CF-011

ドローン配送問題のための安全制約付き マルチエージェント強化学習

Safety-constrained Multi-agent Reinforcement Learning for Drone Routing Problems

加地 正拡[†] 林 冬惠[‡‡]
Masahiro Kaji Donghui Lin

1 はじめに

近年，ネット通販などによる運送業の需要の高まり，ドローン配送が注目されている．ドローン配送は無人機による省コスト化，省人化の効果が大きく，物流業界の人手不足や郊外地域の買い物の不便さの解消が期待できる．

ドローン配送では複数台のドローンによる配送が想定されるが，衝突事故を回避するためには各々が自由に飛行するのではなく，飛行計画や航空管制を考える必要がある．また，墜落のリスクを考えると安全に飛行できる領域は限られる．この限られた空路を衝突することなく，できるだけ効率的に移動することを考える必要がある．この問題は，複数のドローンでの配送の経路計画問題として，ドローン配送問題（Drone Routing Problem，DRP）と定義されている [1] [2]. DRP は複数のドローンを扱う問題であるため，複数のドローンを複数のエージェントとみなし，マルチエージェント強化学習（Multi-Agent Reinforcement Learning，MARL）を適用し，学習済みのモデルを利用することでオンデマンド配送による秒単位の経路計画への適用が可能である．そのため，本研究では MARL で DRP を解くことを検討する．

DRP を MARL で解く研究は進められているが，試行錯誤によって学習を進めるため，その過程で衝突が多く発生してしまう [1]．また，学習済みモデルについても，衝突を完全になくすことが困難であり，一定数の衝突が発生してしまう．学習によってある程度の衝突は避けるようになるが，学習中のエージェントや，学習済みモデルのエージェントの十分な安全性は保障されていない．そのため，特にドローン配送時の安全性を重視した，DRP のための安全な MARL の設計を考える必要がある．また，安全を確保するだけではできるだけ効率的に移動するという本来の目的を達成できないため，学習済みのモデルの性能をなるべく保つことも必要となる．

そこで，本稿では，DRP に MARL を適用する際に安全を確保するための手法として，視野の追加とエージェントの危険な行動の制限について提案する．視野の追加の目的は，各エージェントの状態の潜在的な危険を学習させることである．視野によって，エージェントは近い位置にいる他のエージェントの情報を与えられ，潜在的な危険を学習する．危険な行動の制限の目的は，エージェントが危険な行動を行うときに，その行動を制限して安全な行動をとらせることにより，衝突の原因となる行動をさせないことである．危険な行動を定義し，危険な行動を検知した場合は，その行動を安全な行動に変更する．本研究で扱う DRP はマルチエージェントシステムを想定しているため，エージェント間で協力的な行動をとることにより学習パフォーマンスの向上が期待できる．そのため，危険な行動を制限する際に，エージェント間で協調する仕組みを取り入れる．また，提案手法について安全性，移動の効率の観点から有効性を示す．

本稿は以下のように構成されている．2 章では，ドローン配送を定式化した DRP について説明し，DRP に MARL を適用するための環境，適用する際の課題について述べる．3 章では，DRP に MARL を適用する際に安全を確保するための提案手法について述べる．4 章では，提案手法の有効性を確かめるための実験結果を示し，評価，考察をする．5 章では，本研究のまとめと今後の展望について述べる．

[†]岡山大学大学院 環境生命自然科学研究科
Graduate School of Environmental, Life, Natural Science and Technology, Okayama University
[‡‡]岡山大学 学術研究院環境生命自然科学学域
Faculty of Environmental, Life, Natural Science and Technology, Okayama University

