

УСКОРЕНИЕ ОБМЕНА ДАННЫМИ ВОЗЛЕ ПРЕПЯТСТВИЙ ПРИ ГРУППОВОЙ МНОГОАГЕНТНОЙ ФУРАЖИРОВКЕ¹

Д.И. Зворыкин (*zuev.di@phystech.edu*)

П.С. Сорокоумов (*petr.sorokoumov@yandex.ru*)

Национальный исследовательский центр «Курчатовский институт»,
Москва

При решении задачи фуражировки группой автономных агентов весьма важно правильно организовать обмен информацией между участниками группы. В децентрализованной системе можно выполнять его путём попарных взаимодействий близкорасположенных агентов. В работе рассматривается роль препятствий и узких мест в формировании благоприятной для многочисленных обменов обстановки. Проведенное моделирование обучения с подкреплением групп агентов с разными организациями процессов обмена и при разных формах препятствий показало слабое воздействие на результаты случайно расположенных препятствий и более значительный эффект протяжённых преград. Роль обмена данными оказалась более высокой на начальном этапе работы системы и в условиях изменяющейся среды функционирования. Результаты работы могут использоваться для организации обмена данными при обучении групп агентов.

Ключевые слова: многоагентная фуражировка, обучение с подкреплением, обмен данными.

Введение

Задача фуражировки — одна из типичных проблем для систем групповой робототехники, имеющая множество теоретических и практических приложений. Среди прочих важных факторов её эффективности особую роль играет механизм обмена информацией. При децентрализованном управлении знания каждого отдельного агента о среде ограничены, по-

¹ Работа выполнена за счет государственного задания НИЦ «Курчатовский институт».

этому обмен позволяет активнее изучать окружение и эффективнее планировать действия. Логично предположить, что повышение числа обменов может в отдельных случаях ускорить оптимизацию поведения. В данной работе путём моделирования агентов, обучающихся с подкреплением по различным алгоритмам, оценивается ускорение обмена, получаемое при введении в окружение препятствий, повышающих интенсивность трафика в малых областях пространства.

На практике такая конфигурация часто встречается в постройках социальных насекомых – муравьёв и термитов. Узкие проходы внутри их построек, а также проложенные по грунту «дороги» позволяют многим особям постоянно встречаться и, потенциально, воздействовать друг на друга. Весьма полезно было бы оценить влияние такого сосредоточения на распространение информации внутри децентрализованной группы. Сделанные выводы можно применять в искусственных системах, выполняющих сходную с фуражировкой работу, чтобы оптимизировать их функционирование. Понятно, что объяснять таким образом поведение насекомых было бы неправильно, потому что биологические системы разнообразнее и сложнее абстрактных моделей, но во многих задачах робототехники полученные результаты могли бы оказаться полезными.

Далее в работе сначала приведён краткий обзор подходов к обмену данными между агентами децентрализованных систем, затем описаны варианты организации взаимодействия и постановка задачи обучения с подкреплением. Выполненное моделирование процессов обучения в разных условиях позволило оценить влияние разных вариантов препятствий на эффективность.

1. Методы обмена данными между агентами децентрализованных систем

Проблематика обмена данными между агентами при обучении рассматривалась неоднократно, при этом отмечалась значительная сложность этой задачи. Даже коммуникация между людьми, которые способны планировать свои действия для достижения общей цели и объяснять решения друг другу, оказывается весьма проблематичной [Wu et al., 2025]. В обзоре задач, стоящих перед разработчиками мультиагентных систем [Wong et al., 2023], коммуникация указана среди наиболее интересных проблем наряду с комбинированием централизованного обучения и децентрализованного исполнения, моделированием оппонента, групповой координацией и формированием функции награды. Задача коммуникации при этом может ставиться в виде обучения протоколу взаимодействия [Zhu et al., 2024]. Его выполняют разными способами, например на глубокой Q-сети общего вида [Foerster et al., 2016] или на более специализированной – трансформере [Yang et al., 2022]. При этом создать общий протокол взаи-

модействия для всех агентов весьма сложно, потому что децентрализация системы запутывает процедуру его согласования. Для целей данной работы обучение протоколу представляется излишне затратным, хотя оно оправдано при помехах в каналах связи, проблемах с идентификацией и защитой информации [Ahmed et al., 2024].

