УДК 004.855
EDN: TTWHJB

Экспериментальное исследование мультиагентных подходов к обучению с подкреплением в задаче планирования пути покрытия

Луканов С. Ю. ✉, **Тимошевская О. Ю.**

Псковский государственный университет,
Псков, 180000, Российская Федерация

Постановка задачи. *Задача планирования пути покрытия группами автономных агентов определенной целевой области является актуальной для широкого класса прикладных систем. Рост размерности среды и числа взаимодействующих агентов приводит к усложнению процессов координации и увеличению времени достижения полного покрытия. Дополнительную сложность вносит ограниченность наблюдений агентов, в результате чего задача управления формулируется в условиях частичной наблюдаемости и стохастической динамики. В связи с этим весьма востребованной является разработка специализированных подходов к интеллектуальному управлению, обеспечивающих минимизацию времени покрытия целевой области при децентрализованном принятии решений.* ***Целью исследования*** *является разработка инструментария, который позволяет обучать модели управления группой однородных автономных агентов в задаче планирования пути покрытия целевой области, обеспечивающих минимизацию математического ожидания времени достижения полного покрытия в условиях частично наблюдаемого марковского процесса принятия решений.* ***Методы.*** *Для формализации задачи используется модель частично наблюдаемого марковского процесса принятия решений, включающая описание состояний системы, пространств действий и наблюдений агентов, вероятностной динамики среды и функции вознаграждения. Решение задачи синтеза групповой стратегии основано на методах глубокого обучения с подкреплением для мультиагентных систем, ориентированных на децентрализованное исполнение при обучении с централизованным критиком. Для оценки эффективности применяются методы имитационного моделирования в дискретном двухмерном клеточном пространстве.* ***Новизна.*** *Разработана унифицированная экспериментальная среда для сопоставления мультиагентных архитектур в задаче планирования пути покрытия целевой области. Показано, что использование карт малого размера ограничивает статистическую значимость ряда метрик координации агентов, что обосновывает переход к картам большего размера. Выявлены типовые источники деградации качества покрытия, связанные с граничными эффектами и малыми фрагментами целевой области, и предложена модификация среды, снижающая влияние указанных факторов.* ***Результаты.*** *Предложен подход к синтезу групповой стратегии управления, позволяющий обеспечить достижение полного покрытия целевой области за конечное время. Проведенное моделирование подтверждает возможность эффективной координации действий агентов и сокращения времени покрытия по сравнению с некоординированными стратегиями при сохранении децентрализованного характера управления. Экспериментальные исследования показали различия в динамике покрытия и координации агентов для рассматриваемых архитектур.* ***Теоретическая значимость*** *работы заключается в развитии методов формализации и решения задач мультиагентного покрытия в условиях частичной наблюдаемости.* ***Практическая значимость*** *определяется возможностью применения полученных результатов при разработке интеллектуальных систем управления группами автономных мобильных агентов, в том числе для*

Библиографическая ссылка на статью:
Луканов С. Ю., Тимошевская О. Ю. Экспериментальное исследование мультиагентных подходов к обучению с подкреплением в задаче планирования пути покрытия // Вестник СПбГУТ. 2026. Т. 4. № 1. С. 2. EDN: TTWHJB

Reference for citation:
Lukanov S., Timoshevskaya O. Experimental Study of Multi-Agent Reinforcement Learning Approaches for the Coverage Path Planning Problem // Herald of SPbSUT. 2026. Vol. 4. Iss. 1. P. 2. EDN: TTWHJB

задач мониторинга, разведки и робототехнических систем. Полученные результаты могут быть использованы при проектировании мультиагентных систем покрытия и при сравнительном анализе методов управления агентами в сложных дискретных средах.

Ключевые слова: *глубокое обучение с подкреплением, система управления, планирование пути покрытия, мультиагентная система, искусственный интеллект*

Актуальность

Задача планирования пути покрытия целевой области группами автономных агентов занимает важное место в современных исследованиях в области интеллектуальных систем управления и мультиагентных технологий. Она является востребованной для многих прикладных систем, реализующих мониторинг, разведку, поисково-спасательные операции, картографирование и обслуживание распределенной инфраструктуры. Повышение автономности таких систем и снижение зависимости от централизованных средств управления являются ключевыми требованиями при их практическом применении.

