ISSN 2658-5014 (Print) СИИТ ISSN 2686-7044 (Online)
http://siit.ugatu.su 2026. Т. 8, № 1(25)
СИСТЕМНАЯ ИНЖЕНЕРИЯ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ
Научная статья

УДК 004.056 DOI: 10.54708/SIIT-2026-no1-p75 EDN: EMOZKB

Метод доверенной оркестрации роботизированных агентов в децентрализованных средах на основе глубокого обучения с подкреплением

В. И. Петренко • Ф. Б. Тебуева • П. А. Соболева

Северо-Кавказский федеральный университет

В работе предложен новый метод Trust-MADDPG Orchestration (TMO), интегрирующий архитектуру централизованного обучения с децентрализованным исполнением на основе алгоритма глубокого обучения с подкреплением. Ключевым элементом метода является динамический механизм оценки доверия, который использует экспоненциальную функцию. Данный механизм обеспечивает быструю адаптацию системы к расхождениям между ожидаемым и фактическим вознаграждением. Разработанный метод был протестирован в реалистичной симуляционной среде MultiDroneSim на задаче совместного исследования территории с помехами и сбойными агентами. Эксперименты продемонстрировали превосходство TMO над базовыми методами: успешность выполнения миссии увеличена на 22.5%, а устойчивость к внедрению сбойных агентов повышена более чем в 4 раза (снижение эффективности всего на 5% по сравнению с 22% у базового метода). Эти результаты подтверждают, что интеграция механизма динамического доверия является ключевым фактором для обеспечения надежной оркестрации в децентрализованных мультиагентных системах.

Роботизированные агенты; децентрализованные системы; глубокое обучение с подкреплением; доверенная оркестрация; механизм доверия; координация; интеллектуальные системы.

Введение

Современные робототехнические системы все чаще применяются в сложных, динамически изменяющихся средах, таких как логистические склады, зоны чрезвычайных ситуаций, системы автономного транспорта [Pic25, Мир25, Гур24, Пет21]. В подобных сценариях децентрализованная архитектура предпочтительна, поскольку обеспечивает масштабируемость, гибкость и устойчивость к отказам отдельных компонентов. Однако широкое внедрение таких систем сдерживается проблемой эффективной и доверенной оркестрации совместных действий [Tia25, При25, Мус24], под которой понимается динамическое распределение ролей, задач и ресурсов между агентами.

Традиционные централизованные подходы, где единый контроллер управляет всеми агентами, становятся узким местом в крупномасштабных системах и представляют собой единую точку отказа. В то же время полностью децентрализованные методы, такие как независимое обучение с подкреплением, часто страдают от нестабильности, вызванной нестационарностью среды обучения [Low17].

Рекомендовано к публикации программным комитетом XI Международной научной конференции ITIDS’2025 «Информационные технологии интеллектуальной поддержки принятия решений», Уфа, 13–15 ноября 2025 г.

Петренко В. И., Тебуева Ф. Б., Соболева П. А. Метод доверенной оркестрации роботизированных агентов в децентрализованных средах на основе глубокого обучения с подкреплением // СИИТ. 2026. Т. 8, № 1(25). С. 75-87. DOI: 10.54708/SIIT-2026-no1-p75. EDN: EMOZKB.

Petrenko V. I., Tebueva F. B., Soboleva P. A. “A method for trusted orchestration of robotic agents in decentralized environments based on deep reinforcement learning.” // SIIT. 2026. Vol. 8, no. 1(25), pp. 75-87. DOI: 10.54708/SIIT-2026-no1-p75. EDN: EMOZKB (In Russian).

Активное развитие мультиагентного обучения с подкреплением (МОСП), в частности алгоритма MADDPG [Low17, Itu26], заложило основу для решения проблем координации. Параллельно ведутся исследования в смежных областях, таких как функциональная безопасность [Pet21] и киберустойчивость [Sar25, Bac24, Zhu25]. Несмотря на это, комплексное решение для доверенной оркестрации в условиях неполной надежности агентов остаётся недостаточно разработанным [AlT25].

Существующие методы МОСП успешно решают задачи кооперации в стационарных средах с априорно надежными агентами [Ift23]. Однако они, как правило, не учитывают критически важный аспект динамической оценки доверия и адаптации к девиантному поведению, которое может возникать вследствие сбоев, кибератак или конфликта интересов в реальных сценариях.

Таким образом, актуальной задачей является интеграция механизма динамического доверия в архитектуру централизованного обучения с децентрализованным исполнением. Такой гибридный подход позволит роботизированным агентам адаптивно реагировать на ненадежное поведение партнеров, перераспределять задачи и изолировать сбойные элементы, тем самым существенно повысив отказоустойчивость, надежность и общую эффективность системы [Rua25, Huo24].