2 背景

2.1 ドローン配送問題

複数のドローンが限られた空路を衝突することなく，できるだけ効率的に移動することを考える問題が DRP と定義されている [1]．DRP では，ドローンの移動可能な領域を図 1 のような 2 次元の非グリッドマップ $G =< V, E >$ で表現する[1]．ここで，$V = \{v_1, \ldots, v_{|V|}\}$ は，位置情報 $l_k = (l_k^x, l_k^y)$ を持つノード v_k の集合として表現される．また，E は，$E = \{(v_j, v_k)\}$ として，v_j と v_k をつなぐエッジの集合を表す．各ノードのすべての隣接ノードの集合は $v_k^{nei} = \{v_j | (v_j, v_k) \in E\}$ となる．ドローンの集合は，$N = \{1, \ldots, i, \ldots, |N|\}$ と表現される．各ドローン i には，エピソードごとにスタート地点 $s^i \in V$ とゴール地点 $g^i \in V$ が与えられる．エッジ上での各ドローンの座標を $l_k = (l^x, l^y)$ と表現する．各ドローン i の T ステップまでの行動の軌跡を $path^i = (l^i[0], l^i[1], \ldots, l^i[T])$ のように表す．ここで，$l^i[0] = s^i$ であり，t ステップでゴール地点に到達したなら $l^i[t] = g^i$ となる．ドローンがゴール地点に到達した場合，それ以降，そのドローンはゴール地点で待機を続ける．つまり，$l^i[t']|_{t'>t} = g^i$ となる．

行動の軌跡 $path^i$ のコスト関数 $cost$ は次のように計算される．

$$cost(path^i) = \sum_{t=0}^{T-1} \|l^i[t+1] - l^i[t]\|_2$$

DRP の目的は，上記の制約を満たしながら総移動コストを最小にする各ドローン i の経路を発見することである．これは次のように定式化できる．

$$\min \sum_{epi} \sum_{i} cost(path^i_{epi})$$

$$st. \quad \forall i \in N, \quad l^i[T] = g^i_{epi}$$

ここで，$l^i[T] = g^i_{epi}$ は，ドローンが最大ステップ T にてゴール地点 g^i_{epi} にとどまることを意味する．g^i_{epi} の表記は，エピソードごとにゴール地点が動的に変化することを表している．

2.2 マルチエージェント強化学習

前章で述べたように DRP で想定されているオンデマンド配送による秒単位での経路計画に対応するため，MARL を適用することについて検討する．

[1]ドローンは同じ高さを飛行する想定であり，モデルを簡単化するため 2 次元での定義をしている．また，高さの概念を追加することで 3 次元への拡張は可能である．

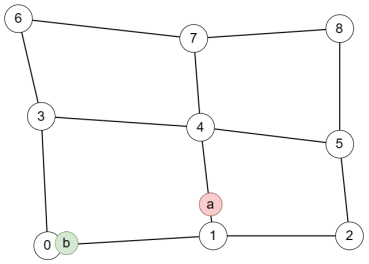


図 1: DRP で表現されるドローンの移動可能な領域

MARL はマルコフ決定過程の枠組みを用いてエージェントが最適な方策を見つける問題を解決する [3]．MARL に対応するマルコフ決定過程はタプル (N, S, A, P, P_0, R) で表される．ここで，$N = \{1, \ldots, n\}$ はエージェントの集合，S は全てのエージェントによって観測される状態空間，$A = A^1 \times \cdots \times A^N$ は全てのエージェントの共同行動空間である．A^i はエージェント i の行動空間を表す．P は状態遷移確率であり，$P(s_{t+1}|s_t, a_t)$ となる．時間ステップ t での状態 s_t とエージェントの共同行動 $a_t = \{a_t^1, \ldots, a_t^n\}$ によって次の時間ステップ $t+1$ の状態 s_{t+1} に遷移する確率を表す．P_0 は状態の初期状態確率分布である．R は報酬関数である．状態 s_t で行動 a_t を取り，次状態が s_{t+1} となったときに得る共同報酬関数 $R(s_t, a_t, s_{t+1})$ を規定する．MARL の目的は，将来の累積報酬 $R_t = \sum_{t=0}^{\infty} \gamma^t R(s_t, a_t, s_{t+1})$ を最大化する，各エージェントの方策 $\pi^i(a^i|s)$ を学習することである．ここで，γ は報酬の割引率を表す．

MARL の代表的な手法として IQL や QMIX，MADDPG などがあるが [4]，これらのような MARL ではエージェントの安全性は保障されておらず試行錯誤によって学習を進めるため，その過程で衝突が多く発生してしまう．また，学習済みモデルについても，衝突を完全になくすことが困難であり，一定数の衝突が発生する．実際のドローンを用いて学習する際や，学習済みモデルを用いて運航する際に衝突が発生することでドローンの墜落の危険や，配送計画への悪影響が想定される．そこで，本研究ではエージェント間の衝突を防ぐために DRP に MARL を適用する際に安全を確保するための手法を提案する．

