УДК 004.85

Обзорная статья

https://doi.org/10.32603/2071-8985-2025-18-3-65-77

Оптимизация стратегии в алгоритмах обучения с подкреплением в логистических системах принятия решений

А. Р. Салиева✉, Н. А. Верзун, М. О. Колбанев

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В. И. Ульянова (Ленина), Санкт-Петербург, Россия

✉ rustamovna.a3@gmail.com

Аннотация. В обзорной статье ставится задача анализа и систематизации современных исследований в области оптимизации стратегий алгоритмов обучения с подкреплением, применяемых в логистических системах принятия решений. В ходе работы над обзором были рассмотрены научные публикации за последние 5 лет, индексируемые в ведущих базах данных, посвященные применению методов обучения с подкреплением в логистике. Особое внимание уделено работам, описывающим алгоритмы Policy Gradient и Proximal Policy Optimization (PPO). Методология обзора включает сравнительный анализ, классификацию подходов и оценку их эффективности. Выявлены основные тенденции в развитии методов оптимизации стратегий для логистических систем. Определены ключевые преимущества и ограничения различных подходов. Установлено, что методы на основе PPO демонстрируют наибольшую эффективность в сложных динамических средах. Обнаружен растущий интерес к гибридным подходам, сочетающим обучение с подкреплением и классические методы оптимизации. Выделены перспективные направления дальнейших исследований, включая адаптацию алгоритмов к специфическим задачам логистики и повышение их интерпретируемости. Полученные результаты могут служить основой для разработки новых алгоритмов и их практического применения в различных секторах логистики и управления цепями поставок.

Ключевые слова: логистические системы принятия решений, обучение с подкреплением, оптимизация стратегии, Policy Gradient методы, Proximal Policy Optimization

Для цитирования: Салиева А. Р., Верзун Н. А., Колбанев М. О. Оптимизация стратегии в алгоритмах обучения с подкреплением в логистических системах принятия решений // Изв. СПбГЭТУ «ЛЭТИ». 2025. Т. 18, № 3. С. 65–77. doi: 10.32603/2071-8985-2025-18-3-65-77.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Review article

Strategy Optimization in Reinforcement Learning Algorithms in Logistic Decision-Making Systems

A. R. Salieva✉, N. A. Verzun, M. O. Kolbanev

Saint Petersburg Electrotechnical University, Saint Petersburg, Russia

✉ rustamovna.a3@gmail.com

Abstract. This review paper aims to analyze and systematize the current research in the field of strategy optimization of reinforcement learning algorithms used in logistic decision-making systems. In the course of the review we have considered scientific publications for the last 5 years, indexed in the leading databases, devoted to the application of reinforcement learning methods in logistics. Particular attention is paid to papers describing Policy Gradient and Proximal Policy Optimization (PPO) algorithms. The methodology of the review includes

comparative analysis, classification of approaches and evaluation of their effectiveness. The main trends in the development of policy optimization methods for logistics systems are identified. The key advantages and limitations of different approaches are identified. It is found that PPO-based methods demonstrate the highest efficiency in complex dynamic environments. A growing interest in hybrid approaches combining reinforcement learning and classical optimization methods is found. Promising directions for further research are highlighted, including adapting algorithms to specific logistics problems and improving their interpretability. The results obtained can serve as a basis for the development of new algorithms and their practical application in various sectors of logistics and supply chain management.

Keywords: logistic decision-making systems, reinforcement learning, strategy optimization, Policy Gradient methods, Proximal Policy Optimization

For citation: Verzun N. A., Kolbanev M. O. Salieva A. R. Strategy Optimization in Reinforcement Learning Algorithms in Logistic Decision-Making Systems // LETI Transactions on Electrical Engineering & Computer Science. 2025. Vol. 18, no. 3. P. 65–77. doi: 10.32603/2071-8985-2025-18-3-65-77.

Conflict of interest. The authors declare no conflicts of interest.

Введение. Логистические системы принятия решений – это комплексные системы, предназначенные для управления и оптимизации логистических процессов и используемые для принятия решений на основе анализа большого объема данных с целью улучшения эффективности и производительности логистических операций – транспортировки, складирования, управления запасами и распределения [1] и т. п.

С развитием технологий и ростом объема данных логистические процессы становятся все более сложными и требуют внедрения передовых решений для оптимизации. В этом контексте методы машинного обучения, в частности методы обучения с подкреплением (Reinforcement Learning, RL), играют все более значимую роль.

Логистические системы принятия решений находят свое применение на практике при решении разнообразных логистических задач:

• *Оптимизация маршрутов доставки*. Планирование и оптимизация маршрутов транспортных средств направлены на снижение затрат на транспортировку и улучшение времени доставки [2]. Примеры подобных систем включают программное обеспечение для маршрутизации грузовиков, системы навигации и логистические платформы, которые анализируют дорожные условия в реальном времени.

• *Управление складскими запасами*. Управление запасами на складе предполагает: контроль уровня запасов, прогнозирование потребности и оптимизация размещения товаров и пр. [3]. Примеры: автоматизированные системы хранения и извлечения (AS/RS), программы для управления запасами и системы прогнозирования спроса.

• *Планирование производства и поставок*. Координация производственных процессов и управление цепочками поставок с целью обеспечения своевременного снабжения материалами и компонентами, необходимыми для производства [4]. Примеры систем данного типа: системы планирования ресурсов предприятия (ERP), системы планирования потребностей в материалах (MRP) и программное обеспечение для управления цепочкой поставок (SCM).

• *Прогнозирование спроса*. На основе обработки и анализа накопленных исторических данных формируются тренды для предсказания будущего спроса на продукцию, что позволяет лучше планировать производство и управление запасами [5]. Примеры систем такого типа включают аналитическое программное обеспечение, системы машинного обучения и платформы больших данных.

Основные компоненты системы поддержки принятия решений в логистике показаны на обобщенной схеме (рис. 1).

Интеллектуальная логистическая система поддержки принятия решений – это система, использующая методы искусственного интеллекта и машинного обучения для анализа больших объемов данных и выработки оптимальных решений в логистике, она позволяет автоматизировать и улучшать процесс принятия решений, снижать затраты и повышать эффективность логистических операций.