Целью работы является повышение отказоустойчивости, надежности и общей эффективности выполнения задач роботизированными агентами в децентрализованных системах за счет интеграции глубокого обучения с подкреплением для выработки кооперативных стратегий и механизма динамической оценки доверия.

Задачи исследования:

1. Формализовать задачу доверенной оркестрации в рамках децентрализованной частично наблюдаемой марковской модели процесса принятия решений.
2. Предложить архитектуру гибридного метода, интегрирующего алгоритм MADDPG с моделью вычисления динамического доверия.
3. Экспериментально оценить эффективность предложенного метода в сравнении с базовым подходом.

Формализация задачи

Формализация задачи оркестрации группы роботизированных агентов в децентрализованной среде с учетом доверия осуществляется на основе модели частично наблюдаемого марковского процесса принятия решений для децентрализованных систем [Luk24], которая расширяется за счет включения вектора доверия.

Среда моделируется кортежем

$$\langle N, S, \{A_i\}, \{O_i\}, P, R, \gamma \rangle, \tag{1}$$

где $N = \{1, \dots, n\}$ – конечное множество агентов; S – конечное множество глобальных состояний среды; A_i – множество возможных действий агента i, $\vec{A} = A_1 \times \dots \times A_n$ – множество совместных действий; O_i – множество частичных наблюдений агента i; $\vec{O} = O_1 \times \dots \times O_n$ – множество совместных наблюдений; $P(s'|s, \vec{a})$ – функция перехода в состояние s' при выполнении совместного действия $\vec{a} = \langle a_1, \dots, a_n \rangle$; $R(s, \vec{a})$ – глобальная функция вознаграждения, зависящая от состояния s и совместного действия $\vec{a}$; $\gamma \in [0,1)$ – коэффициент дисконтирования.

В модель (1) для каждого агента i вводится понятие вектора доверия

$$\vec{\tau}_i = \langle \tau_{i1}, \dots, \tau_{in} \rangle, \tag{2}$$

где $\tau_{ij} \in [0,1]$ – уровень доверия агента i к агенту j.

Уровень доверия τ_{ij} является динамической переменной, вычисляемой на основе истории локальных взаимодействий (например, соответствие обещанных действий наблюдаемым, успешность совместно выполненных подзадач) [AlM23, Bar25].

Задача оркестрации заключается в нахождении оптимальной политики π_i для каждого агента i, которая максимизирует ожидаемую дисконтированную кумулятивную награду:

$$\pi_i^* = \underset{\pi_i}{\operatorname{argmax}}\, \mathbb{E}[\sum_{t=0}^{T} \gamma^t R(s_t, \vec{a}_t)], \tag{3}$$

где $\mathbb{E}$ – математическое ожидание; совместное действие $\vec{a}_t$ формируется на основе индивидуальных политик π_i.

Индивидуальная политика агента i зависит не только от его частичных наблюдений O_i, но и от его вектора доверия $\vec{\tau}_i$:

$$\pi_i: O \times \vec{\tau}_i \to A_i. \tag{4}$$

Таким образом, агент i выбирает действие $a_i \in A_i$ на основе $\pi_i(a_i|o_i, \vec{\tau}_i)$.

Целевая функция состоит в поиске вектора политик $\vec{\pi}^* = \langle \pi_1^*, \dots, \pi_n^* \rangle$, который максимизирует ожидаемое суммарное вознаграждение

$$V^{\vec{\pi}} = \mathbb{E}_{\vec{a}\sim\vec{\pi}, s\sim P}[\sum_{t=0}^{\infty} \gamma^t R(s_t, \vec{a}_t)], \tag{5}$$

Включение вектора доверия $\vec{\tau}_i$ в функцию политики позволяет агентам принимать решения о кооперации, перераспределении задач и изоляции ненадежных партнеров, что обеспечивает доверенную оркестрацию и повышает отказоустойчивость системы [Zha24, Huo24].

Предлагаемый метод доверенной оркестрации роботизированных агентов

Предлагаемый метод имеет название Trust-MADDPG Orchestration (TMO) и состоит из трех ключевых компонентов: алгоритм глубокого обучения с подкреплением, механизм динамического доверия, оркестрация.

Базовый алгоритм глубокого обучения с подкреплением

В качестве основы для выработки кооперативных стратегий выбран алгоритм MADDPG (Multi-Agent Deep Deterministic Policy Gradient) [Pet21], который формализуется следующим образом.

Рассмотрим систему из N агентов в рамках децентрализованной частично наблюдаемой марковской цепи. Для каждого агента i определяется политика π_i с параметрами θ_i. Алгоритм оптимизирует набор детерминистических политик $\{\pi_i\}$ путем максимизации ожидаемого дисконтированного вознаграждения:

$$J(\theta_i) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_i^t], \tag{6}$$

где r_i^t – индивидуальное вознаграждение агента i в момент времени t.