Современные тенденции развития автономных робототехнических систем характеризуются ростом размерности среды, усложнением ее структуры и увеличением числа одновременно действующих агентов. Это приводит к существенному усложнению процессов координации и росту вычислительной сложности алгоритмов управления. Дополнительным сдерживающим фактором является частичная наблюдаемость среды отдельными агентами, обусловленная ограниченными возможностями сенсоров и отсутствием глобальной информации, что требует формализации задачи управления в виде частично наблюдаемого марковского процесса принятия решений.

Классические методы планирования пути покрытия, основанные на детерминированных алгоритмах, эвристиках или централизованных подходах, плохо масштабируются при увеличении числа агентов и размерности пространств состояний и действий, а также слабо адаптируются к стохастической динамике и задачам с неполной информацией о среде. В связи с этим все большее внимание уделяется методам интеллектуального управления, в частности подходам глубокого обучения с подкреплением, позволяющим автоматически синтезировать стратегии поведения агентов на основе взаимодействия со средой.

Несмотря на активное развитие мультиагентного обучения с подкреплением, сохраняется ряд нерешенных проблем, связанных с корректной постановкой задачи покрытия, сопоставимостью экспериментальных результатов, выбором адекватных метрик эффективности и влиянием параметров экспериментальной среды на динамику координации агентов. В частности, использование сред малого размера может искажать статистические характеристики процесса покрытия и не отражать особенности коллективного поведения агентов, что снижает обобщающую способность получаемых выводов.

В связи с вышеизложенным актуальными являются разработка и исследование унифицированной экспериментальной среды и подхода к синтезу групповых стратегий управления, обеспечивающих децентрализованное принятие решений, устойчивую координацию агентов и минимизацию времени достижения полного покрытия в условиях частичной наблюдаемости и стохастической динамики. Решение данных задач имеет как теоретическое значение для развития методов мультиагентного обучения с подкреплением, так и практическую ценность для создания интеллектуальных систем управления автономными агентами в сложных дискретных средах.

Постановка задачи

Рассмотрим задачу планирования пути покрытия целевой области группой автономных агентов с дискретным временны́м шагом. Среда моделируется в виде конечного двумерного клеточного пространства $\mathcal{X} \subset \mathbb{Z}^2$, содержащего множество целевых ячеек $\mathcal{X}^* \subseteq \mathcal{X}$, которые требуется покрыть. В системе действует N однородных агентов, каждый из которых в момент времени $t \in \mathbb{N}$ занимает одну ячейку $x_t^i \in \mathcal{X}$ и обладает фиксированным радиусом покрытия ρ, определяющим локальную зону охвата. Динамика системы описывается как частично наблюдаемый марковский процесс принятия решений, заданный кортежем:

$$\langle S, U, Z, P, O, R, \gamma \rangle,$$

где S – множество глобальных состояний, включающих совокупность позиций всех агентов и текущее состояние покрытия целевой области; U – множество допустимых действий агента; Z – пространство локальных наблюдений агента; P – функция переходов $P(s' \mid s, \boldsymbol{u})$, задающая вероятностную динамику среды при совместном векторе действий $\boldsymbol{u} = (u^1, u^2 \ldots u^N)$; O – функция наблюдений $O(z^i \mid s, u^i)$, которая определяет вероятность получения i-м агентом наблюдения z^i в состоянии s; R – функция вознаграждения; $\gamma \in (0,1]$ – коэффициент дисконтирования.

Целью является нахождение групповой стратегии, которая отображает историю наблюдений агента в распределение над действиями, таких что математическое ожидание времени достижения полного покрытия множества минимизируется.

Инструментарий

Для проведения вычислительных экспериментов и получения результатов, представленных в данной работе, был разработан и использован специализированный программный комплекс [1], предназначенный для имитационного моделирования системы «агент – среда», обучения агентов с применением методов глубокого обучения с подкреплением, а также последующего анализа и визуализации показателей эффективности.

Программный комплекс имеет модульную архитектуру и включает три логически взаимосвязанных подсистемы:

– имитационного моделирования и обучения агентов;

– управления экспериментами и конфигурациями;

– расчета и визуализации метрик эффективности.

Ключевым компонентом комплекса является программная среда, реализованная на базе платформы Unity версии 2023.2.18f1 с использованием пакета Unity ML-Agents версии v2.3.0-exp.3. Данная подсистема обеспечивает моделирование задач покрытия дискретной целевой области агентами с различными архитектурными решениями. В рамках среды реализована ее абстрактная базовая модель, определяющая общие механизмы генерации карт, подсчета покрытой площади, вычисления вознаграждений и управления эпизодами обучения.