Для корректной выработки решений системе принятия решений требуются *большие объемы разнообразных цифровых данных*, описывающих логистическую систему и ее окружение. Источниками цифровых данных в логистике могут выступать:

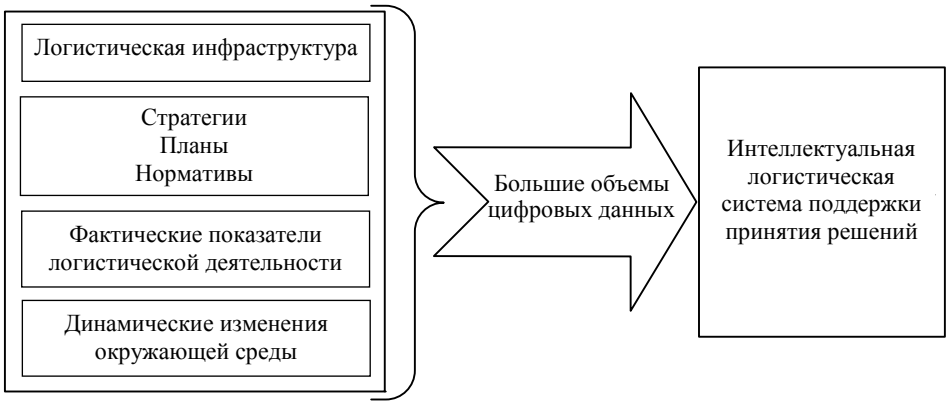

Рис. 1. Обобщенная схема системы принятия решений в логистике
Fig. 1. Generalized scheme of the decision-making system in logistics

• *Логистическая инфраструктура*, которая представляет собой физические и организационные структуры и средства, необходимые для функционирования логистических процессов. Это транспортные сети, склады, оборудование и информационные системы, которые обеспечивают эффективное перемещение и хранение товаров и сопровождающих их информационных потоков [6].

• *Планы* – предварительно разработанные действия и последовательности шагов для достижения целей в логистике. Например, предполагаемые маршруты транспортировки, графики доставки и пр.

• *Стратегии* – долгосрочные планы и подходы к управлению логистикой, направленные на повышение эффективности и снижение затрат. Могут включать инновации в транспортировке, оптимизацию складских процессов и интеграцию новых технологий.

• *Нормативы* – это стандарты и правила, регулирующие логистические операции: законодательные требования, отраслевые стандарты и внутренние регламенты компании.

• *Результаты мониторинга и фактические показатели логистической деятельности* – данные, полученные в результате постоянного наблюдения за логистическими процессами. Включают в себя информацию о фактических временах доставки, состоянии запасов, использовании транспортных средств и других ключевых показателях эффективности логистической деятельности.

• *Динамические изменения окружающей среды*. На любую логистическую систему влияет множество внешних факторов, не зависящих от ее деятельности. Внешние факторы имеют различную природу возникновения, могут изменяться, к ним необходимо приспосабливаться, учитывать при принятии решений. Это, например, изменения в экономических условиях, политической обстановке, погодные условия, природные катаклизмы, дорожно-транспортная обстановка и другие, зачастую непредсказуемые события и явления, которые могут повлиять на логистические процессы.

Системы поддержки принятия решений меняются, приспосабливаясь к новым вызовам и требованиям цифровой трансформации логистической отрасли, которая предполагает: внедрение новых бизнес-моделей, цифровых технологий, цифровых платформ и пр. Так, например, можно отметить следующие особенности принятия решений в операционных логистических задачах:

• *Большие объемы цифровых данных*. Логистические системы ежедневно обрабатывают огромные объемы цифровых данных, которые поступают из различных внутренних (генерируемых в ходе деятельности организаций) и внешних (поставщики, клиенты и другие участники цепочки поставок) источников. Сегодня на смену традиционным информационным системам, автоматизирующим различные направления логистики приходят цифровые платформы, позволяющие слить в единое информационное пространство данные, например: ERP-систем (Enterprise Resource Planning – система планирования ресурсов предприятия), CRM-систем (Customer Relationship Management – система управления взаимоотношениями с клиентами), WMS-систем (Warehouse Management System – система управления складом), TMS-систем (Transportation Management System – система управления транспортом), и пр. Кроме того, большие объемы данных генерируются автоматически: навигационные данные, поступающие от аппаратуры спутниковой навигации, данные, поступающие от датчиков, сенсоров на автоматизированных складах и пр.

• *Работа в реальном режиме времени* – в условиях высокой динамичности рынка логисти-

ческие решения должны приниматься и реализовываться мгновенно. Это требует от систем способности быстро обрабатывать поступающую информацию и генерировать оптимальные решения «на лету», учитывая текущие изменения в дорожной обстановке, погодных условиях и другие влияющие факторы.

• *Возможная неполнота и противоречивость данных* – в логистике данные часто поступают автоматически из различных сенсоров и датчиков, таких как GPS-трекеры, устройства мониторинга состояния транспортных средств и складского оборудования. Эти данные могут быть неточными, запаздывающими или противоречивыми, что создает дополнительные сложности для принятия решений.

В связи с этими особенностями компонент схемы рис. 1, непосредственно принимающий решение, разрабатывается на основе технологий искусственного интеллекта. Один из наиболее подходящих методов – это машинное обучение с подкреплением.

Обучение с подкреплением в логистических системах принятия решений. В основе *обучения с подкреплением* (*Reinforcement Learning*, *RL*) лежит идея обучения на основе взаимодействия *агента* (*agent*) с *окружающей средой* (*environment*) для достижения определенных целей [7]. Агент принимает решения на основе текущего *состояния* (*states*) среды, выполняет действия (*actions*) и получает обратную связь в виде *вознаграждений* (*reward*) или *наказаний* в зависимости от результата его действий. Цель агента – максимизировать суммарное вознаграждение за определенный период времени, корректируя свои действия.

RL идеально подходит для логистических задач, так как позволяет системам адаптироваться к динамическим изменениям и неопределенности, эффективно обрабатывать большие объемы данных и работать в реальном времени.

Примеры применения RL в логистике включают оптимизацию маршрутов доставки, управление складскими запасами, планирование производства и распределения ресурсов [8]. Преимущества использования методов RL:

– *адаптивность*. RL-агенты могут адаптироваться к изменяющимся условиям и новым ситуациям;

– *автономность*. Способность обучаться без необходимости точного моделирования среды;

– *оптимизация долгосрочных стратегий*. Максимизация совокупного вознаграждения за длительный период, а не за краткосрочные цели;

– *решение сложных задач*. Способность справляться с задачами, где классические методы оптимизации менее эффективны.

Обучение с подкреплением в логистических системах принятия решений позволяет:

1. Принимать решения в условиях *сложных и динамических сред*. Логистические задачи часто характеризуются изменяющимися условиями, большим количеством переменных и необходимостью оперативного принятия решений.

2. Подходит для задач *долгосрочного планирования*. В логистике важно не только мгновенное оптимальное решение, но и стратегическое планирование для минимизации издержек и максимизации эффективности в долгосрочной перспективе.

3. Позволяет учитывать *случайные события*. Логистические процессы подвержены неопределенностям (например, задержки в поставках, изменение спроса). RL дает возможность учитывать такие случайности и адаптироваться к ним.

4. Повышает *эффективность использования ресурсов*. RL помогает оптимизировать распределение ресурсов, таких как транспортные средства, склады, персонал, что ведет к повышению общей эффективности логистической системы.

5. *Интеграция с другими технологиями*. RL легко интегрируется с другими методами машинного обучения и анализа данных, что позволяет создавать комплексные и мощные системы поддержки принятия решений.

Методов обучения с подкреплением очень много и классифицируют их по различным критериям. С точки зрения подхода к обучению методы *Reinforcement Learning* делятся на [9]:

– оптимизацию политики (стратегии);

– оптимизацию функции ценности;

– смешанные.