Решением проблем повторного использования полученных знаний в рамках обучения с подкреплением занимается отдельная область исследований – обучение путём переноса (transfer learning). В обзоре [Silva et al., 2019] описаны многие проверенные подходы данной области, например метрики успешности переноса знаний и методы их повторного использования как внутри агентов, так и при коммуникации между ними. Среди вариантов передаваемых между агентами знаний упоминаются функции награды, политики, оценки чужих состояний, советы по выбору действий (т.е. действие, выбранное бы агентом, если бы его состояние было тем же, что у партнёра по коммуникации), опыт выбора действий (тройки из состояний, выбранных действий и полученных за них наград) и др. Нарботки такого рода вполне применимы к решаемой задаче [Azmani et al., 2023].

Также хорошо известно использование среды для неявной коммуникации путём оставления следов (stigmergy) [Shaw et al., 2022], но его сложнее воплотить из-за технических проблем с реализацией требуемых следов на реальных роботах.

В целом многие механизмы обмена данными между фуражирующими агентами хорошо известны и проверены, но отношения между их эффективностью и расположением препятствий в среде изучены недостаточно.

2. Организация межагентного взаимодействия

При обмене данными между агентами хотелось бы найти оптимальный баланс между полнотой передаваемых данных, затратами на коммуникацию, скоростью и качеством распространения информации. На данный момент выделяют несколько семейств методов, направленных на достижение этой цели, которые условно можно разделить на три основные группы.

К реактивным (стигмергическим) подходам относятся механизмы косвенной координации через изменение среды. Агенты оставляют в среде маркеры, которые влияют на поведение других агентов, что создает обратную связь, основанную на принципах самоорганизации через усиление/затухание маркеров. Вычислительная сложность таких решений минимальна. Можно выделить такие подклассы следов, как феромоны, градиентные поля, трассировки. Агент, обнаружив ресурс, оставляет феромонный след с интенсивностью, пропорциональной качеству ресурса; получаемое градиентное поле обновляется по некоторому заданному правилу. Другие агенты двигаются вдоль градиента поля и корректируют маршрут при обнаружении более сильных сигналов.

К рыночным механизмам относится распределение задач через аукционы с применением рассчитанной на основе локальной информации предполагаемой награды как «валюты». Разрешение конфликтов можно реализовать, например, через модифицированный аукцион Викри [Vickrey, 1961], в котором сравниваются «общественные полезности» без агента и при его участии.

Третий подход основан на обмене обученными моделями и их частями. Агенты, использующие машинное обучение, совместно корректируют политики, периодически обмениваясь параметрами моделей для передачи опыта. Можно организовать, например, обмен Q-значениями (для табличных методов) или тройками <состояние, действие в нём, полученная награда>. В данной работе применяется именно этот принцип.

Возможно и сочетание указанных подходов, например использование феромонов для быстрого реагирования на важные события и обучения с подкреплением для долгосрочного планирования.

Проанализируем эффективность обмена данными на примере группы агентов, выполняющих фуражировку на клеточном поле (рис. 1). В центральной ячейке среды находится база агентов, в которую они должны доставлять пищу из её источников. Если агент войдёт на ячейку источника, он получит пищу, количество которой – случайная величина с заданным распределением. Эти распределения со временем не меняются (т.е. пища в источниках не иссякает), но у разных источников параметры распределений различны, причём агенты заранее их не знают. Полученную пищу агенты доставляют на базу, после чего снова двигаются к тому же или другому источнику.