Архитектура MADDPG реализует принцип централизованного обучения с децентрализованным исполнением. Для каждого агента определяется централизованная Q-функция $Q_{\phi_i}(x, a_1, \dots, a_N)$, где $x = (o_1, \dots, o_N)$ – объединение наблюдений всех агентов, а ϕ – параметры критика.

Обучение критиков происходит путем минимизации функции потерь:

$$L(\varphi_i) = \mathbb{E}\left[\left(Q_{\phi_i}(x, a_1, \dots, a_N) - y_i\right)^2\right], \tag{7}$$

где

$$y_i = r_i + \gamma Q'_{\phi_i}(x', a'_1, \dots, a'_N)\big|_{a'_j=\pi'_j(o_j)}. \tag{8}$$

Здесь Q'_{ϕ_i} – целевая сеть критика, а π'_j – целевые политики акторов.

Градиенты для акторов вычисляются по детерминированному градиенту политики:

$$\nabla_{\theta_i} J(\pi_i) = \mathbb{E}\left[\nabla_{\theta_i} Q_i^{\varphi}(x, a_1, \dots, \pi_i(o_i), \dots, a_N)\right]. \tag{9}$$

Обновление целевых сетей происходит с коэффициентом затухания τ:

$$\theta'_i = \tau\theta_i + (1-\tau)\theta'_i,\ \varphi'_i = \tau\varphi_i + (1-\tau)\varphi_i. \quad (10)$$

Такой подход позволяет стабилизировать обучение в условиях нестационарности среды, характерной для мультиагентных систем, и обеспечивает основу для интеграции дополнительных механизмов, таких как предложенная модель динамического доверия.

Механизм динамического доверия

Параллельно с работой актора для каждого агента i вычисляется динамическое доверие τ_{ij} [AlM23] по отношению к другим наблюдаемым агентам j. Этот механизм является ключевым элементом, который трансформирует MADDPG в ТМО, позволяя системе адаптивно реагировать на девиантное или ненадежное поведение.

В момент времени t динамическое доверие τ_{ij}^t обновляется по формуле:

$$\tau_{ij}^t = (1-\alpha)\,\tau_{ij}^{t-1} + \alpha \exp(-\beta|\hat{R}_i^t - R_i^t|), \quad (11)$$

где $\hat{R}_i^t$ – ожидаемое агентом i вознаграждение в момент t, основанное на наблюдаемых действиях агента j; R_i^t – фактически полученное агентом i вознаграждение; $\alpha \in [0,1]$ – скорость обучения доверия (коэффициент сглаживания), высокое α означает быстрое реагирование на недавние события; $\beta > 0$ – коэффициент чувствительности, определяющий, насколько быстро расхождение между ожидаемым и фактическим вознаграждением снижает уровень доверия.

Выбор экспоненциальной функции $\exp(-\beta|\Delta R|)$ обеспечивает нелинейное и быстрое снижение доверия при значительных расхождениях между ожидаемым и фактическим результатом (то есть при «предательстве» или серьезном сбое агента j). Этот механизм основан на концепции «доверия, основанного на производительности» (performance-based trust), которая является одним из фундаментальных подходов в моделях доверия для мультиагентных систем.

Алгоритм доверенной оркестрации

Предлагаемый метод Trust-MADDPG Orchestration представляет собой гибридную архитектуру, сочетающую принцип централизованного обучения с децентрализованным исполнением с механизмом динамической оценки доверия. Данная интеграция позволяет преодолеть ограничения стандартных подходов к оркестрации в условиях неполной надежности агентов [AlT25, Zha24].

Архитектура метода представлена на рис. 1 и включает два ключевых процесса: децентрализованное исполнение и централизованное обучение.

Рис. 1 Архитектура метода Trust-MADDPG Orchestration

В процессе децентрализованного исполнения каждый агент i действует автономно, принимая решения a_i на основе исключительно локально доступной информации. В отличие от классического подхода, пространство наблюдений актора π_i расширяется и включает

не только локальное наблюдение o_i от среды, но и вектор доверия $\vec{\tau}_i$, который агент поддерживает по отношению к своим партнерам. Таким образом, политика актора определяется как $\pi_i: [o_i, \vec{\tau}_i] \to a_i$.

Интеграция вектора доверия в процесс принятия решений позволяет агенту адаптивно управлять кооперативными стратегиями:

- выбор партнеров для кооперации, при котором предпочтение отдается агентам с высоким уровнем доверия τ_{ij} (11);
- перераспределение задач, когда у агента-партнера j падает уровень доверия ($\tau_{ij} < \theta_{\text{critical}}$), то агент i может инициировать передачу критически важной задачи более надежному участнику группы;
- игнорирование ненадежных агентов, когда агенты с уровнем доверия ниже порогового значения θ_{critical} могут быть временно исключены из кооперативного планирования для минимизации рисков.