Наиболее эффективны для логистических систем принятия решений методы с оптимизацией стратегии, что обусловлено их ключевыми преимуществами:

– адаптивностью к динамическим условиям;

– оптимизацией долгосрочных результатов;

– балансом между исследованием и использованием;

– способностью справляться с неопределенностью;

– масштабируемостью для сложных систем.

Примеры применения RL на основе оптимизации стратегии в логистических системах принятия решений:

1. *Системы оптимизации маршрутов*. RL-алгоритмы, основанные на оптимизации стратегии, могут адаптироваться к текущим дорожным условиям и выбирать наиболее эффективные маршруты в реальном времени.

2. *Управление запасами.* Оптимизация стратегии позволяет системе принимать решения о пополнении запасов, учитывая долгосрочные потребности и текущие условия, что помогает избежать как недостатка, так и избытка товаров на складе.

3. *Планирование производства и цепочек поставок*. С помощью методов RL можно оптимизировать планирование производственных процессов и управление цепочками поставок, учитывая множество факторов и обеспечивая бесперебойную работу всей системы.

Методы обучения с подкреплением на основе оптимизации стратегии. При обучении с подкреплением на основе оптимизации стратегии агент учится принимать решения, взаимодействуя с окружающей средой, чтобы максимизировать некоторую кумулятивную награду [7]. Основные элементы такого подхода показаны на рис. 2.

Компоненты подхода: агент (agent) – объект, который принимает решения; окружающая среда (environment) – все, с чем взаимодействует агент; действия (actions) – множество возможных действий, которые агент может выполнять; состояния (states) – различные ситуации или состояния, в которых может находиться агент; награда (reward) – оценка, которую получает агент за выполнение определенных действий в конкретных состояниях.

Процесс обучения с подкреплением предполагает следующие шаги взаимодействия:

1. *Состояние* – агент получает текущее состояние от окружающей среды.

2. *Действие* – агент выбирает действие на основе текущей политики.

3. *Награда* – окружающая среда реагирует на действие и предоставляет новое состояние и награду.

4. *Оценка* (*critic's evaluation*) – критик оценивает ценность текущей политики. Эта оценка идет от критика к политике.

5. *Обновление политики* (*policy update*) – актор обновляет параметры политики на основе оценки критика.

Основные компоненты:

– политика (*policy*, π) – стратегия, определяющая поведение агента. Политика может быть детерминированной или стохастической и описывает вероятность выбора определенного действия в данном состоянии;

– функция ценности (*value function*, *V*) – оценка ожидаемой кумулятивной награды, которую агент может получить, находясь в определенном состоянии и следуя определенной политике;

– функция ценности действия (*action-value function*, *Q*) – оценка ожидаемой кумулятивной награды, которую агент может получить, выполняя конкретное действие в определенном состоянии и далее следуя определенной политике.

Оптимизация стратегии в RL включает использование следующих основных двух классов методов:

Политика градиентного подъема – Policy Gradient Methods. Обновляет параметры политики напрямую на основе градиента ожидаемой награды.

Обучении стратегии с учетом границ (*ограничений*) *для обновлений – Proximal Policy Optimization* (*PPO*). Метод, который ограничивает изменения политики, чтобы стабилизировать обучение и улучшить производительность.

Рассмотрим подробнее методы этих классов.

Policy Gradient Methods. *Policy Gradient* – это класс алгоритмов RL, которые оптимизируют политику напрямую, используя градиентные методы

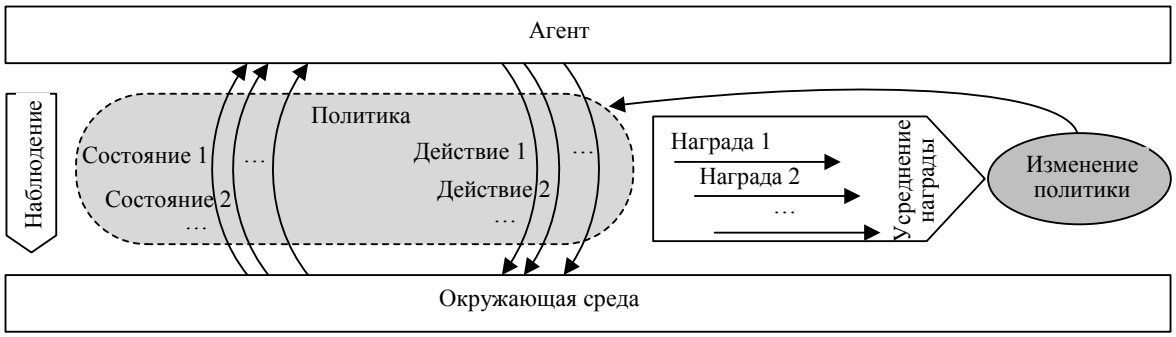

Рис. 2. Обучение с подкреплением на основе оптимизации стратегии
Fig. 2. Reinforcement learning based on strategy optimization

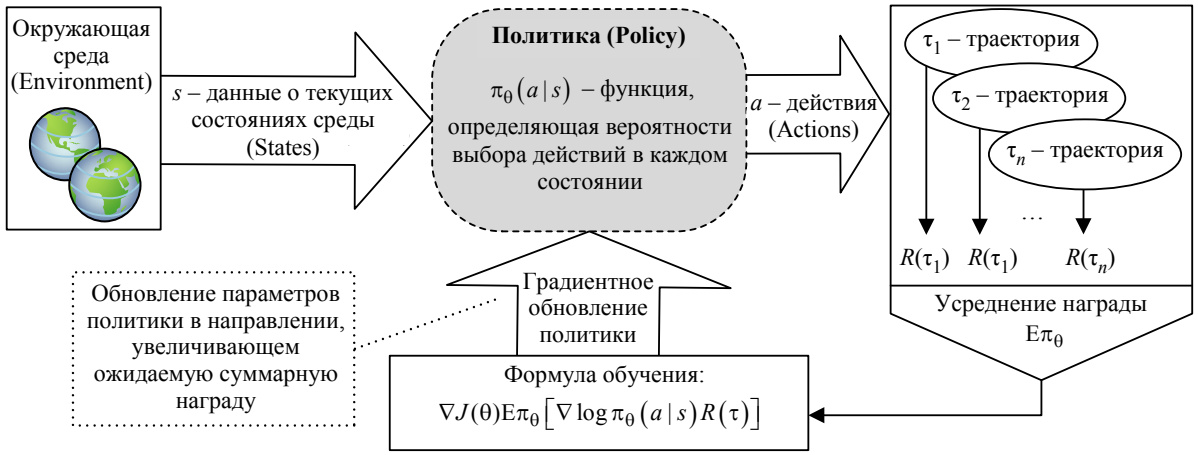

Рис. 3. Схема работы алгоритма с оптимизацией политики типа Policy Gradient
Fig. 3. The scheme of the algorithm with optimization of the policy type Policy Gradient

для максимизации ожидаемого вознаграждения [10]. В отличие от методов Q-learning и SARSA, методы Policy Gradient напрямую задают политику и обучают ее с помощью градиентного спуска. Примером метода типа Policy Gradient может служить REINFORCE, предусматривающий обновление параметров политики на основе накопленного вознаграждения [11].

