kaishun Wu, and Haodi Zhang. 2020. Embedding high-level knowledge into dqn to learn faster and more safely. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 13608–13609.
- [22] Assaf Hallak, Dotan Di Castro, and Shie Mannor. 2015. Contextual markov decision processes. *arXiv preprint arXiv:1502.02259* (2015).
- [23] John C Harsanyi. 1967. Games with incomplete information played by “Bayesian” players, I–III Part I. The basic model. *Management science* 14, 3 (1967), 159–182.
- [24] Nicholas R Jennings, Peyman Faratin, Alessio R Lomuscio, Simon Parsons, Charles Sierra, and Michael Wooldridge. 2001. Automated negotiation: prospects, methods and challenges. *International Journal of Group Decision and Negotiation* 10, 2 (2001), 199–215.
- [25] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. 1998. Planning and acting in partially observable stochastic domains. *Artificial intelligence* 101, 1-2 (1998), 99–134.
- [26] Ehud Kalai and Ehud Lehrer. 1993. Rational learning leads to Nash equilibrium. *Econometrica: Journal of the Econometric Society* (1993), 1019–1045.
- [27] Levente Kocsis and Csaba Szepesvári. 2006. Bandit based monte-carlo planning. In *European conference on machine learning*. Springer, 282–293.
- [28] Sarit Kraus. 1997. Negotiation and cooperation in multi-agent environments. *Artificial intelligence* 94, 1-2 (1997), 79–97.
- [29] Jean-Michel Lasry and Pierre-Louis Lions. 2007. Mean field games. *Japanese journal of mathematics* 2, 1 (2007), 229–260.
- [30] Jiaoyang Li, Wheeler Ruml, and Sven Koenig. 2021. Ecbs: A bounded-suboptimal search for multi-agent path finding. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 12353–12362.
- [31] Michael I Littman, Anthony R Cassandra, and Leslie Pack Kaelbling. 1995. Learning policies for partially observable environments: Scaling up. In *Machine Learning Proceedings 1995*. Elsevier, 362–370.
- [32] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems* 30 (2017).
- [33] Chengdong Ma, Aming Li, Yali Du, Hao Dong, and Yaodong Yang. 2024. Efficient and scalable reinforcement learning for large-scale network control. *Nature Machine Intelligence* (2024), 1–15.
- [34] Omid Madani, Steve Hanks, and Anne Condon. 1999. On the undecidability of probabilistic planning and infinite-horizon partially observable Markov decision problems. In *Proceedings of the sixteenth national conference on Artificial intelligence and the eleventh Innovative applications of artificial intelligence conference innovative applications of artificial intelligence*. 541–548.
- [35] Christos H Papadimitriou and John N Tsitsiklis. 1987. The complexity of Markov decision processes. *Mathematics of operations research* 12, 3 (1987), 441–450.
- [36] Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. 2021. Stable-Baselines3: Reliable Reinforcement Learning Implementations. *Journal of Machine Learning Research* 22, 268 (2021), 1–8. <http://jmlr.org/papers/v22/20-1364.html>
- [37] Arrasy Rahman, Ignacio Carlucho, Niklas Höpner, and Stefano V Albrecht. 2023. A general learning framework for open ad hoc teamwork using graph-based policy learning. *Journal of Machine Learning Research* 24, 298 (2023), 1–74.
- [38] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. 2020. Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research* 21, 178 (2020), 1–51.
- [39] Christopher D Rosin. 2011. Multi-armed bandits with episode context. *Annals of Mathematics and Artificial Intelligence* 61, 3 (2011), 203–230.
- [40] Spyridon Samothrakis, David Robles, and Simon Lucas. 2011. Fast approximate max-n monte carlo tree search for ms pac-man. *IEEE Transactions on Computational Intelligence and AI in Games* 3, 2 (2011), 142–154.
- [41] Mikayel Samvelyan, Tabish Rashid, Christian Schroeder de Witt, Gregory Farquhar, Nantas Nardelli, Tim G. J. Rudner, Chia-Man Hung, Philip H. S. Torr, Jakob Foerster, and Shimon Whiteson. 2019. The StarCraft Multi-Agent Challenge. *CoRR abs/1902.04043* (2019).
- [42] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. 2020. Mastering atari, go, chess and shogi by planning with a learned model. *Nature* 588, 7839 (2020), 604–609.
- [43] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).