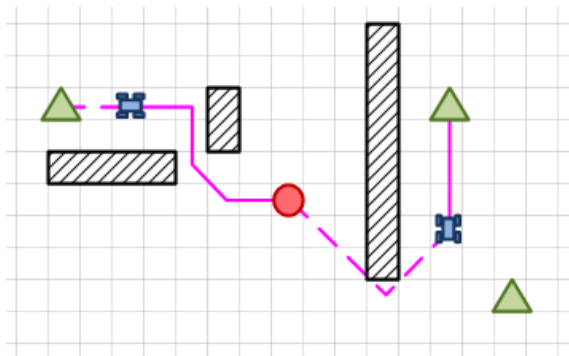


Рис. 1. Среда функционирования агентов: клеточное поле, по которому агенты-роботы доставляют пищу из источников (зелёных треугольников) к базе (красному кругу), избегая препятствий (заштрихованных прямоугольников)

Чтобы сосредоточиться на высокоуровневых взаимодействиях агентов, будем далее считать, что агенты знают заранее координаты всех источников (но не их обильность) и способны проложить маршрут к каждому из них, если препятствия (непроходимые ячейки) это позволяют, и маршрут не блокируется другими агентами. Это значит, что в задаче обучения с подкреплением в качестве действий рассматриваются не элементарные движения, а целые процедуры сбора пищи из источника. Каждая из них включает движение к источнику с базы по кратчайшему маршруту, получение пищи, движение обратно на базу и автоматическая выгрузка пищи. Количество этих действий равно числу источников пищи, доступных с базы.

Без обмена данными можно было бы свести задачу в такой постановке к простейшему случаю «многорукого бандита», но обмен существенно меняет ситуацию. Будем считать, что обмен между двумя агентами происходит, если они находятся поблизости друг от друга, т.е. расстояние между ними не превышает некоторый фиксированный порог. Обмен выполняется мгновенно, без дополнительных затрат времени, и не является отдельным действием, которое агент должен был бы намеренно выбирать. Это снижает гибкость системы, но устраняет проблемы при изучении агентом протокола взаимодействия: обмен данными и последующие дообучение моделей происходят в фоновом режиме, не прерывая выполнение текущей поведенческой процедуры агента.

Таким образом, действиями в задаче обучения с подкреплением являются комплексные поведенческие процедуры сбора пищи из источников; число таких действий равно числу источников пищи. Наградой агента является среднее число добытой пищи за такт моделирования, т.е. число собранной пищи, разделённое на длину маршрута. В состояние агента включаются текущие оценки обильности источников. При каждом акте обмена данными любые агенты передают друг другу координаты и обильность последнего посещенного участка, сведения о ранее совершённых действиях (прежнее состояние, действие, награда, следующие состояние) и текущие параметры своих моделей (Q-таблицы – при их наличии).

Это позволяет агентам корректировать свои политики. Если, например, при Q-обучении значение Q для некоторых аргументов у текущего агента было равно Q_1 , а у другого агента Q_2 , то новое значение будет получено по формуле:

$$Q_{\text{нов}} = \frac{Q_1 + Q_2}{2},$$

где Q_1 – скорость внесения корректив, N_1 – число посещений соответствующего участка текущим агентом, N_2 – число посещений его же другим агентом. Это значит, что степень сдвига зависит от относительного опыта агентов в совершении конкретного действия.

Если один агент мешает другому пройти к нужной точке, применяется трехуровневый протокол разрешения конфликтов: 1) агенты с грузом получают приоритет, 2) в узких проходах используется правило правой руки для расхождения, 3) при невозможности разрешения конфликта один из агентов случайно отступает на свободную клетку. Ситуации полной блокировки предотвращаются принудительным сбросом цели при превышении лимита шагов миссии.

Таким образом, задача обучения с подкреплением формализована следующим образом:

- состояние агента – две координаты, количество имеющейся пищи, расстояние до базы, номер группы агентов, текущий градиент (для глубокого Q-обучения);
- набор допустимых действий – перемещения к каждому из кормовых участков с автоматическим сбором пищи и её доставкой на базу;
- награда при доставке пищи на базу – количество пищи, делённое на пройденное расстояние;
- методы обмена данными: дообучение каждого агента на последнем действии другого (как если бы он сам совершил это действие в том же состоянии и с тем же результатом) и, для классического Q-обучения, усреднение Q-значений описанным выше способом. Обмен выполняется каждый такт для каждой пары, сблизившейся на расстояние меньше порогового.