На рис. 3 представлен принцип работы обучения с подкреплением с оптимизацией политики типа Policy Gradient.

На вход подаются текущие состояния среды. Основные компоненты включают:

– *политику* – функцию, которая определяет вероятности выбора действий в каждом состоянии;

– *функцию ценности* – оценивает ожидаемую награду для каждого состояния или пары «состояние–действие»;

– *градиент политики* – вычисляется для обновления параметров политики.

Процесс работы предполагает выполнение следующих шагов:

– агент взаимодействует со средой, выбирая действия согласно текущей политике;

– собираются траектории (последовательности состояний, действий и наград);

– вычисляется градиент политики на основе полученных наград;

– параметры политики обновляются в направлении, увеличивающем ожидаемую суммарную награду.

Цель алгоритма – найти оптимальную политику, максимизирующую ожидаемую награду в долгосрочной перспективе.

Эта схема сочетает идеи генеративных моделей и обучения с подкреплением. Генератор создает данные или действия, которые затем оцениваются и используются для обновления стратегии с целью улучшения будущих результатов.

Основная формула обучения алгоритма:

$$\nabla J(\theta) E\pi_\theta \left[\nabla \log \pi_\theta(a|s) R(\tau)\right],$$

где $J(\theta)$ – ожидаемое вознаграждение при параметрах политики θ; $\pi_\theta(a|s)$ – параметризованная политика, которая определяет вероятность выбора действия a в состоянии s с параметрами θ; $R(\tau)$ – суммарное вознаграждение, полученное за траекторию τ, где траектория τ состоит из последовательности состояний, действий и вознаграждений; $E\pi_\theta$ – математическое ожидание по всем траекториям τ, сгенерированным согласно политике π_θ; $\nabla \log \pi_\theta(a|s)$ – градиент логарифма вероятности действия a, выбранного политикой в состоянии s.

Архитектура алгоритма обучения:

– *политика* (*нейронная сеть*) аппроксимирует распределение вероятностей для действий в каждом состоянии;

– *градиентное обновление* параметров политики на основе градиента ожидаемого вознаграждения.

Методы Policy Gradient, например REINFORCE, обеспечивают гибкость и мощность при решении сложных задач обучения с подкреплением, особенно в тех случаях, когда пространство состояний и действий велико или непрерывно.

Примеры применения Policy Gradient RL в логистических системах принятия решений.

Методы Policy Gradient (REINFORCE) подходят для задач с дискретными действиями или простыми непрерывными задачами, где вариативность градиента и стабильность обновлений менее критичны. Примеры:

1. *Динамическое планирование производства*. Policy Gradient RL [12] применяется для динамического планирования в гибком производственном

процессе. Алгоритм учится оптимально распределять задачи по машинам, учитывая текущую загрузку, приоритеты заказов и сроки выполнения. Это позволяет повысить эффективность производства и сократить время выполнения заказов.

Формулировка проблемы (Flexible Job Shop Problem, FJSP) включает в себя набор заданий, каждое из которых состоит из последовательности операций. Каждая операция может быть выполнена на одной из нескольких машин с разным временем обработки.

Представление состояния. Состояние системы может включать текущее распределение заданий по машинам, оставшееся время выполнения текущих операций, очередь ожидающих заданий и т. д.

Действия. Действия могут включать назначение следующей операции на конкретную машину или изменение приоритетов заданий.

Функция вознаграждения. Вознаграждение может быть основано на таких факторах, как общее время выполнения всех заданий, загруженность машин, соблюдение сроков и т. д.

Архитектура нейронной сети. Авторы используют многослойный перцептрон или рекуррентную нейронную сеть для представления политики.

Обучение. Процесс обучения включает множество эпизодов, где агент пытается оптимизировать расписание, получая обратную связь в виде вознаграждений.

Улучшение алгоритма. Статья описывает специфические улучшения стандартного алгоритма Policy Gradient для лучшей работы с FJSP, например, использование базовых эвристик для инициализации политики или специальные методы исследования пространства состояний.

Эксперименты и результаты. Авторы сравнивали свой метод с традиционными алгоритмами планирования и другими подходами машинного обучения, демонстрируя улучшения в таких метриках, как общее время выполнения заданий или адаптивность к изменениям в производственной среде.

2. *Решение задачи маршрутизации транспорта.* Исследование [13] демонстрирует применение Policy Gradient RL для решения задачи маршрутизации транспортных средств (Vehicle Routing Problems, VPR). Алгоритм учится эффективно строить маршруты доставки, минимизируя общее расстояние и время в пути, что особенно полезно для компаний, занимающихся доставкой товаров.

Policy Gradient применяется к VRP следующим образом: агент (система принятия решений) наблюдает текущее состояние: расположение транспортных средств и клиентов. На каждом шаге агент выбирает следующего клиента для посещения. После выбора клиента агент получает вознаграждение, основанное на эффективности маршрута. Алгоритм обновляет политику выбора клиентов, чтобы максимизировать общую эффективность маршрутов.

Входными данными выступают: координаты депо и клиентов, требования клиентов (объем, вес груза и пр.), характеристики транспортных средств (вместимость), матрица расстояний или функция расчета расстояний до пунктов доставки.

Обработка данных осуществляется следующим образом: алгоритм использует нейронную сеть (обычно на основе архитектуры внимания) для представления политики выбора клиентов. Сеть обрабатывает текущее состояние системы и выдает вероятности выбора каждого клиента. Клиент выбирается на основе этих вероятностей (возможно, с использованием «жадной» стратегии или выборки). После каждого решения алгоритм вычисляет градиент политики и обновляет параметры сети для улучшения будущих решений.

На выходе алгоритм предоставляет: оптимизированную политику выбора клиентов, последовательность посещения клиентов для каждого транспортного средства, оценку общей длины маршрута или стоимости решения

Этот алгоритм позволяет находить близкие к оптимальным решения для задачи VRP, адаптируясь к различным конфигурациям клиентов и требованиям. Он особенно эффективен для задач VRP большой размерности и может быть легко адаптирован к их различным вариантам.

3. *Оптимизация управления запасами*. В [14] рассмотрено применение RL к управлению запасами в цепочке поставок крови. Policy Gradient алгоритм обучается принимать решения о заказе и распределении крови, учитывая ее ограниченный срок годности, неопределенность спроса и важность избегания дефицита. Этот подход может быть адаптирован для управления запасами, например для скоропортящихся товаров.

Policy Gradient в контексте управления запасами крови работает следующим образом: агент (система принятия решений) наблюдает текущее состояние запасов крови и спроса. На основе этой информации агент принимает решение о количестве крови для заказа. После принятия решения система переходит в новое состояние, и агент получает вознаграждение, основанное на эффектив-

ности решения. Алгоритм обновляет политику принятия решений, чтобы максимизировать долгосрочное вознаграждение.

Входными данными служит: текущий уровень запасов и прогноз спроса на кровь, информация о сроках годности крови, стоимость хранения и утилизации, штрафы за нехватку крови.

Обработка данных происходит следующим образом: алгоритм использует нейронную сеть для представления политики принятия решений. Сеть принимает текущее состояние системы и выдает рекомендацию по объему заказа крови. После каждого решения алгоритм вычисляет градиент политики и обновляет параметры сети для улучшения будущих решений.