На основе базового класса реализованы специализированные среды для одноагентного, конкурентного, кооперативного и централизованного обучения. Каждая из сред поддерживает параметризуемое количество агентов, различные размеры карт, нормализацию вознаграждений. Поведение агентов описывается дискретным пространством действий, а процесс взаимодействия со средой формализован в виде частично наблюдаемого марковского процесса принятия решений.

Для обеспечения воспроизводимости экспериментов в среде реализованы механизмы детерминированной инициализации эпизодов, записи временны́х рядов покрытия целевой области, а также сохранения траекторий движения агентов. Сбор статистик осуществляется встроенными модулями, которые формируют выходные файлы в специализированных форматах с целью последующего анализа. Для автоматизации подготовки и запуска вычислительных экспериментов разработано отдельное программное средство – менеджер конфигураций [2], реализованный на языке C# с использованием платформы .NET Framework 4.7.2 и фреймворка Windows Forms. Данное приложение обеспечивает централизованное управление параметрами обучения, среды моделирования и экспериментальных сценариев.

Менеджер конфигураций предоставляет графический интерфейс для редактирования параметров обучения, параметров среды моделирования, а также конфигураций алгоритмов обучения. Приложение поддерживает сохранение, импорт и экспорт конфигураций, контроль согласованности параметров между различными файлами настроек и автоматическую генерацию сценариев запуска. Анализ результатов обучения осуществляется с помощью специально разработанной программы расчета и визуализации показателей эффективности [3], реализованной на языке Python с использованием фреймворка PyQt5 версии 5.9.2. Подсистема предназначена для обработки файлов статистик покрытия и траекторий, собранных в процессе тестирования обученных моделей.

Архитектура программы разделена на два уровня: функциональный и интерфейсный. Первый уровень обеспечивает чтение и агрегацию данных, вычисление метрик эффективности и формирование матриц попарных сравнений между моделями. Реализованный механизм регистрации метрик позволяет расширять набор показателей без изменения основной логики программы. Второй уровень предоставляет интерактивный графический интерфейс для отображения временны́х графиков покрытия, маршрутов агентов, тепловых карт значений метрик и матриц сравнений.

Подсистема визуализации обеспечивает наглядный сравнительный анализ различных алгоритмов и режимов обучения, что существенно упрощает интерпретацию экспериментальных результатов и повышает достоверность выводов, полученных в работе. Интерфейс программы представлен на рисунке 1.

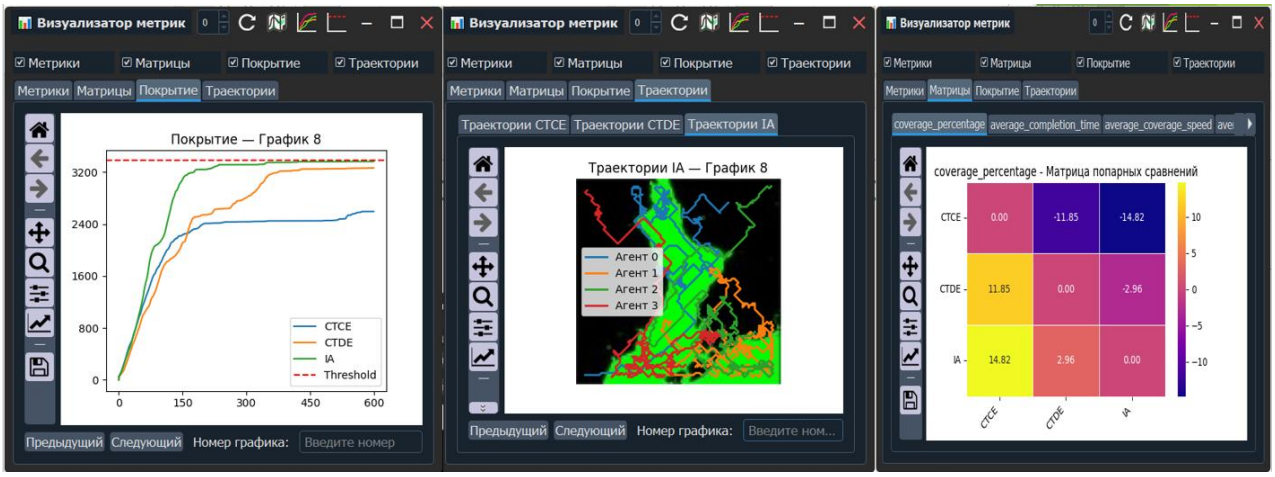

Рис. 1. Интерфейс подсистемы визуализации и расчета метрик эффективности и траекторий агентов