Выбранный метод обращения к алгоритму обучения с подкреплением осложняет применение готовых средств обучения многоагентных систем (PettingZoo и подобных), потому что в них, как правило, обращение к обучающему алгоритму происходит каждый шаг моделирования. К данной задаче из-за сочетания непрерывного обмена информацией и дообучения в произвольные моменты применение таких подходов затруднено, поэтому пришлось реализовывать небольшую собственную среду многоагентного моделирования, поддерживающую алгоритмы обычного и глубокого Q-обучения:

- для классического Q-обучения применялись скорость обучения $\alpha = 0.1$, коэффициент дисконтирования $\gamma = 0.95$; вероятность исследования ϵ меняется от 1.0 вплоть до 0.1, уменьшаясь каждый шаг в 0.997 раз;
- глубокое Q-обучение использовало для предсказания Q-функции персептрон с двумя скрытыми слоями по 64 нейрона с функциями активации ReLU, вычисляющий прогнозируемое Q-значение для каждого действия; использован оптимизатор Adam со скоростью обучения 0.0001.

В разработанном ПО реализованы также алгоритмы глубокого обучения PPO и PolicyGradient. Поскольку предварительные тестовые измерения показали, что они обучаются значительно медленнее Q-обучения, детальные исследования их поведения не проводились.

3. Моделирование и обсуждение результатов

Если взаимодействия между агентами ограничены некоторой дальностью связи, то целесообразно оценить влияние разных препятствий на расстояние между ними. Наличие большого количества препятствий в среде заставит агентов обходить их, что приведёт к появлению зон повышенной плотности агентов. Так как конфигурации препятствий могут быть весьма разнообразными, оценим влияние только двух простейших вариантов: случайно расположенных равномерно распределённых препятствий и сплошных непроходимых стенок с маленькими проходами. Некоторые другие конфигурации можно получить как комбинацию двух указанных, чтобы распространить на них сделанные выводы.

Чтобы оценить расстояния между агентами в разных окружениях, рассмотрим группу из двух агентов, которые независимо друг от друга перемещаются по клеточной карте из базы к случайно выбранной свободной точке и обратно. При этом возможные пункты назначения лежат внутри некоторой области досягаемости, скорости агентов одинаковы, задержек в пунктах назначения нет, и агенты не препятствуют движению друг друга (считаются точечными). Для сбора данных о расстояниях была реализована вспомогательная имитационная модель. Оказалось, что плотность распределения вероятности расстояния между агентами имеет вид, показанный на рис. 2. При этом доля времени, когда агенты могут обмениваться информацией, определяется как часть площади под полученной кривой, ограниченная порогом дальности коммуникации.

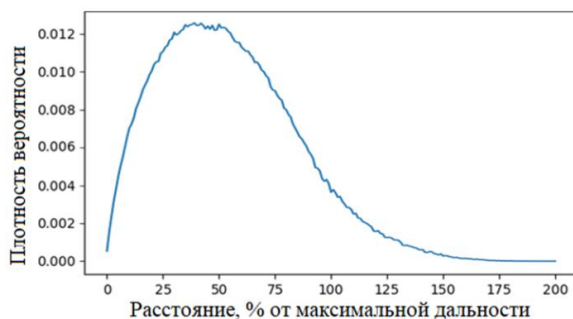


Рис. 2. Плотность распределения расстояния между агентами при отсутствии препятствий. Максимальное удаление одного робота от базы принято за 100%

Введение препятствий модифицирует данное распределение, потому что агенты располагаются друг к другу в среднем несколько ближе. При случайно расположенных препятствиях кривая слегка смещается к малым значениям, сохраняя форму (рис. 3). При этом росту вероятности появле-

ния препятствия в каждой ячейке с 0 до 0.5 соответствует падение медианного расстояния между агентами от 54% до 52% от дальности хода робота. Таким образом, влияние таких препятствий на расстояние слабо.