На выходе алгоритм предоставляет: оптимизированную политику управления запасами крови, рекомендации по объему заказа крови для каждого периода, оценку ожидаемых затрат и уровня обслуживания.

Proximal Policy Optimization (PPO) – это результат усовершенствования методов Policy Gradient, разработанный для улучшения стабильности и эффективности обучения. PPO вводит ограничение на размер изменения политики между обновлениями, предотвращая слишком большие обновления, которые могут привести к ухудшению производительности [15].

На рис. 4 представлен алгоритм Proximal Policy Optimization (PPO). На вход алгоритму подаются данные о текущем состоянии среды и предыдущих взаимодействиях агента со средой.

Основные компоненты алгоритма PPO включают:

– *политику* – нейронную сеть, которая определяет вероятности выбора действий в каждом состоянии;

– *функцию ценности*, оценивающую ожидаемую награду для каждого состояния;

– *механизм ограничения обновлений политики*, предотвращающий слишком большие изменения политики за один шаг обучения.

Процесс работы предполагает выполнение следующих шагов:

1. Агенты получают текущее состояние из среды.
2. Состояние подается на вход нейронной сети (политики).
3. Политика выбирает действия для агентов.
4. Действия выполняются в среде SUMO.
5. Среда возвращает новое состояние и награды.
6. Политика обновляется с использованием алгоритма PPO на основе полученного опыта.

Основная формула обучения алгоритма [16]:

$$L^{\mathrm{CLIP}}(\theta) = \\ = \mathrm{E}_t\left[\min\left(r_t(\theta)\widehat{A}_t, \mathrm{clip}\left(r_t(\theta), 1-\varepsilon, 1+\varepsilon\right)\widehat{A}_t\right)\right],$$

где $r_t(\theta)$ – вероятностное соотношение между новой и старой политикой; $\widehat{A}_t$ – оценка преимущества (advantage estimate) для времени t; ε – гиперпараметр, определяющий допустимое изменение политики; $\mathrm{clip}\left(r_t(\theta), 1-\varepsilon, 1+\varepsilon\right)$ – ограничение веро-

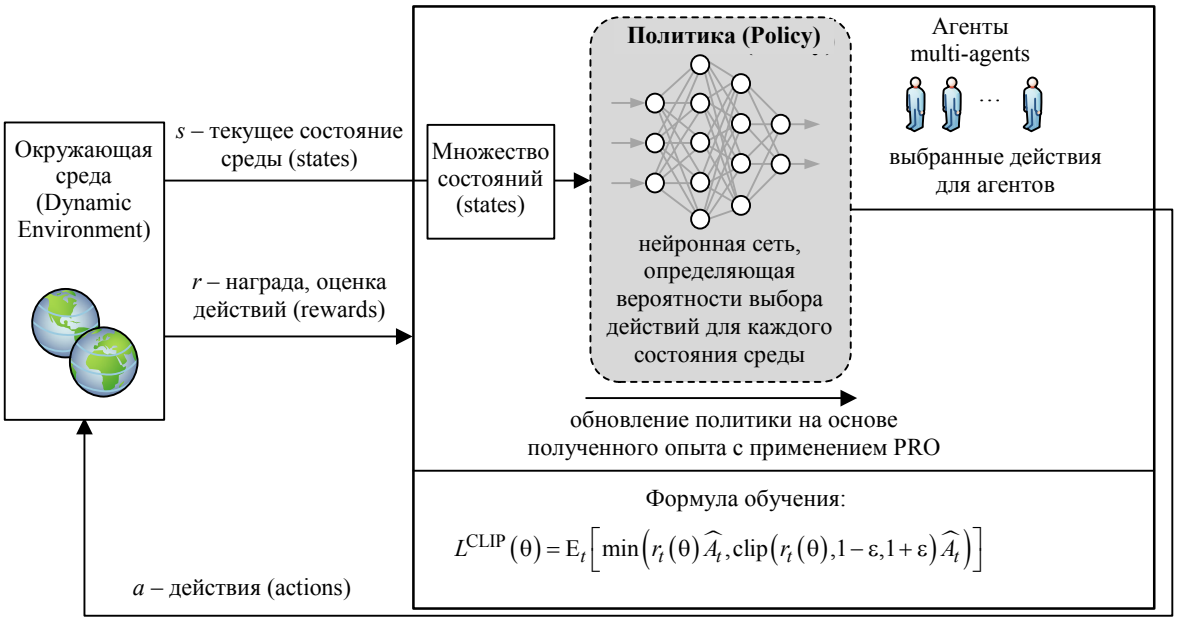

Рис. 4. Схема работы алгоритма с оптимизацией политики типа Proximal Policy Optimization
Fig. 4. The scheme of the algorithm with optimization of the policy type Proximal Policy Optimization

ятностного соотношения $r_t(\theta)$ в диапазоне от $1 - \varepsilon$ до $1 + \varepsilon$.

В архитектуре алгоритма обучения выделяют:

– *политику* – нейронную сеть, аппроксимирующую распределение вероятностей для действий;

– *критику* – нейронную сеть, оценивающую ценность состояния для вычисления преимущества;

– *clipping* – ограничение изменения политики между обновлениями.

Примеры применения Proximal Policy Optimization в логистических системах принятия решений.

Proximal Policy Optimization (PPO) идеален для сложных задач, требующих точной и стабильной политики, например управление движением автономных транспортных средств в сложных городских условиях, где требуется стабильное и эффективное обучение при больших пространствах состояний и действий. Примеры:

1. *Оптимизация маршрутов доставки* [17]. PPO может быть использован для обучения агентов, оптимизирующих маршруты доставки в режиме реального времени. Агенты учитывают такие факторы, как текущий трафик, погодные условия и сроки доставки.

2. *Управление автоматизированными транспортными средствами* (Automated Guided Vehicle, AGV) на складе [18]. PPO применяется для оптимизации планирования и маршрутизации AGV в сложных складских средах.

Входными данными в данном случае являются: текущее положение всех AGV, список задач, требующих выполнения, с их местоположением и приоритетом, карта логистического центра или склада, характеристики AGV (скорость, грузоподъемность и т. д.), текущее состояние выполнения задач.

Обработка данных происходит следующим образом: алгоритм использует нейронную сеть для представления политики принятия решений. Сеть обрабатывает текущее состояние системы и выдает вероятности выбора каждого возможного действия. Действие выбирается на основе этих вероятностей. PPO использует метод клипированной суррогатной функции для обновления параметров сети, что обеспечивает более стабильное обучение.

На выходе алгоритм предоставляет: оптимизированную политику планирования AGV, решения о назначении задач конкретным AGV, оценку эффективности текущей политики (например, общее время выполнения задач, использование ресурсов).

Этот алгоритм позволяет эффективно оптимизировать планирование AGV в сложных логистических системах, адаптируясь к динамическим изменениям в среде и обеспечивая баланс между производительностью и стабильностью обучения.