Тестируемые реализации мультиагентных архитектур

В рамках исследования были рассмотрены основные концепции архитектурного построения систем мультиагентного обучения с подкреплением. Наибольшее распространение в современной литературе получили три подхода: обучение независимых агентов (IA, *аббр. от англ.* Independent Agents), централизованное обучение с децентрализованным выполнением (CTDE, *аббр. от англ.* Centralized Training with Decentralized Execution) и централизованное обучение с централизованным выполнением (CTCE, *аббр. от англ.* Centralized Training with Centralized Execution).

Подход IA предполагает автономное обучение каждого агента без явного учета структуры мультиагентного взаимодействия: остальные агенты воспринимаются как часть окружающей среды. Данный метод отличается простотой реализации, однако характеризуется повышенной нестационарностью среды, что может негативно сказываться на сходимости процесса обучения. В архитектуре CTDE этап обучения осуществляется с привлечением глобальной информации о состоянии системы и действиях всех агентов, тогда как на этапе исполнения каждая стратегия функционирует на основе локальных наблюдений. Такой компромисс позволяет повысить качество координации при сохранении масштабируемости системы.

Подход CTCE использует полностью централизованную стратегию как во время обучения, так и при выполнении действий акторами, что потенциально обеспечивает более высокую эффективность в задачах, требующих плотного взаимодействия агентов, но существенно ограничивает возможность применения в распределенных и ресурсно-ограниченных системах [4, 5]. В качестве базового алгоритма обучения использовался метод PPO (*аббр. от англ.* Proximal Policy Optimization – оптимизация проксимальной политики), обладающий устойчивостью к резким обновлениям стратегии и хорошо зарекомендовавший себя в задачах управления с дискретными и непрерывными пространствами действий [6, 7].

Для реализации архитектуры CTDE был выбран алгоритм MA-POCA (*аббр. от англ.* Multi-Agent POsthumous Credit Assignment – мультиагентное апостериорное распределение заслуг), представляющий собой развитие метода COMA (*аббр. от англ.* Counterfactual Multi-Agent – мультиагентный контрфактический метод). В MA-POCA применяется механизм внимания для формирования совместного

представления наблюдений, что обеспечивает адаптацию архитектуры к переменному числу кооперирующих агентов. Алгоритм использует централизованного критика и децентрализованных акторов, а также контрфактическую базисную линию, позволяющую оценивать индивидуальный вклад агента в общее вознаграждение за счет исключения его действия при вычислении Q-функции [8]. Такой подход способствует формированию согласованных кооперативных стратегий в мультиагентной системе.

Экспериментальное тестирование

Для обучения агентов был создан набор из предварительно обработанных спутниковых карт, на которых водные массивы – целевые зоны, помечены зеленым цветом. Каждый агент обладает фиксированным радиусом покрытия и получает визуальные наблюдения с двух источников: глобальной камеры, охватывающей всю карту, и локальной камеры, фиксирующей окрестность агента. На каждом временно́м шаге агент может осуществлять перемещение на одну ячейку в одном из восьми возможных направлений. Функция вознаграждения подвергается нормализации. При попытке выхода за пределы допустимой области агенту начисляется штраф, после чего текущий эпизод завершается. Оценка качества обученных стратегий выполняется на отдельном наборе тестовых карт, не участвовавших в процессе обучения. Параметры конфигурации обучающей среды представлены в таблице 1.

Таблица 1. Параметры конфигурации обучающей среды

Параметр	Значение	Пояснение
cameraResolution	300	Разрешение камеры для сбора визуальных наблюдений в пикселях
mapSize	100	Размер стороны карты в пикселях
wallPenalty	0,5	Штраф за столкновение с границей области
discoveryReward	20	Награда за покрытие одной ячейки
timePenalty	0,002	Штраф за один временно́й шаг
visionRange	3	Радиус покрытия области агентом в пикселях
maxSteps	600	Максимальное число шагов в эпизоде
resetOnWallHit	True	Прерывание эпизода при выходе агента за границу
numAgents	4	Количество агентов
rewardNormalization	True	Нормализация награды за покрытие

Для обучения агентов в сценариях IA и CTCE применялся алгоритм PPO, тогда как в конфигурации CTDE использовался алгоритм MA-POCA. С целью обеспечения корректного сравнительного анализа значения гиперпараметров алгоритмов были унифицированы для всех экспериментов (таблица 2).