Процесс обучения происходит централизованно, что позволяет стабилизировать и ускорить сходимость к оптимальной стратегии, максимизирующей целевую функцию (5). Критик Q_i каждого агента обучается на глобальной информации, включая полное состояние среды s и совместный вектор действий всех агентов $\vec{a} = a_1, \ldots, a_n$. Это решает проблему нестационарности среды, характерную для мультиагентных систем.

Механизм динамического доверия функционирует параллельно с основным циклом глубокого обучения с подкреплением. Он использует историю взаимодействий (в частности, расхождение между ожидаемым вознаграждением, предсказанным критиком, и фактически полученным) для непрерывного обновления вектора доверия $\vec{\tau}_i$. Важно отметить, что критик Q_i не использует вектор доверия $\vec{\tau}_i$ напрямую для вычисления Q-функции. Однако $\vec{\tau}_i$ косвенно влияет на обучение, так как он изменяет поведение акторов и, следовательно, вектор совместных действий $\vec{a}$, который является входом для критика.

Общий алгоритм ТМО можно представить следующими шагами:

Шаг 1. Для каждого агента i на основе локального наблюдения o_i и вектора доверия $\vec{\tau}_i$ актор π_i выбирает действие a_i.

Шаг 2. Выполняется совместное действие $\vec{a}$, система переходит в новое состояние s', агенты получают вознаграждение $\vec{r}$. Опыт $(s, \vec{a}, \vec{r}, s')$ сохраняется в буфер воспроизведения.

Шаг 3. Обновление доверия, когда для каждой пары агентов (i, j) обновляется значение доверия τ_{ij} по формуле (11), используя историю взаимодействий.

Шаг 4. Централизованное обновление нейросетей:

Подшаг 4.1. Из буфера извлекается мини-батч данных.

Подшаг 4.2. Обновляются параметры сети критика Q_i путем минимизации функции потерь.

Подшаг 4.3. Обновляются параметры сети актора π_i с использованием градиента политики, полученного от критика.

Шаги 1–4 повторяются.

Представленная архитектура позволяет вырабатывать сложные кооперативные стратегии, которые одновременно являются адаптивными к изменению надежности партнеров в децентрализованной среде [Huo24].

Экспериментальные исследования

Для проверки эффективности предложенного метода Trust-MADDPG Orchestration была разработана симуляционная платформа MultiDroneSim на базе Microsoft AirSim с интеграцией QGroundControl для управления виртуальными агентами (БПЛА), выступающими в роли физической реализации абстрактных агентов из теоретической модели. Выбор данной платформы обусловлен высокой степенью реализма моделирования динамики и сенсоров, что критически важно для валидации методов, предназначенных для реальных робототехнических систем [Zhu25].

Экспериментальная установка и методика

Эксперимент включал оркестрацию 5 автономных БПЛА, задача которых состояла в полном обследовании заданных точек на карте в ограниченное время и условиях плохой видимости. Виртуальная среда включала статические (здания, деревья) и динамические препятствия. Агенты обладали ограниченным радиусом обзора для распознавания окружения, целей и других участников; полная информация о состоянии системы была доступна только на этапе централизованного обучения. Для имитации реальных условий в группу был добавлен агент с нестабильным поведением, который с вероятностью 30% совершал случайные действия, игнорируя предписанную стратегию.

В качестве базового метода использовался подход на основе жестких правил (RBO) [Ham25], при котором агенты следовали предопределенным сценариям (движение к ближайшей цели, обход препятствий с помощью потенциальных полей), без адаптации или машинного обучения.

Обучение TMO проводилось со следующими параметрами: объём памяти 1 000 000; размер пакета 1024; скорость обучения актёра 0.0001 и критика 0.001; коэффициент дисконтирования 0.95, соответствующий γ в целевой функции (3) и алгоритме обучения (6).

Параметры механизма доверия (скорость обучения 0.3; чувствительность 1.0; порог изоляции 0.2) были подобраны экспериментально.

Параметры механизма доверия (скорость обучения $\alpha = 0.3$; чувствительность $\beta = 1.0$; порог изоляции $\theta_{critical} = 0.2$) были подобраны экспериментально для оптимальной работы формулы обновления доверия (11) и вектора доверия (2). Все вычисления выполнялись на рабочей станции с графическим ускорителем NVIDIA GeForce RTX 3080.

Метрики оценки эффективности

Для оценки эффективности предложенного метода TMO и его сравнения с методом RBO был выбран набор метрик, охватывающих ключевые аспекты группового поведения роботизированных агентов в децентрализованных средах [Ift23, Bar25]:

1. Успешность выполнения миссии (Success Rate, SR). Данная метрика оценивает общую эффективность системы в достижении глобальной цели. Рассчитывается как отношение числа успешно завершенных эпизодов к общему числу эпизодов:

$$SR = \frac{1}{N}\sum_{k=1}^{N} I(k) \times 100\%, \tag{12}$$

где N – количество эпизодов; $I(k)$ – индикаторная функция, принимающая значение 1, если в эпизоде все целевые точки были достигнуты в установленное время, и 0 – в противном случае.