3 提案手法

この章では，DRP において安全性を確保するための手法として，状態表現への視野の追加と危険な行動の制限について提案する．危険な行動の制限は，行動の実行時に安全制約を設けることでエージェントの安全

を確保する．

3.1 視野の追加

各エージェントの状態は，現在地，目的地をベクトルとして表す．これを [現在地，目的地] と表現する．現在地，目的地はそれぞれノード数と同じサイズの配列で表現され，図 1 でのノードの番号は配列の番号に対応している．現在地の表現は，エージェントがノード v_i 上にあるなら，現在地の配列の i 番目の要素を 1 とする．エージェントがエッジ上にあるとき，エッジがつなぐノード v_i，v_j に対応する配列の要素の和が 1 になるように，それぞれ 0 から 1 の間の数値となる．それぞれの要素はエージェントがノードに近いほど，そのノードに対応する配列の要素が大きくなる．目的地の表現は，エージェントの目的地がノード v_k なら，目的地の配列の k 番目の要素を 1 とする．

この状態表現に視野を追加する．視野は，ロボットのナビゲーション機能のための強化学習などに用いられており，エージェントに周囲の情報を与える [5]．ここでの視野は，現在地のノードと隣接するノード付近に他のエージェントがいる場合に，その情報をエージェントに与える．視野の情報は，現在地の情報に合わせて与えられる．近くのノード付近に他のエージェントがいる場合，そのノードの番号を表す部分を -1 とすることで，他のエージェントが存在するという情報を与える．

図 1 上のエージェント a を例に状態表現を説明する．それぞれの状態表現の例を表 1 に示す．図 1 では，エージェント a はノード 1 と 4 の間のエッジ上で，ノード 1 に近い位置にある．そのため，エージェント a の現在地の表現は $[0, 0.8, 0, 0, 0.2, 0, 0, 0, 0]$ となる．エージェント a に視野を追加した場合，ノード 0 の近くに他のエージェントがいるため，現在地の表現の 0 番目を -1 にする．それによって，ノード 0 付近に他のエージェントがいることを感知する．

視野によって，エージェントは他の近くのエージェントの情報を得ることができる．それによって，他のエージェントの位置による潜在的な危険を学習し，衝突を避けるように行動することが期待できる．

表 1: 図 1 でのエージェント a の状態表現

視野なし	$[0, 0.8, 0, 0, 0.2, 0, 0, 0, 0]$
視野あり	$[-1, 0.8, 0, 0, 0.2, 0, 0, 0, 0]$

3.2 危険な行動の制限

ElSayed-Aly らは安全な MARL の手法として，危険な行動を制限する手法を提案している [6]．この手法では，エージェントの行動を監視し，危険な行動をとらせないことで安全を確保する．危険な行動を制限する動作を説明する．まず，危険な行動を定義しておく．そして，学習を行う全てのエージェントの行動を確認する．その行動が危険な行動であることを検知した場合，安全な行動に変換する．それ以外の場合は，そのまま環境に作用させる．環境は安全な行動を受け取り，次状態に移行し，エージェントに報酬を与える．また，MARL アルゴリズムが安全でない行動について学習できるように，エージェントにペナルティとして，負の報酬を割り当てる．安全でない行動を環境に作用させる前に安全な行動に修正することで，学習中の安全を確保する．

本研究ではこの手法に着目し，DRP に反映させる．それに加えて，マルチエージェントシステムを想定しているため，エージェント間の協調が必要となる．そのため，危険な行動を制限する際にエージェント間で協調する仕組みを取り入れる．

図 2 は危険な行動を制限する手法を適用した MARL を表す．従来の MARL は，エージェントは状態を環境から受け取り，その状態と方策からエージェントがとるべき行動が出力される．そして，エージェントがその行動を行い，環境から報酬と新たな状態を与えられるという流れを繰り返す．危険な行動を制限する MARL では，方策によって得られた行動を安全制約によって確認し，その行動が危険な行動である場合は安全な行動に変更する．その際，危険な行動であることを学習させるために，エージェントに負の報酬をペナルティとして与える．