Таблица 2. Гиперпараметры алгоритмов обучения

Гиперпараметр	Значение
batch_size	256
buffer_size	2560
learning_rate	0,0003
beta	0.01
epsilon	0,2
lambd	0,97
num_epoch	3
learning_rate_schedule	linear
beta_schedule	linear
normalize	false
hidden_units	256
num_layers	3
vis_encode_type	nature_cnn
max_steps	6000000
time_horizon	128

Обозначим ключевые различия между программными реализациями рассматриваемых архитектур. Во-первых, при использовании подхода CTCE модифицируется структура входных и выходных данных нейросетевой модели: визуальные наблюдения всех четырех агентов агрегируются, а выходной слой формирует совместное управляющее воздействие, включающее действия всех агентов, что приводит к существенному росту размерности модели и вычислительных затрат. Во-вторых, в архитектуре CTDE применяется специализированный алгоритм, учитывающий межагентные зависимости за счет контрфактического оценивания и механизма внимания, а также использующий общий сигнал вознаграждения для группы агентов.

Для каждого архитектурного варианта было выполнено несколько независимых обучающих прогонов. Модели, демонстрировавшие средние значения суммарного вознаграждения, близкие к медианным по серии экспериментов, использовались для последующего расчета и анализа метрик эффективности. Также была произведена отдельная серия обучающих прогонов для карт, размером 300×300 пикселей с длиной эпизода в 6000 шагов. Значения собранных метрик эффективности для карт малых размеров приведены в таблице 3, для карт больших размеров – в таблице 4.

Обучающие прогоны осуществлялись на вычислительной машине со следующими аппаратными характеристиками: процессор Intel Core i7-12650H, видеокарта NVIDIA GeForce RTX 3060 Laptop 6GB VRAM (драйвер CUDA 11.8), оперативная память 16GB DDR4 3200MHz. Нейросетевые вычисления осуществлялись с использованием графического процессора. Основными проблемами для обученных агентов стали сложности при покрытии ячеек вблизи границ карты и точное распознавание небольших участков целевой области, что привело к увеличению среднего времени завершения эпизода.

Таблица 3. Метрики эффективности для малых карт

Метрика	IA	CTCE	CTDE
Средняя длительность схождения, эпизод	4600	5500	8300
Затраченное на обучение время, час	4,232	25,3	8,254
Доля покрытия целевой области, %	98,76	92,1	98,6
Среднее время завершения, шаг	586,44	594,2	575,8
Средняя скорость покрытия, ячейка за шаг	4,03	3,71	4,1
Средняя скорость покрытия в первой половине, ячейка за шаг	9,14	7,79	8,3
Среднее межагентное расстояние, ячейка	46,3	47,41	50,2

Таблица 4. Метрики эффективности для больших карт

Метрика	IA	CTCE	CTDE
Средняя длительность схождения, эпизод	420	540	820
Затраченное на обучение время, час	4,566	26,1	8,83
Доля покрытия целевой области, %	97,49	88,3	98,91
Среднее время завершения, шаг	5353,45	5361,9	5040,98
Средняя скорость покрытия, ячейка за шаг	3,92	3,55	4,22
Средняя скорость покрытия в первой половине, ячейка за шаг	7,61	7,73	9,66
Среднее межагентное расстояние, ячейка	140,9	152,81	166,1

Полученные результаты подтверждают, что указанные эффекты носят системный характер и проявляются независимо от выбранной архитектуры обучения, однако их влияние в наибольшей степени сглаживается при использовании подхода с централизованным обучением и децентрализованным исполнением. В частности, применение контрфактической оценки и общего сигнала вознаграждения способствует более равномерному распределению агентов в пространстве и снижению избыточных перекрытий локальных зон охвата агентов. Дополнительная серия экспериментов на картах увеличенного размера показала, что рост размерности среды позволяет более отчетливо выявить различия в динамике координации агентов и снижает влияние граничных эффектов на интегральные метрики эффективности, что подтверждает целесообразность использования карт большого масштаба при сравнительном анализе мультиагентных стратегий.