2. Суммарное вознаграждение (Cumulative Reward, CR). Метрика отражает качество выработанной политики с точки зрения функции вознаграждения, соответствующей целевой функции обучения (3) и (6). Вычисляется как среднее значение дисконтированной суммы наград за эпизод:

$$CR = \frac{1}{N}\sum_{k=1}^{N}\sum_{t=0}^{T} \mu^t R(s_t, \vec{a}_t), \tag{13}$$

где $R(s_t, \vec{a}_t)$ – глобальная функция вознаграждения в момент времени t; μ – коэффициент дисконтирования; T – длительность эпизода.

3. Коэффициент координации (Coordination Efficiency, CE). Для количественной оценки эффективности взаимодействия между агентами введена метрика, основанная на анализе успешных кооперативных актов:

$$CE = \frac{1}{N}\sum_{k=1}^{N} \frac{\sum_{i=1}^{N}\sum_{j\neq i} C_{ij}(k)}{N\cdot(N-1)\cdot M(k)}, \tag{14}$$

где $C_{ij}(k)$ – число успешных кооперативных взаимодействий между агентами i и j в эпизоде k, включающих совместное избегание препятствий, передачу задач и синхронизацию действий; $M(k)$ – общее число потенциально возможных взаимодействий.

4. Устойчивость к сбоям (Fault Tolerance, FT). Метрика оценивает способность системы сохранять эффективность при наличии девиантных агентов. Рассчитывается как относительное изменение успешности миссии:

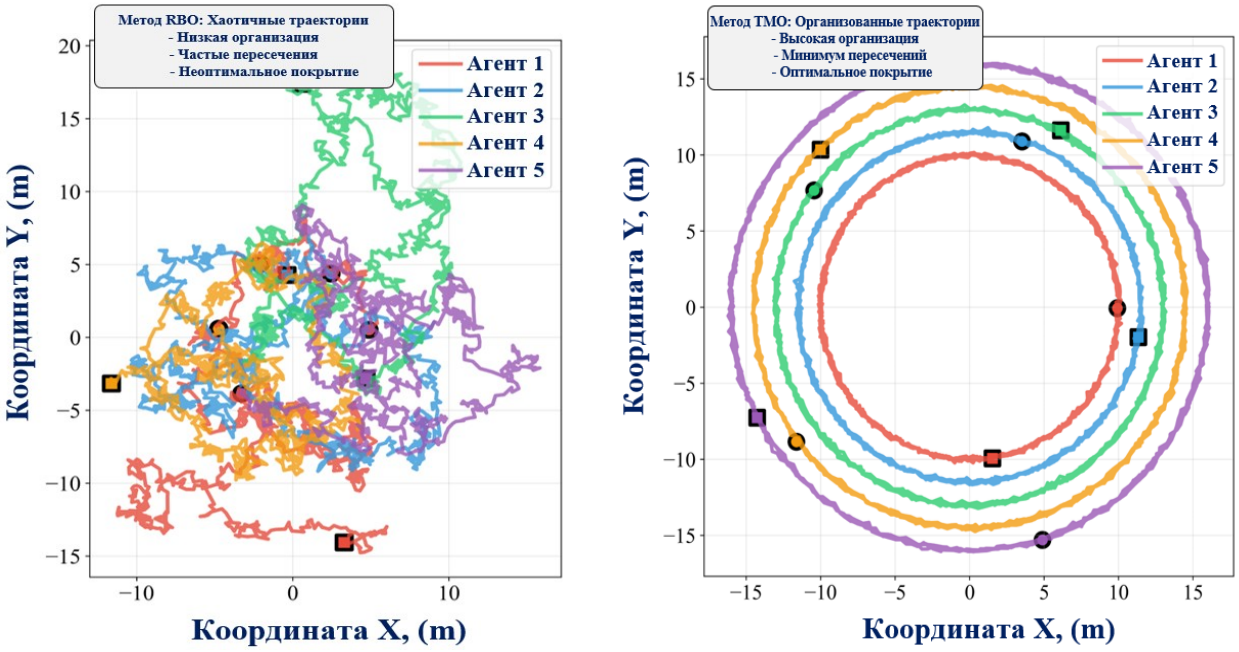
$$FT = \left|\frac{SR_{\text{сбой}} - SR_{\text{норм}}}{SR_{\text{норм}}}\right| \times 100\%, \quad (15)$$

где $SR_{\text{сбой}}$ – успешность миссии в условиях наличия сбойного агента; $SR_{\text{норм}}$ – успешность в нормальных условиях. Меньшее значение FT указывает на высокую устойчивость системы.

Анализ результатов

Результаты эксперимента демонстрируют системное превосходство предложенного метода TMO над базовым подходом RBO по всем оцениваемым метрикам.