3. *Оптимизация загрузки транспортных средств* [19]. PPO может быть использован для разработки стратегий эффективной загрузки грузовиков или контейнеров, максимизируя использование пространства и учитывая ограничения по весу.

Алгоритм может принимать различные данные в зависимости от конкретной задачи, например *входными данными* могут быть: уровни запасов в различных узлах цепочки поставок, прогнозы спроса, производственные мощности, транспортные ограничения, затраты на хранение, производство и транспортировку, сроки доставки грузов.

Порядок обработки данных: PPO использует нейронную сеть для представления политики принятия решений. Сеть обрабатывает текущее состояние системы и выдает распределение вероятностей для возможных действий. Действие выбирается на основе этого распределения. PPO использует клипированную суррогатную функцию цели для обновления параметров сети, что обеспечивает более стабильное обучение.

На выходе алгоритм предоставляет: оптимизированную политику принятия решений в цепочке поставок. Конкретные решения (например, объемы заказов, маршруты доставки). Оценку эффективности текущей политики (например, общие затраты, уровень обслуживания клиентов).

PPO обеспечивает хороший баланс между исследованием новых стратегий и использованием уже найденных эффективных решений, что делает его особенно подходящим для сложных и динамичных сред, характерных для современных цепочек поставок.

4. *Управление запасами* [20]. PPO позволяет создавать адаптивные стратегии управления запасами, учитывающие колебания спроса и оптимизирующие уровни запасов.

5. *Оптимизация энергопотребления в логистических системах* [21]. PPO может быть использован для разработки энергоэффективных стратегий управления автономными транспортными средствами и складскими системами.

6. *Многоагентное управление складом* [22]. PPO применим в системах управления крупными автоматизированными складами, где множество агентов (роботов) должны координировать свои действия для эффективного выполнения задач.

Алгоритм в данном случае принимает следующие данные на входе: текущее положение всех роботов, карта склада с расположением товаров и зарядных станций, список заказов, требующих выполнения, уровень заряда батареи каждого ро-

бота, состояние загруженности роботов, информация о препятствиях и других роботах в зоне видимости.

Обработка данных происходит следующим образом: для каждого робота используется нейронная сеть, представляющая его политику принятия решений. Сеть обрабатывает локальное состояние робота и выдает распределение вероятностей для возможных действий. Действие выбирается на основе этого распределения. PPO использует клипированную суррогатную функцию цели для обновления параметров сети каждого робота, обеспечивая стабильное обучение в многоагентной среде.

На выходе алгоритм предоставляет: оптимизированные политики принятия решений для каждого робота, конкретные решения о действиях роботов (перемещение, подбор товаров, зарядка), оценку эффективности текущей коллективной политики (например, время выполнения заказов, энергоэффективность).

PPO в многоагентном контексте обеспечивает стабильное обучение и эффективное сотрудничество между роботами, что критически важно для оптимизации складских операций.

Сравнительный анализ рассмотренных алгоритмов обучения представлен в таблице.

Рекомендации по применению методов обучения типа Policy Optimization для логистических систем принятия решений.

Исследование неизвестной среды. Для задач, связанных с исследованием неизвестной среды, где структура и функционирование логистической среды не полностью известны, рекомендуется использовать методы Policy Optimization, например Proximal Policy Optimization. Эти алгоритмы могут эффективно исследовать различные действия и обновлять свои политики на основе получаемых вознаграждений.

Обучение оффлайн. В случае, когда доступен большой объем данных или симуляций, методы Policy Optimization могут быть эффективны для обучения без активного взаимодействия с реальной логистической средой. Это позволяет минимизировать риски и оптимизировать стратегии до внедрения в реальные условия.

Эксплуатация и исследование. Методы Policy Optimization, например PPO, имеют механизм для балансировки между эксплуатацией известных знаний (максимизация вознаграждений) и исследованием новых действий (обновление политики для улучшения стратегии).

Выбор подходящего метода обучения.

Для начала использования. В случае, когда точная модель окружающей среды неизвестна или она меняется со временем, целесообразно начать с методов Policy Optimization, таких как PPO, что может быть разумным выбором для первоначального исследования и определения базовых стратегий.

Для стабильного обучения. Если необходимо обеспечить стабильность и сходимость в процессе обучения и учитывать текущую стратегию, предпочтительно использование методов Actor-Critic.

Для сложных сред с большим пространством действий. Для задач с высокой размерностью данных или сложными взаимодействиями, где необходима адаптация и обучение на основе больших объемов информации, подходящим выбором будет использование методов Policy Optimization, основанных на глубоких нейронных сетях, особенно если есть доступ к мощным вычислительным ресурсам для обучения.

Ключевые различия между алгоритмами типа Policy Optimization
Key differences between the algorithms Policy Optimization

Характеристика	Policy Gradient методы (REINFORCE)	Proximal Policy Optimization (PPO)
Обучение политики	Прямое обучение политики	Прямое обучение политики с ограничением изменения
Градиентное обновление	На основе градиента ожидаемого вознаграждения	На основе функции потерь с ограничением (клиппинг)
Вариативность	Высокая вариативность градиента	Сниженная вариативность благодаря ограничению
Стабильность	Меньше стабильности, может требовать тонкой настройки	Высокая стабильность, менее чувствителен к выбору гиперпараметров
Применение	Простые задачи, задачи с меньшим пространством состояний	Сложные задачи, задачи с большим пространством состояний
Преимущества	Простота реализации, минимальные требования к ресурсам	Улучшенная стабильность и эффективность, подходит для непрерывных действий
Недостатки	Может приводить к нестабильным обновлениям и требует большого количества данных	Сложность реализации, большие вычислительные затраты

Заключение. Развитие технологий и рост объема генерируемых и обрабатываемых в ходе логистических процессов данных ведет к их усложнению и требует внедрения передовых решений для повышения их эффективности. В этом контексте методы машинного обучения, в частности методы обучения с подкреплением, играют все более значимую роль. Проведенный обзор и анализ показали, что оптимизация стратегий в алгоритмах обучения с подкреплением активно развивается и имеет высокий потенциал для улучшения эффективности логистических систем принятия решений, что обусловлено их ключевыми преимуществами: адаптивностью к динамическим условиям; оптимизацией долгосрочных результатов; способностью справляться с неопределенностью; масштабируемостью для сложных систем. Подробно проанализированы два наиболее популярных класса методов обучения с подкреплением и оптимизацией стратегии: политика градиентного подъема (Policy Gradient Methods) и обучение стратегии с учетом ограничений для обновлений (Proximal Policy Optimization). Проанализированы известные примеры применения данных методов в логистике, выявлены их особенности, преимущества и недостатки, типы решаемых задач и пр.

Перспективны следующие направления дальнейших исследований: адаптация алгоритмов к специфическим задачам логистики, повышение их интерпретируемости. Полученные в их ходе результаты могут стать основой для разработки новых алгоритмов и их практического применения в различных секторах логистики и управления цепями поставок в задачах принятия решений.

Список литературы

1. Развитие науки и научно-образовательного трансфера логистики / И. Л. Андреевский, И. Д. Афанасенко, С. Е. Барыкин, В. В. Борисова, Н. А. Верзун и др.; под науч. ред. д-ра экон. наук проф. В. В. Щербакова. СПб.: изд-во СПбГЭУ, 2019. 220 с.