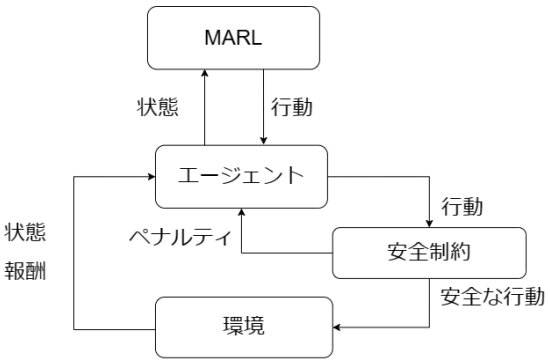


図 2: 危険な行動を制限する MARL

ここでは，危険な行動を以下のように定義する．

- 他のエージェントと同じエッジを通る行動

- 他のエージェントと次に目指すノードが同じである行動

本研究はマルチエージェントシステムを想定しているが，マルチエージェントシステムでは各エージェントと他のエージェントとの協調が必要である．複数のエージェントが同じノードを目指したときに，1 つのエージェントにはその行動を許可し，他のエージェントを停止させる．このようにあるエージェントに道を譲るような協調的な行動をさせる．このエージェント間の協調によって，不要な停止行動の削減による移動の効率の上昇が期待できる．

4 実験・評価

この章では，実験のための条件や設定について説明する．その後，前章で述べた提案手法の有効性を確認するために，エピソードにおける衝突率，ゴール率，タイムアップ率，コスト C を比較する．エピソード終了時の状態は衝突，ゴール，タイムアップの 3 つが存在しており，衝突率，ゴール率，タイムアップ率はそれぞれの状態で終了するエピソードの割合である．2 章にて距離でコストを定義したが，本研究の MARL では，ステップ数が距離に比例するため，ステップ数によってコストを再定義する．コスト C は，各エージェントが目的地に到達するまでのステップ数の和と定義する．コスト C は $C = \sum_{i=1}^{N} c_i$ と表せる．ここでは，N はエージェント数，c_i はエージェント i が目的地に到着するまでにかかったステップ数を表す．衝突することや設定している上限ステップ数を超えることによって全てのエージェントが目的地に到達できなかった場合，コストをエージェント数と上限ステップ数の積としており，上限ステップ数を T とすると，コストは $C = N \times T$ となる．このコストによって，効率的に移動できているか，安全性をどの程度確保しているかという点を総合的に評価する．

4.1 実験の設定

エピソードの状態は，エピソード中にエージェントが衝突した場合は衝突，全てのエージェントが目的地に到着した場合はゴールとなる．エピソードには上限ステップ数が定められており，エピソード中にステップ数が上限ステップ数に到達した場合，エピソードの状態はタイムアップとなる．また，1 ステップに進む距離を $speed$ として定義する．報酬は，エージェントが停止行動をしたときに $-10 \times speed$，移動したときに $-1 \times speed$，衝突したときに $-10 \times speed$，目的地に到達したときに 100 を与える．本研究では，スピード $speed$ は 5 としている．また，本研究の実験では MARL の代表的な手法である QMIX [7] を利用する．実験では，図 1 のようなグリッドに近い形で，ノード数が 8×5 の 40 個となるマップと，実在する街路をモデル化したノード数が 43 個の図 3 のマップを用いる．

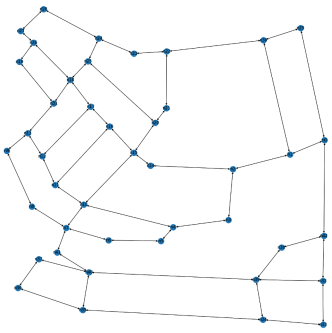


図 3: 実在する街路をモデル化したマップ

4.2 視野の効果の評価

図 4 は視野の追加によるコストの比較である．この実験はエージェント数を 4，マップをノード数が 8×5 のものとし，エピソードの上限ステップ数を 100 にして行った．視野の有無と危険な行動の制限（安全制約）の有無をそれぞれ組み合わせて学習させた．

安全制約なしの場合，視野ありの方がコストが 100 ほど低く，移動の効率と安全性を総合的に評価すると視野ありの方がよいといえる．安全制約ありの場合も同様に，視野ありの方がコストが低いため，こちらも視野ありの方が良いと言える．