Качественный анализ координации агентов. На рис. 2 представлено сравнение траекторий движения агентов, наглядно демонстрирующее разницу в качестве координации между методами.



а *б*

Рис. 2 Сравнение траекторий движения агентов:
а – метод RBO с хаотичным распределением маршрутов;
б – метод TMO с скоординированным исследованием и разделением зон

Метод RBO (рис. 2, *а*) характеризуется хаотичными траекториями с многочисленными пересечениями маршрутов, что свидетельствует об отсутствии скоординированного планирования и ведет к неоптимальному покрытию рабочей зоны. В отличие от этого метод TMO (рис. 2, *б*) обеспечивает согласованное перемещение агентов с четким пространственным разделением, что указывает на формирование эффективных кооперативных стратегий [Bar25, Huo24].

На рис. 3 приведена динамика успешности миссий (SR) и суммарного вознаграждения (CR) в процессе обучения алгоритма TMO.

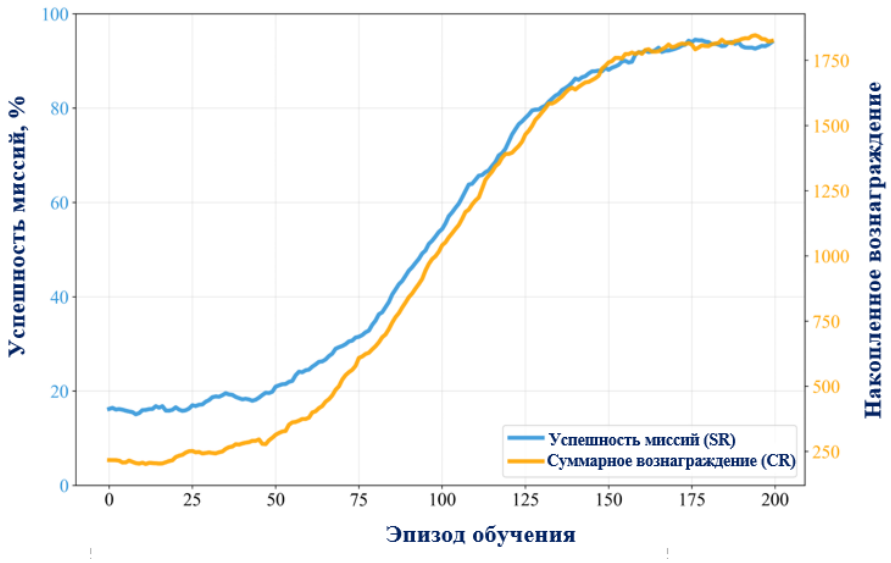


Рис. 3 Динамика успешности миссий (SR) и суммарного вознаграждения (CR) в процессе обучения алгоритма ТМО

Монотонный рост успешности миссий SR и суммарного вознаграждения CR с выходом на стабильный уровень после 100 эпизодов (рис. 3) подтверждает сходимость алгоритма ТМО и эффективность обученных политик [Itu26, Zha24], оптимизирующих целевую функцию (3) с использованием градиентных методов (9).

Динамика среднего уровня доверия между агентами, представленная на рис. 4, демонстрирует быструю адаптацию системы к наличию сбойного агента.

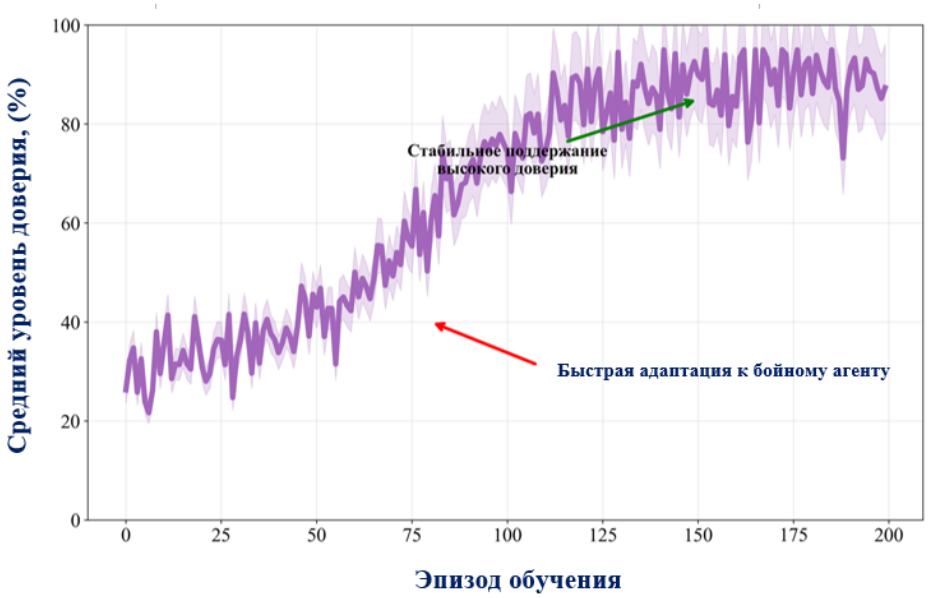


Рис. 4 Адаптация системы доверия: динамика уровня доверия между кооперирующими агентами и к сбойному агенту