Для снижения влияния граничных эффектов на эффективность агентов целесообразно модифицировать среду путем создания буферной зоны по периметру карты и перехода к использованию недопустимого действия вместо досрочного завершения эпизода. Буферная зона шириной 3 пикселя доступна для посещения и не содержит целевых ячеек. Это позволяет минимизировать искажение стратегии агентов, вызванное граничными эффектами. Прерывание эпизода при попытке выхода за границу заменяется на недопустимое действие: при выборе такого действия агент получает штраф и остается в прежней ячейке. Это позволяет агенту на следующих шагах вернуться в рабочую область и скорректировать траекторию. Предварительный визуальный анализ в последующих тестах показал, что предложенные модификации снижают вероятность неполного покрытия приграничных ячеек и уменьшают влияние единичных ошибок управления у границ. Однако количественная оценка эффекта модификаций требует отдельного экспериментального сравнения и рассматривается как направление дальнейшей работы.

Выводы

В работе рассмотрена задача планирования пути покрытия целевой области группой однородных автономных агентов в условиях частичной наблюдаемости и стохастической динамики среды. Задача формализована в виде частично наблюдаемого марковского процесса принятия решений, что позволило использовать методы глубокого обучения с подкреплением для синтеза децентрализованных стратегий управления.

Разработан программный комплекс для имитационного моделирования, обучения агентов и анализа показателей эффективности, обеспечивающий воспроизводимость экспериментов и сопоставимость результатов для различных мультиагентных архитектур. Реализованная среда позволяет варьировать размер карт, число агентов и параметры вознаграждения, а также проводить детальный анализ динамики покрытия и координации агентов.

В ходе экспериментального исследования выполнено сравнение подходов обучения независимых агентов, централизованного обучения с централизованным выполнением и централизованного обучения с децентрализованным выполнением. Показано, что архитектура с централизованным критиком и децентрализованным выполнением обеспечивает более высокие показатели скорости и полноты покрытия, а также более устойчивую координацию агентов по сравнению с независимыми стратегиями при сохранении децентрализованного характера управления.

Установлено, что использование карт малого размера ограничивает информативность ряда метрик эффективности и может искажать оценку координации агентов. Эксперименты на картах увеличенного масштаба позволяют более отчетливо выявить различия между архитектурами и снизить влияние граничных эффектов и локальных особенностей целевой области на итоговые результаты.

Полученные результаты подтверждают перспективность применения методов мультиагентного обучения с подкреплением для решения задач покрытия в сложных дискретных средах и могут быть использованы при разработке интеллектуальных систем управления группами автономных мобильных агентов. Направлениями дальнейших исследований являются усложнение модели среды, учет динамических препятствий и развитие методов адаптивной координации для масштабируемых мультиагентных систем.

Литература

1. Луканов С. Ю. Программный комплекс для имитационного моделирования и обучения агентов в задачах покрытия области в виртуальных средах на основе методов глубокого обучения с подкреплением. Свидетельство о государственной регистрации программы для ЭВМ RU 2025686965, опубл. 07.10.2025.

2. Луканов С. Ю. Программное средство «Менеджер конфигураций RLCPP» для работы с программным комплексом глубокого обучения с подкреплением в задачах покрытия области. Свидетельство о государственной регистрации программы для ЭВМ RU 2025693955, опубл. 03.12.2025.

3. Луканов С. Ю. Программа для расчета и визуализации показателей эффективности в задачах покрытия области. Свидетельство о государственной регистрации программы для ЭВМ RU 2025685840, опубл. 26.09.2025.

4. Nguyen T. T., Nguyen N. D., Nahavandi S. Deep Reinforcement Learning for Multiagent Systems: A Review of Challenges, Solutions, and Applications // IEEE Transactions on Cybernetics. 2020. Vol. 50. Iss. 9. PP. 3826–3839. DOI: 10.1109/TCYB.2020.2977374. EDN: OKVIPT

5. 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. PP. 2845–2847. DOI: 10.1007/s10458-023-09633-6

6. Schulman J., Wolski F., Dhariwal P., Radford A., Klimov O. Proximal Policy Optimization Algorithms // ArXiv. 20.07.2017. DOI: 10.48550/arXiv.1707.06347

7. Andrychowicz M., Raichuk A., Stańczyk P., Orsini M., Girgin S. et al. What Matters in On-Policy Reinforcement Learning? A Large-Scale Empirical Study // ArXiv. 10.06.2020. DOI: 10.48550/arXiv.2006.05990

8. Cohen A., Teng E., Berges V.-P., Dong R.-P., Henry H. et al. On the Use and Misuse of Absorbing States in Multi-Agent Reinforcement Learning // ArXiv. 10.11.2022. DOI: 10.48550/arXiv.2111.05992

Материалы статьи были представлены на VI Всероссийской научно-технической и научно-методической конференции магистрантов, аспирантов и их руководителей «Перспективные телекоммуникационные технологии и развитие цифровых кластеров в России и мире (ПКМ 2025)».