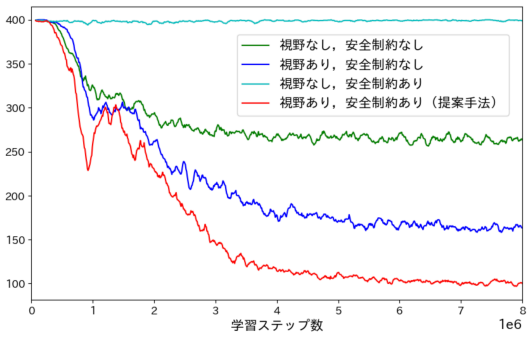


図 4: 視野の効果：コスト

安全制約の有無に関わらず，視野ありの方がよい学

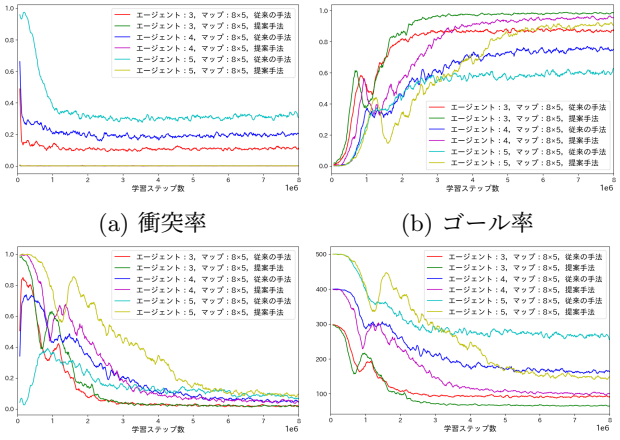
習結果となるため，以降の危険な行動の制限についての実験では，視野を追加した状態での実験結果を示す．

4.3 危険な行動の制限の評価

MARL に安全制約を取り入れた手法の実験，評価を行う．ここでは，従来手法を QMIX，提案手法を安全制約を取り入れた QMIX としている．

危険な行動を制限する際にエージェント間で協調する手法が，様々な条件でも有効なものであるか検証するための実験を行なった．まずは，マップをノード数が 8×5 のものに固定して，エージェント数を 3，4，5 にして実験した．エージェント数が多くなるほどマップ中のエージェントの密度は大きくなり，エージェント同士の衝突が起こりやすくなる．衝突の起こりやすさの違いによる，提案手法の効果の変化を確認した．

その結果は図 5 となった．衝突率に注目すると，従来の手法では，エージェント数が 3 の場合は 0.15，4 の場合は 0.2，5 の場合は 0.35 程度となり，エージェント数が多いほど衝突率が高いことが確認できる．しかし，提案手法では，エージェント数に関わらず，衝突率は 0 となり安全の確保ができている．ゴール率に注目すると，どのエージェント数で比較しても，提案手法での結果は従来の手法の結果よりもゴール率が高くなっていることがわかる．しかし，提案手法と従来の手法のどちらでも，エージェント数が多いほどゴール率が低くなっている．コストに注目すると，どのエージェント数においても提案手法は従来の手法よりも低くなっているため，移動の効率と安全性を総合的に評価するとエージェント数に関わらず，提案手法の方が良い結果であるといえる．

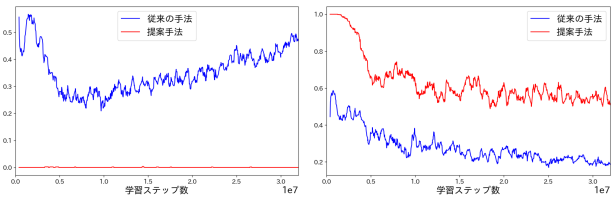


(a) 衝突率　(b) ゴール率

(c) タイムアップ率　(d) コスト

図 5: 様々な条件での危険な行動の制限の効果

次に，実在する街路をモデル化したマップでの実験を行なった．使用するマップは図 3 のマップである．この実験では，エージェント数を 4，エピソードの上限ステップ数を 200 としている．結果は図 6 となった．衝突率に注目すると，従来の手法では衝突率は 0.5 程度，提案手法の実験では 0 となり，安全が確保されていることがわかる．しかし，タイムアップ率に注目すると，従来の手法ではその値は 0.2 程度であるのに対し，提案手法では 0.5 程度である．そのため，この実験では，提案手法は移動の効率を保つことができていないといえる．また，コストは提案手法の方が良いものの，大きくは変わらなかった．安全の確保と移動の効率の総合的な評価は，提案手法と従来の手法では大きくは変わらないという結果となる．