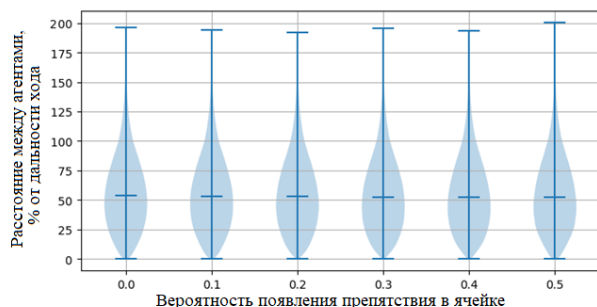


Рис. 3. Плотность распределения расстояния между агентами при независимо расположенных препятствиях; размер карты – 100х100 ячеек, дальность хода агента – 50 ячеек

Если препятствия образуют сплошную стенку с одним проходом вокруг базы, форма распределения существенно меняется (рис. 4).

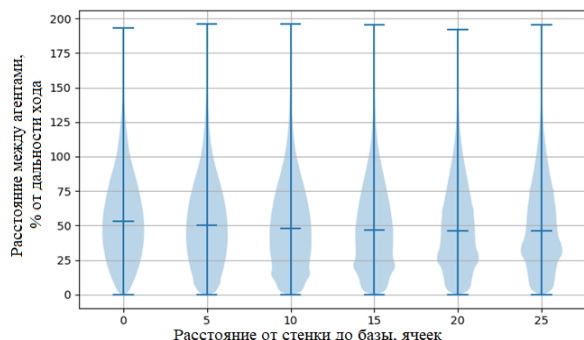


Рис. 4. Плотность распределения расстояния между агентами при препятствиях, образующих вокруг базы квадрат со стороной 20 ячеек с одним отверстием посередине одной стороны; размер карты – 100х100 ячеек, дальность хода агента – 50 ячеек

Видно, что медианное значение уменьшается за счёт повышения вероятности малого расстояния между агентами. При этом доля самых малых расстояний существенно растёт: на карте без препятствий роботы находились на расстоянии меньше 5 ячеек с вероятностью 1.5%, при наличии стенки в 10 ячейках от базы – с вероятностью 4%.

С учётом полученных результатов было проведено моделирование обучения многоагентных групп, участники которых автоматически обмениваются информацией при сближении. Квадратная карта имела сторону в 150 ячеек, кормовых участков на ней – 16, дальность связи – 5 ячеек. Обмены ускоряют сбор пищи даже без препятствий (рис. 5).

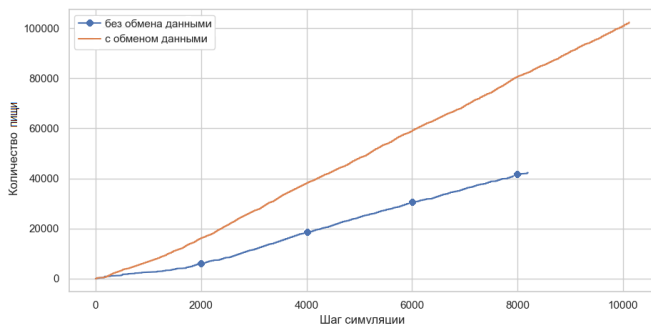


Рис. 5. Сравнение производительности групп с обменом данными и без него. Препятствия отсутствуют. В модели участвовали два робота, обучающиеся по алгоритму глубокого Q-обучения; среднее количество пищи в участках распределено равномерно от 1 до 100

Также при наличии обменов система быстрее находит наилучший участок из имеющихся (рис. 6) и активнее его эксплуатирует.

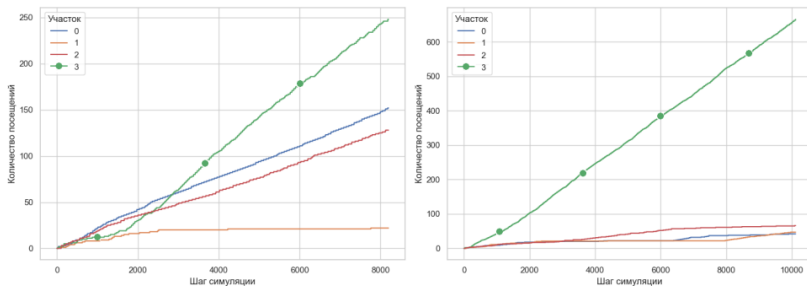


Рис. 6. Количество посещений агентами каждого из четырёх случайно расположенных кормовых участков. Один из них (№3) обильнее других; обмен информацией позволяет быстрее определить его. В модели участвовали два робота, применяющие алгоритм классического Q-обучения с обменом данными

Моделирование работы нескольких алгоритмов обучения с подкреплением в среде со случайно расположенными препятствиями показало, что при небольшой доле препятствий (до 40%) результаты у всех реализованных алгоритмов менялись незначительно, при большем числе препятствий – несколько падали из-за удлинения путей к пище.