2. Полуэктова З. С. Логистические системы в АПК: Основные элементы и факторы, влияющие на выбор модели логистической системы // Матер. XIX Междунар. науч.-практ. конф. «Логистика – Евразийский мост». Красноярск: Красноярский гос. аграрный ун-т, 2024. С. 331–315.

3. Кудрявцева С. С., Шинкевич А. И. Основные тренды развития логистики складирования в условиях цифровой экономики // Логистические системы в глобальной экономике. 2019. № 9. С. 126–129.

4. Гулягина О. С. Развитие логистического потенциала национальных цепей поставок с целью их интеграции в глобальные цепи поставок // Вестн. Полоцкого гос. ун-та. Сер. D. Экономические и юридические науки. 2021. № 14. С. 49–52. doi: 10.52928/2070-1632-2021-59-14-49-52.

5. Лебедев Е. А., Карцева Е. С., Зверева А. Г. Организация цифровых цепей поставок // Евразийский союз ученых. 2018. № 4–6(49). С. 59–62.

6. Байбик Г. Л. Стратегическая роль логистики в повышении экономической безопасности и операционной эффективности предприятия // Отходы и ресурсы. 2024. Т. 11, № 1. С. 1–10. doi: 10.15862/21ECOR124. URL: https://resources.today/PDF/21ECOR124.pdf (дата обращения: 23.01.2025).

7. Черкасов Д. Ю., Иванов В. В. Машинное обучение // Наука, техника и образование. 2018. № 5(46). С. 85–87.

8. Бождай А. С., Евсеева Ю. И., Артамонов Д. В. Использование машинного обучения с подкреплением в создании самоадаптивного программного обеспечения // Изв. вузов. Поволжский регион. Технические науки. 2019. № 3(51). С. 58–68. doi: 10.21685/2072-3059-2019-3-5.

9. Ротова О. М., Шибанова А. Д. Обучение с подкреплением: Введение // Теория и практика современной науки. 2020. № 1(55). С. 477–482.

10. Szepesvári C. Algorithms for reinforcement learning. San Rafael: Morgan & Claypool Publishers, 2010. 103 p.

11. Policy gradient methods for reinforcement learning with function approximation / R. S. Sutton, D. McAllester, S. Singh, Y. Mansour // Advances in Neural Information Processing Systems. 2000. Vol. 12. P. 1057–1063.

12. Deep Reinforcement learning for dynamic flexible job shop scheduling with random job arrival / J. Chang, D. Yu, Y. Hu, W. He, H. Yu // Proc. 2022. Vol. 10(4). P. 1–20. doi: 10.3390/pr10040760.

13. Reinforcement learning for solving the vehicle routing problem / M. Nazari, A. Oroojlooy, L. Snyder, M. Takác // 32nd conf. on Neural Information Proc. Systems (NeurIPS 2018). Montréal, Canada. 2018. P. 9861–9871. URL: https://arxiv.org/abs/1802.04240 (дата обращения: 23.01.2025).

14. Chen X., Wang X. Applying reinforcement learning to inventory management in a blood supply chain // European J. of Operational Research. 2018. Vol. 270(3). P. 1029–1040. URL: https://www.sciencedirect.com/science/article/pii/S0377221718306344 (дата обращения: 23.01.2025).

15. Williams R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning // Machine Learning. 1992. Vol. 8. P. 229–256.

16. Understanding policy gradient algorithms: A sensitivity-based approach / Shuang Wu, Ling Shi, Jun Wang, Guangjian Tian // Proc. of the 39th Int. Conf. on Machine Learning, PMLR 162. Baltimore, Maryland, USA. 2022. P. 24131–24149.

17. Pan W., Liu Sh. Q. Deep reinforcement learning for the dynamic and uncertain vehicle routing problem // Appl. Intelligence. 2023. Vol. 53(1). P. 405–422. doi: 10.1007/s10489-022-03456-w.

18. Wang F., Lu Z., Zhang Yu. Research on intelligent dynamic scheduling algorithm for automated guided vehicles in container terminal based on deep reinforcement learning // 2023 IEEE Intern. Conf. on Mechatronics and Automation (ICMA). Harbin, Heilongjiang, China: IEEE, 2023. doi: 10.1109/ICMA57826.2023.10215866.

19. Deep reinforcement learning for demand driven services in logistics and transportation systems: A survey / Z. Zong, J. Wang, T. Feng, T. Xia, D. Jin, Y. Li // ACM Comput. Surv. 2024. Vol. 1, no. 1. P. 1–21. URL: https://www.semanticscholar.org/reader/517f6fe6dd05a891d837fd2a9fd81aa7954c56fd (дата обращения: 23.01.2025).

20. Li Yang, Sathishkumar V. E., Adhiyaman Manickam M. Information retrieval and optimization in distribution and logistics management using deep reinforcement learning // Intern. J. of Information Systems and Supply Chain Management (IJISSCM). 2023. Vol. 16, no. 1. P. 1–19. doi: 10.4018/IJISSCM.316166 (дата обращения: 23.01.2025).

21. A Survey of deep reinforcement learning algorithms for motion planning and control of autono-mous vehicles / F. Ye, Sh. Zhang, P. Wang, Ch.-Y. Chan // IEEE Intelligent Vehicles Symp. (IV). Nagoya, Japan: IEEE, 2021. P. 1–8. doi: 10.1109/IV48863.2021.9575880.

22. Scalable multi-agent reinforcement learning for warehouse logistics with robotic and human co-workers / A. Krnjaic, R. D. Steleac, J. D. Thomas, G. Papoudakis, L. Schäfer, A. W. Keung To, K.-Ho Lao, M. Cubuktepe, M. Haley, P. Börsting, S. V. Albrecht. 2022. P. 1–8. URL: https://arxiv.org/html/2212.11498v3 (дата обращения: 23.01.2025).

Информация об авторах

Салиева Аделина Рустамовна – аспирант гр. 3933, кафедра информационных систем СПбГЭТУ «ЛЭТИ».
E-mail: rustamovna.a3@gmail.com

Верзун Наталья Аркадьевна – канд. техн. наук, доцент кафедры информационных систем СПбГЭТУ «ЛЭТИ».
E-mail: verzun.n@unecon.ru
https://orcid.org/0000-0002-0126-2358

Колбанев Михаил Олегович – д-р техн. наук, профессор кафедры информационных систем СПбГЭТУ «ЛЭТИ».
E-mail: mokolbanev@mail.ru
https://orcid.org/0000-0003-4825-6972

References

1. Razvitie nauki i nauchno-obrazovatel'nogo transfera logistiki / I. L. Andreevskij, I. D. Afanasenko, S. E. Barykin, V. V. Borisova, N. A. Verzun i dr.; pod nauchnoj red. d-ra jekon. nauk prof. V. V. Shherbakova. SPb.: izd-vo SPbGJeU, 2019. 220 s. (In Russ.).