- [44] Jonathon Schwartz and Hanna Kurniawati. 2023. Bayes-Adaptive Monte-Carlo Planning for Type-Based Reasoning in Large Partially Observable, Multi-Agent Environments. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*. 2355–2357.
- [45] Jonathon Schwartz, Hanna Kurniawati, and Marcus Hutter. 2023. Combining a Meta-Policy and Monte-Carlo Planning for Scalable Type-Based Reasoning in Partially Observable Environments. *arXiv preprint arXiv:2306.06067* (2023).
- [46] Lloyd S Shapley. 1953. Stochastic games. *Proceedings of the national academy of sciences* 39, 10 (1953), 1095–1100.
- [47] Guni Sharon, Roni Stern, Ariel Felner, and Nathan R Sturtevant. 2015. Conflict-based search for optimal multi-agent pathfinding. *Artificial Intelligence* 219 (2015), 40–66.
- [48] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharmashan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. 2018. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* 362, 6419 (2018), 1140–1144. <https://doi.org/10.1126/science.aar6404> <https://www.science.org/doi/pdf/10.1126/science.aar6404>
- [49] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. 2017. Mastering the game of go without human knowledge. *nature* 550,

- 7676 (2017), 354–359.
- [50] David Silver and Joel Veness. 2010. Monte-Carlo planning in large POMDPs. *Advances in neural information processing systems* 23 (2010).
  - [51] Richard D Smallwood and Edward J Sondik. 1973. The optimal control of partially observable Markov processes over a finite horizon. *Operations research* 21, 5 (1973), 1071–1088.
  - [52] Eilon Solan and Nicolas Vieille. 2015. Stochastic games. *Proceedings of the National Academy of Sciences* 112, 45 (2015), 13743–13746.
  - [53] Edward J Sondik. 1978. The optimal control of partially observable Markov processes over the infinite horizon: Discounted costs. *Operations research* 26, 2 (1978), 282–304.
  - [54] Dale O Stahl. 1993. Evolution of smartn players. *Games and Economic Behavior* 5, 4 (1993), 604–617.
  - [55] Roni Stern. 2019. Multi-agent path finding—an overview. *Artificial Intelligence* (2019), 96–115.
  - [56] Roni Stern, Nathan R Sturtevant, Ariel Felner, Sven Koenig, Hang Ma, Thayne T Walker, Jiaoyang Li, Dor Atzmon, Liron Cohen, TK Satish Kumar, et al. 2019. Multi-agent pathfinding: Definitions, variants, and benchmarks. In *Twelfth Annual Symposium on Combinatorial Search*.
  - [57] Pingzhong Tang. 2017. Reinforcement mechanism design. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 5146–5150.
  - [58] Chao Yu, Akash Velu, Eugene Vinitzky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. 2022. The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games. In *Thirty-sixth Conference on Neural Information Processing Systems*.
- Datasets and Benchmarks Track.* <https://openreview.net/forum?id=YVXaxB6L2Pl>
- [59] Jingjin Yu and Steven LaValle. 2013. Structure and intractability of optimal multi-robot path planning on graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 27. 1443–1449.
  - [60] Han Zhang, Jiaoyang Li, Pavel Surynek, TK Satish Kumar, and Sven Koenig. 2022. Multi-agent path finding with mutex propagation. *Artificial Intelligence* 311 (2022), 103766.
  - [61] Nevin Lianwen Zhang and Weihong Zhang. 2001. Speeding up the convergence of value iteration in partially observable Markov decision processes. *Journal of Artificial Intelligence Research* 14 (2001), 29–51.
  - [62] Ming Zhou, Ziyu Wan, Hanjing Wang, Muning Wen, Runzhe Wu, Ying Wen, Yaodong Yang, Yong Yu, Jun Wang, and Weinan Zhang. 2023. MALib: A Parallel Framework for Population-based Multi-agent Reinforcement Learning. *Journal of Machine Learning Research* 24, 150 (2023), 1–12. <http://jmlr.org/papers/v24/22-0169.html>
  - [63] Fengming Zhu and Fangzhen Lin. 2024. On Computing Universal Plans for Partially Observable Multi-Agent Path Finding. arXiv:2305.16203 [cs.MA] <https://arxiv.org/abs/2305.16203>
  - [64] Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. 2007. Regret minimization in games with incomplete information. *Advances in neural information processing systems* 20 (2007).