Моделирование среды с протяжённым препятствием - сплошной стеной в 10 ячеек вокруг базы с одним малым отверстием – показало, что несмотря на рост числа обменов данными, производительность агентов в целом упала (рис. 7). Причиной этого стали большие расстояния, которые пришлось проходить для сбора пищи, и временные затраты агентов для разрешения коллизий. Это значит, что интенсивный обмен данными помог системе приспособиться к среде, но одновременно препятствия ухудшили её свойства, удлинив маршруты до цели.

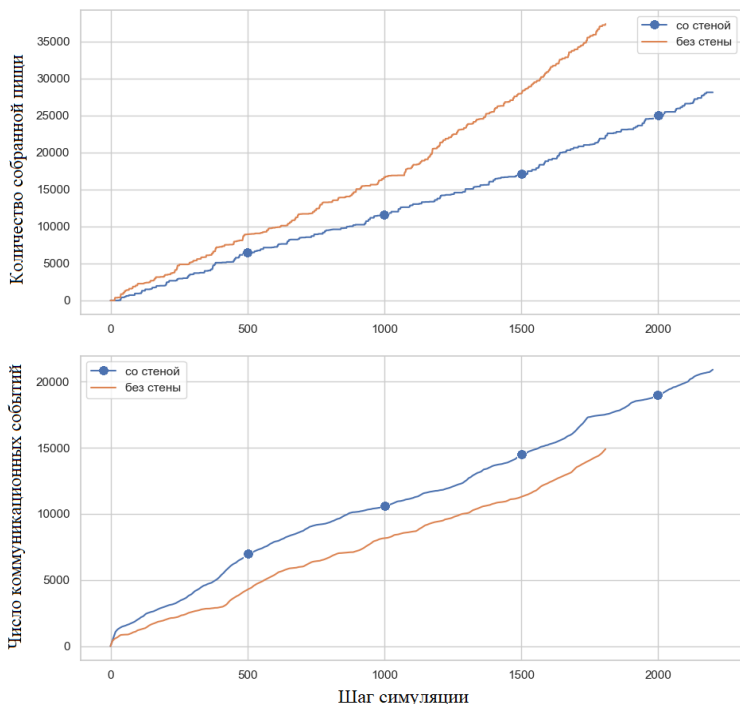


Рис. 7. Количество собранной пищи (сверху) и число обменов данными (внизу) при сборе пищи в среде с протяжённым препятствием - сплошной стенкой вокруг базы с одним проходом. Использован алгоритм глубокого Q-обучения

Таким образом, повышение интенсивности обмена желательно проводить менее жёсткими методами; для поставленной задачи это, например, использование не одного узкого прохода, а четырёх, расположенных в центрах сторон квадратного ограждения. Кроме того, при постоянных параметрах среды обмен играет роль только до нахождения всеми агентами оптимальной стратегии, поэтому можно ожидать, что его роль сильнее

проявится в изменяющейся среде. Для проверки этой гипотезы было проведено моделирование в условиях, когда обилие источников периодически менялось: количество пищи, которое мог получить каждый агент из источника на шаге t , при этом равнялось

$$\frac{A}{T} \sin(\varphi + \frac{2\pi t}{T}),$$

где A – случайно выбранное для каждого источника в начале моделирования максимальное значение, φ – случайная фаза из диапазона $[0, 2\pi)$, также своя для каждого источника; $T=1000$. Если стена достаточно сильно удалена от базы (на 20 ячеек), оказалось, что на начальном этапе скорость сбора пищи с препятствиями превышает скорость без препятствий (рис. 8). Связано это с большим временем, проводимым агентами вместе, и с большими их затратами на движение к далёкой пище.