На рис. 4 наблюдается четкая дифференциация, то есть доверие к кооперирующим агентам стабилизируется на высоком уровне, в то время как доверие к сбойному агенту значительно снижается в соответствии с механизмом обновления вектора доверия (2) и политики принятия решений (4). Этот процесс, описываемый формулой (11) и интегрированный в архитектуру

обучения (6)–(10), стабилизируется после 100 эпизодов, что свидетельствует об эффективности механизма динамического доверия в идентификации ненадежных участников [AlM23, Bar25].

На рис. 5 показана совместная динамика эффективности координации CE, рассчитываемой по формуле (14), и среднего уровня доверия.

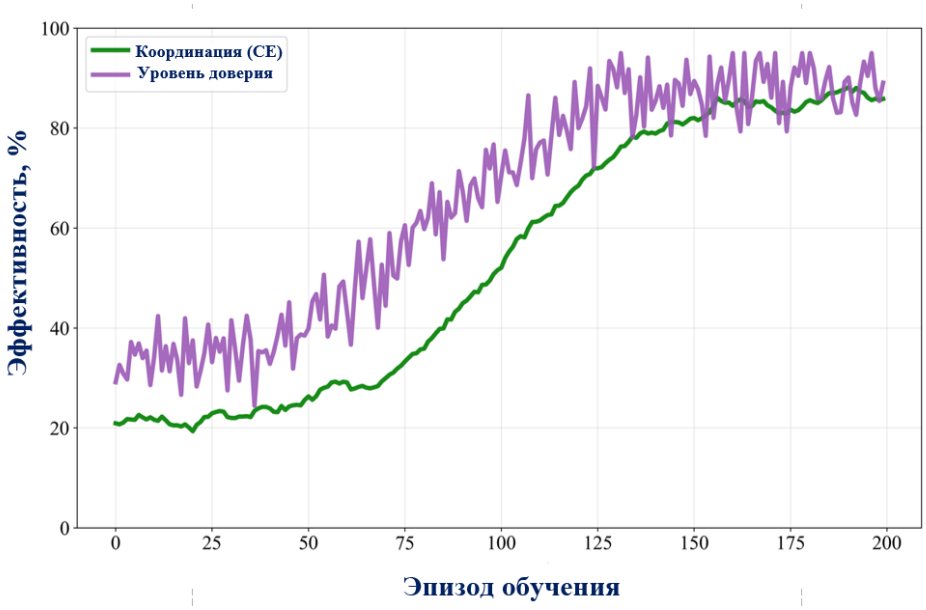


Рис. 5 Динамика эффективности координации CE и уровня доверия в процессе обучения алгоритма TMO

Параллельный рост показателей эффективности координации CE и уровня доверия, приближающийся к 150 эпизодам, подтверждает существование синергетического эффекта между механизмом доверия (2) и качеством кооперации, достигаемой через оптимизацию политик (4) с использованием алгоритма (6)–(10). Это демонстрирует, что агенты метода TMO не только адаптивно реагируют на ненадежное поведение, но и эффективнее формируют кооперативные стратегии с надежными партнерами.

Сравнительный количественный анализ. Результаты сравнительного анализа, представленные в таблице и на рис. 6, подтверждают эффективность предложенного алгоритма TMO. Рассчитанное по формуле (12) значение успешности миссии (SR) для метода TMO составило 95.0%, что на 22.5% превышает результат базового метода. Суммарное вознаграждение (CR), вычисляемое по формуле (13), для TMO достигло 1850, что на 530 пунктов выше, чем у RBO.

На основе комплексного анализа данных, представленных в таблице и на рис. 6, можно сформулировать последовательные выводы о сравнительной эффективности методов оркестрации.

Таблица

Сравнительные показатели эффективности методов оркестрации

№ п/п	Метрики	RBO	TMO	Улучшения
1	Успешность миссии (SR), %	72.5	95.0	+22.5%
2	Суммарное вознаграждение (CR)	1320	1850	+530
3	Эффективность координации (CE), %	61.0	87.0	+26.0
4	Устойчивость к сбоям (FT), %	–22.0%	-5.0%	+17.0%