Статья поступила 11 декабря 2025 г.
Одобрена после рецензирования 26 января 2026 г.
Принята к публикации 03 февраля 2026 г.

Информация об авторах

Луканов Сергей Юрьевич – старший преподаватель отделения информационно-коммуникационных технологий образовательного департамента (Передовой инженерной школы гибридных технологий в станкостроении Союзного государства) Псковского государственного университета. Email: lukanovysergey@gmail.com

Тимошевская Ольга Юрьевна – кандидат технических наук, доцент отделения информационно-коммуникационных технологий образовательного департамента (Передовой инженерной школы гибридных технологий в станкостроении Союзного государства) Псковского государственного университета. Email: olga.tim777@yandex.ru

Experimental Study of Multi-Agent Reinforcement Learning Approaches for the Coverage Path Planning Problem

S. Lukanov✉, O. Timoshevskaya

Pskov State University,
Pskov, 180000, Russian Federation

Purpose. *The coverage path planning problem for groups of autonomous agents over a given target area is relevant to a wide range of applied systems. An increase in the dimensionality of the environment and in the number of interacting agents leads to higher coordination complexity and longer times to achieve complete coverage. Additional difficulty arises from agents' limited observations, which results in a control problem formulated under partial observability and stochastic dynamics. In this context, the development of specialized intelligent control approaches that minimize the time required to cover the target area under decentralized decision-making is of significant interest. The aim of this work is to develop a framework for training control models for groups of homogeneous autonomous agents in the coverage path planning problem, ensuring minimization of the expected time to achieve full coverage under a partially observable Markov decision process.* ***Methods.*** *To formalize the problem, a partially observable Markov decision process model is employed, including a description of system states, agents' action and observation spaces, probabilistic environment dynamics, and a reward function. The synthesis of a group control policy is based on deep reinforcement learning methods for multi-agent systems, oriented toward decentralized execution with centralized training using a critic. Performance evaluation is carried out using simulation-based experiments in a discrete two-dimensional grid environment.* ***Novelty.*** *This work introduces a unified experimental environment for comparing multi-agent architectures in the coverage path planning problem. It is shown that the use of small-scale maps limits the statistical significance of several agent coordination metrics, which motivates the transition to larger maps. Typical sources of coverage performance degradation related to boundary effects and small target area fragments are identified, and an environment modification is proposed to mitigate their impact.* ***Results.*** *A formal problem statement is presented, and an approach to synthesizing a group control policy is proposed that enables complete coverage of the target area within finite time. Simulation results confirm the feasibility of effective agent coordination and a reduction in coverage time compared to uncoordinated strategies, while preserving decentralized control. Experimental studies reveal differences in coverage dynamics and agent coordination across the considered architectures.* ***The theoretical significance*** *of this work lies in the advancement of methods for formalizing and solving multi-agent coverage problems under partial observability.* ***The practical significance*** *is determined by the applicability of the obtained results to the development of intelligent control systems for groups of autonomous mobile agents, including applications in monitoring, reconnaissance, and robotic systems. The results can be used in the design of multi-agent coverage systems and in the comparative analysis of agent control methods in complex discrete environments.*

Key words: *deep reinforcement learning, control system, coverage path planning, multi-agent system, artificial intelligence*

Information about Author

Lukanov Sergey – Senior Lecturer of the Information and Communication Technologies Division, Educational Department of the Advanced Engineering School of Hybrid Technologies in Machine Tool Engineering of the Union State (Pskov State University). E-mail: lukanovysergey@gmail.com

Timoshevskaya Olga – Ph. D. of Engineering Sciences, Associate Professor of the Information and Communication Technologies Division, Educational Department of the Advanced Engineering School of Hybrid Technologies in Machine Tool Engineering of the Union State (Pskov State University).
Email: olga.tim777@yandex.ru