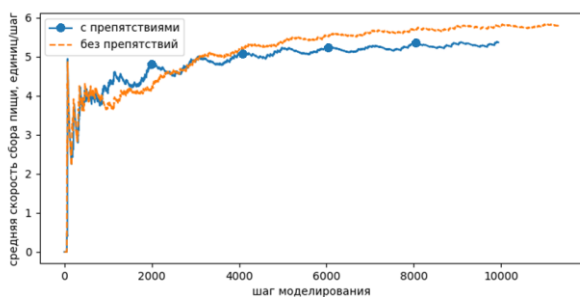


Рис. 8. Скорость сбора пищи при меняющейся среде и 4 отверстиях в ограждении.

Видно, что на отрезке от 1000 до 3000 шагов препятствия увеличивают её, далее – снижают

Исследование показало, что характер получаемых результатов – как по скорости обучения, так и по эффективности системы, – для алгоритмов классического и глубокого Q-обучения совпадает.

Заключение

Таким образом, было показано, что препятствия могут интенсифицировать обмен данными, если для него требуется сближение агентов. Более эффективны при этом протяжённые преграды, тогда как при их случайном расположении эффект незначителен. Иногда после этого повышается общая эффективность работы многоагентной системы. Однако наличие чрезмерно больших препятствий слишком сильно ухудшает свойства среды, затрудняя работу, поэтому следует тщательно планировать положение мест скопления агентов. Также при малых дальности локальной связи весьма важен будет эффективный механизм расхождения агентов в узких проходах, который в данной работе детально не исследовался.

Список литературы

- [Ahmed et al., 2024] Ahmed M.H., Ghasemi M. Privacy-preserving decentralized actor-critic for cooperative multi-agent reinforcement learning // Proceedings of The 27th International Conference on Artificial Intelligence and Statistics Proceedings of Machine Learning Research / eds. S. Dasgupta, S. Mandt, Y. Li.: PMLR, 2024. – P. 2755-2763.
- [Azmani et al., 2023] Azmani H. et al. Cooperative foraging behaviour through multi-agent reinforcement learning with graph-based communication // Sixteenth European Workshop on Reinforcement Learning. Vrije Universiteit Brussel, Brussels, Belgium, 2023.
- [Foerster et al., 2016] Foerster J. et al. Learning to communicate with deep multi-agent reinforcement learning // NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems, 2016. – P. 2145-2153. – DOI: 10.5555/3157096.3157336.
- [Shaw et al., 2022] Shaw S. et al. ForMIC: Foraging via multiagent RL with implicit communication // IEEE Robot. Autom. Lett. – 2022. – Vol. 7. – P. 4877-4884. – DOI: 10.1109/LRA.2022.3152688.
- [Silva et al., 2019] Silva F., Costa A. A Survey on Transfer Learning for Multiagent Reinforcement Learning Systems // J. Artif. Intell. Res. – 2019. – Vol. 64. – DOI: 10.1613/jair.1.11396.
- [Vickrey, 1961] Vickrey W. Counterspeculation, Auctions, and Competitive Sealed Tenders // J. Finance. – 1961. – Vol. 16, Issue 1. – P. 8-37. – DOI: 10.2307/2977633.
- [Wong et al., 2023] Wong A. et al. Deep multiagent reinforcement learning: challenges and directions // Artif. Intell. Rev. – 2023. – Vol. 56, Issue 6. – P. 5023-5056. – DOI: 10.48550/arXiv.2106.15691
- [Wu et al., 2025] Wu C. M. et al. Adaptive mechanisms of social and asocial learning in immersive collective foraging // Nat. Commun. – 2025. – Vol. 16, Issue 1. – P. 1-15. – DOI: 10.1038/s41467-025-58365-6.
- [Yang et al., 2022] Yang Y. et al. Transformer-based working memory for multiagent reinforcement learning with action parsing // NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems. Article. – 2022. – No. 2527. – P. 34874-34886.
- [Zhu et al., 2024] Zhu C., Dastani M., Wang S. A survey of multi-agent deep reinforcement learning with communication // Autonomous Agents and Multi-Agent Systems. – 2024. – Vol. 38, Issue 1. – P. 2845-2847. – DOI: 10.1007/s10458-023-09633-6.