(a) 衝突率　(b) タイムアップ率

図 6: 図 3 のマップでの実験結果

4.4 考察

4.3 節にて，提案手法はエージェント数に関わらず，安全を確保しつつ移動の効率を保つことができるがエージェント数が多いほどゴール率が低いことを明らかにした．提案手法によって衝突の原因となる行動を防ぐためにエージェントを停止させるが，エージェント数が多い場合は衝突が起こりやすい．そのため，エージェントを停止させる機会が多くなり，その結果として上限ステップ数以内に全てのエージェントが目的地に到達できないことが原因であると考えられる．本研究では，エージェント数は最大で 5 としているが，エージェント数が多くなるほど提案手法でのゴール率が低くなるため，エージェント数が多くなる場合は提案手法では十分な学習ができないと考えられる．そのため，提案手法を用いる場合は十分な学習ができるエージェント数の検証が必要となる．

また，実在する街路をモデル化したマップでの実験では，提案手法は安全の確保はできるが移動の効率を保てていないという結果を示した．図 3 をみると確認できるが，図 1 のようなマップよりもノードの数に対してのエッジの数が少ない．エッジの数が少ないこと

により，エージェントがノード上にいるときの行動の選択肢が少なくなる．その状態で提案手法を適用した場合，移動する行動の選択肢で安全だと判断されるものが少なくなり，停止する行動が増えることが原因であると考えられる．提案手法は行動の制限が強く，図 3 のようなマップでは移動の効率が大きく落ちてしまうため今後はバランスを考慮したアルゴリズムが必要である．

5 おわりに

ドローン配送問題（Drone Routing Problem，DRP）にマルチエージェント強化学習（Multi-Agent Reinforcement Learning，MARL）を適用する場合，試行錯誤によって学習を進めるため，その過程で衝突が多く発生してしまうことや，学習済みのモデルにて衝突を完全になくすことは困難であることが問題として挙げられた．本研究では，この問題を解決するために，DRP に MARL を適用する際に安全を確保するため，各エージェントの状態表現に視野を追加することとエージェントの危険な行動を制限することを組み合わせた手法を提案した．

実験では，提案手法によって衝突率を 0 にできることを示した．この結果より，DRP に MARL を適用する際に衝突させずに学習させることが可能となる．

今後の課題として，提案手法がうまく働かなかったマップでも，安全を確保しつつ効率的な移動ができるように改善する必要がある．手法の改善により，多様なマップ，多様な条件で安全を確保しつつ，効率的な移動ができるようになれば，実世界での運用に大きく近づくだろう．

謝辞

本研究は，日本学術振興会科学研究費（B）(21H03561， 24K03001）の補助を受けた．

参考文献

[1] 青山秀紀, 丁世堯, 林冬惠．ドローン配送計画最適化問題のための最短経路情報を利用したマルチエージェント強化学習．人工知能学会全国大会論文集, Vol. JSAI2022, pp. 3O3GS505–3O3GS505, 2022.

[2] Shiyao Ding, Hideki Aoyama, and Donghui Lin. MARL4DRP: Benchmarking Cooperative Multi-Agent Reinforcement Learning Algorithms for Drone Routing Problems. In *PRICAI 2023: Trends in Artificial Intelligence*, pp. 459–465, Singapore, 2024. Springer Nature Singapore.

[3] Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In William W. Cohen and Haym Hirsh, editors, *Machine Learning Proceedings 1994*, pp. 157–163. Morgan Kaufmann, San Francisco (CA), 1994.

[4] Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, and Stefano V. Albrecht. Benchmarking Multi-Agent Deep Reinforcement Learning Algorithms in Cooperative Tasks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS)*, 2021.

[5] Jinyoung Choi, Kyungsik Park, Minsu Kim, and Sangok Seok. Deep Reinforcement Learning of Navigation in a Complex and Crowded Environment with a Limited Field of View. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 5993–6000, 2019.

[6] Ingy ElSayed-Aly, Suda Bharadwaj, Christopher Amato, Rüdiger Ehlers, Ufuk Topcu, and Lu Feng. Safe multi-agent reinforcement learning via shielding. AAMAS '21, p. 483–491, Richland, SC, 2021. International Foundation for Autonomous Agents and Multiagent Systems.

[7] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. *The Journal of Machine Learning Research*, Vol. 21, No. 1, pp. 7234–7284, 2020.