2. Polujektova Z. S. Logisticheskie sistemy v APK: Osnovnye jelementy i faktory, vlijajushhie na vybor modeli logisticheskoj sistemy // Mater. XIX Mezhdunar. nauch.-prakt. konf. «Logistika – Evrazijskij most». Krasnojarsk: Krasnojarskij gos. agrarnyj un-t, 2024. S. 331–315. (In Russ.).

3. Kudrjavceva S. S., Shinkevich A. I. Osnovnye trendy razvitija logistiki skladirovanija v uslovijah cifrovoj jekonomiki // Logisticheskie sistemy v global'noj jekonomike. 2019. № 9. S. 126–129. (In Russ.).

4. Guljagina O. S. Razvitie logisticheskogo potenciala nacional'nyh cepej postavok s cel'ju ih integracii v global'nye cepi postavok // Vestn. Polockogo gos. un-ta. Ser. D. Jekonomicheskie i juridicheskie nauki. 2021. № 14. S. 49–52. doi: 10.52928/2070-1632-2021-59-14-49-52. (In Russ.).

5. Lebedev E. A., Karceva E. S., Zvereva A. G. Organizacija cifrovyh cepej postavok // Evrazijskij sojuz uchenyh. 2018. № 4–6(49). S. 59–62. (In Russ.).

6. Bajbik G. L. Strategicheskaja rol' logistiki v povyshenii jekonomicheskoj bezopasnosti i operacionnoj jeffektivnosti predprijatija // Othody i resursy. 2024. T. 11, № 1. S. 1–10. doi: 10.15862/21ECOR124. URL: https://resources.today/PDF/21ECOR124.pdf (data obrashenija: 23.01.2025). (In Russ.).

7. Cherkasov D. Ju., Ivanov V. V. Mashinnoe obuchenie // Nauka, tehnika i obrazovanie. 2018. № 5(46). S. 85–87. (In Russ.).

8. Bozhdaj A. S., Evseeva Ju. I., Artamonov D. V. Ispol'zovanie mashinnogo obuchenija s podkrepleniem v sozdanii samoadaptivnogo programmnogo obespechenija // Izv. vuzov. Povolzhskij region. Tehnicheskie nauki. 2019. № 3(51). S. 58–68. doi: 10.21685/2072-3059-2019-3-5. (In Russ.).

9. Rotova O. M., Shibanova A. D. Obuchenie s podkrepleniem: Vvedenie // Teorija i praktika sovremennoj nauki. 2020. № 1(55). S. 477–482. (In Russ.).

10. Szepesvári C. Algorithms for reinforcement learning. San Rafael: Morgan & Claypool Publishers, 2010. 103 p.

11. Policy gradient methods for reinforcement learning with function approximation / R. S. Sutton, D. McAllester, S. Singh, Y. Mansour // Advances in Neural Information Processing Systems. 2000. Vol. 12. P. 1057–1063.

12. Deep reinforcement learning for dynamic flexible job shop scheduling with random job arrival / J. Chang, D. Yu, Y. Hu, W. He, H. Yu // Processes. 2022. Vol. 10(4). P. 1–20. doi: 10.3390/pr10040760.

13. Reinforcement learning for solving the vehicle routing problem / M. Nazari, A. Oroojlooy, L. Snyder, M. Takác // 32nd conf. on Neural Information Proc. Systems (NeurIPS 2018). Montréal, Canada. 2018. P. 9861–9871. URL: https://arxiv.org/abs/1802.04240 (data obrashenija: 23.01.2025).

14. Chen X., Wang X. Applying reinforcement learning to inventory management in a blood supply chain // European J. of Operational Research. 2018. Vol. 270(3). P. 1029–1040. URL: https://www.sciencedirect.com/science/article/pii/S0377221718306344 (data obrashenija: 23.01.2025).

15. Williams R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning // Machine Learning. 1992. Vol. 8. P. 229–256.

16. Understanding policy gradient algorithms: A sensitivity-based approach / Shuang Wu, Ling Shi, Jun Wang, Guangjian Tian // Proc. of the 39th Int. Conf. on Machine Learning, PMLR 162. Baltimore, Maryland, USA. 2022. P. 24131–24149.

17. Pan W., Liu Sh. Q. Deep reinforcement learning for the dynamic and uncertain vehicle routing problem // Appl. Intelligence. 2023. Vol. 53(1). P. 405–422. doi: 10.1007/s10489-022-03456-w.

18. Wang F., Lu Z., Zhang Yu. Research on intelligent dynamic scheduling algorithm for automated guided vehicles in container terminal based on deep reinforcement learning // 2023 IEEE Intern. Conf. on Mechatronics and Automation (ICMA). Harbin, Heilongjiang, China: IEEE, 2023. doi: 10.1109/ICMA57826.2023.10215866.

19. Deep reinforcement learning for demand driven services in logistics and transportation systems: A survey / Z. Zong, J. Wang, T. Feng, T. Xia, D. Jin, Y. Li // ACM Comput. Surv. 2024. Vol. 1, no. 1. P. 1–21. URL: https://www.semanticscholar.org/reader/517f6fe6dd05a891d837fd2a9fd81aa7954c56fd (data obrashenija: 23.01.2025).

20. Li Yang, Sathishkumar V. E., Adhiyaman Manickam M. Information retrieval and optimization in distribution and logistics management using deep reinforcement learning // Intern. J. of Information Systems and Supply Chain Management (IJISSCM). 2023. Vol. 16, no. 1. P. 1–19. doi: 10.4018/IJISSCM.316166 (data obrashenija: 23.01.2025).

21. A survey of deep reinforcement learning algorithms for motion planning and control of autonomous vehicles / F. Ye, Sh. Zhang, P. Wang, Ch.-Y. Chan // IEEE Intelligent Vehicles Symp. (IV). Nagoya, Japan: IEEE, 2021. P. 1–8. doi: 10.1109/IV48863.2021.9575880.

22. Scalable Multi-agent reinforcement learning for warehouse logistics with robotic and human co-workers / A. Krnjaic, R. D. Steleac, J. D. Thomas, G. Papoudakis, L. Schäfer, A. W. Keung To, K.-Ho Lao, M. Cubuktepe, M. Haley, P. Börsting, S. V. Albrecht. 2022. P. 1–8. URL: https://arxiv.org/html/2212.11498v3 (data obrashenija: 23.01.2025).

Information about the authors

Adelina R. Salieva – postgraduate student gr. 3933, Department of Information systems, Saint Petersburg Electrotechnical University.
E-mail: rustamovna.a3@gmail.com

Natalia A. Verzun – Cand. Sci. (Eng.), Associate Professor of the Department of Information systems, Saint Petersburg Electrotechnical University.
E-mail: verzun.n@unecon.ru
https://orcid.org/0000-0002-0126-2358

Mikhail O. Kolbanev – Dr Sci. (Eng.), Professor of the Department of Information systems, Saint Petersburg Electrotechnical University.
E-mail: mokolbanev@mail.ru
https://orcid.org/0000-0003-4825-6972

Статья поступила в редакцию 10.12.2024; принята к публикации после рецензирования 26.01.2025; опубликована онлайн 28.03.2025.

Submitted 10.12.2024; accepted 26.01.2025; published online 28.03.2025.