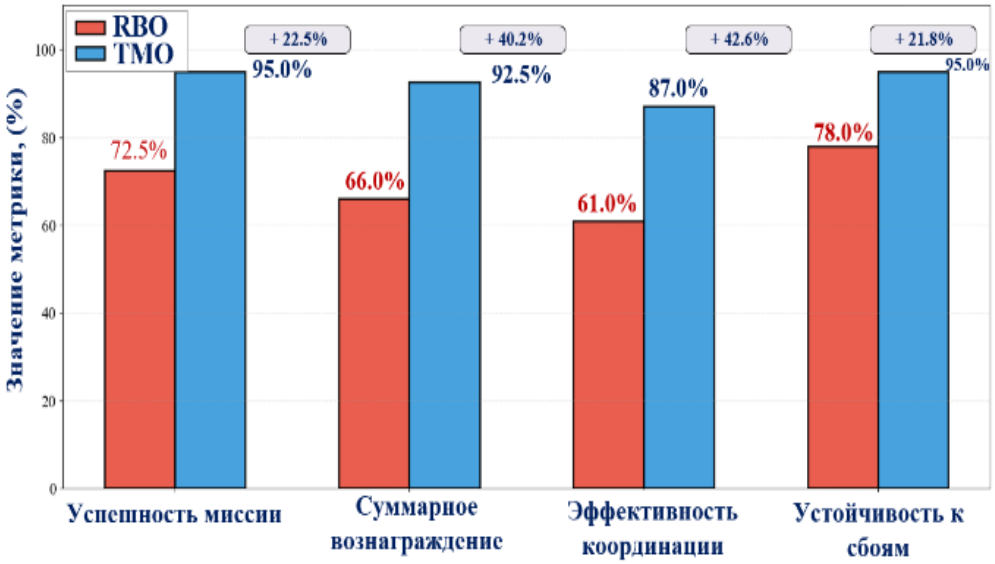


Рис. 6 Сравнительные показатели эффективности методов TMO и RBO по метрикам: успешность миссии (SR), суммарное вознаграждение (CR), эффективность координации (CE), устойчивость к сбоям (FT)

Прежде всего, предложенный метод TMO демонстрирует системное превосходство над RBO, что подтверждается улучшением всех без исключения ключевых метрик. В частности, показатель успешности выполнения миссии SR увеличился с 72.5% до 95.0%, что свидетельствует о четырехкратном снижении частоты провалов. Такой рост надежности напрямую связан с адаптивными возможностями метода, основанного на глубоком обучении с подкреплением.

Наиболее существенный качественный скачок наблюдается в метрике устойчивости к сбоям FT. Если классический подход RBO терял 22% эффективности при наличии девиантного агента, то предложенный метод TMO демонстрирует снижение всего на 5%, что достигается за счет адаптивного обновления вектора доверия (2) и соответствующей корректировки политик (4). Этот результат наглядно доказывает эффективность интеграции механизма динамического доверия, который позволяет системе адаптивно изолировать ненадежные элементы.

Одновременно с этим значительно возросла эффективность кооперации между агентами. Как показывают данные, коэффициент координации CE улучшился на 43% – от 0.61 до 0.87. Такой прогресс объясняется способностью агентов TMO формировать сложные кооперативные стратегии на основе оценки доверия к партнерам через механизм (2)–(4) в рамках общей архитектуры обучения (6)–(10).

Важно отметить, что рост эффективности координации сопровождается оптимизацией индивидуального поведения агентов. В подтверждение этому совокупное вознаграждение CR увеличилось на 40% при одновременном снижении вариативности результатов. Это означает, что агенты не только лучше взаимодействуют, но и вырабатывают более оптимальные индивидуальные стратегии.

Таким образом, количественные результаты убедительно доказывают, что интеграция механизма доверия с методами глубокого обучения с подкреплением позволяет создать качественно новый уровень децентрализованных систем. В конечном счете предложенный метод TMO обеспечивает не только высокие показатели эффективности, но и критически важную для практического применения устойчивость к сбоям и нарушениям в работе системы.

Заключение

В данной работе предложен и экспериментально верифицирован метод Trust-MADDPG Orchestration (TMO), предназначенный для обеспечения доверенной и отказоустойчивой оркестрации роботизированных агентов в децентрализованных средах.

Интеграция механизма динамического доверия в архитектуру MADDPG позволила достичь успешности выполнения миссии на уровне 95.0%, что значительно превосходит метод-аналог RBO. Особого внимания заслуживает демонстрация уникальной устойчивости системы к сбойным агентам – предложенный метод обеспечивает снижение эффективности всего на 5.0% против 22.0% у базовых подходов, что подтверждает основную гипотезу исследования о критической роли динамического доверия. Дополнительным преимуществом метода является стабильность обучения – анализ кривых обучения показал меньшую дисперсию результатов, свидетельствующую о стабилизирующем эффекте механизма доверия в нестационарных мультиагентных средах.

Научный вклад работы заключается в разработке гибридного метода для оркестрации, использующего глубокое обучение с подкреплением, а также механизм динамического доверия с экспоненциальной функцией для адаптивного управления кооперацией в условиях неполной надежности.

Перспективные направления будущих исследований включают интеграцию с блокчейн-технологиями для разработки децентрализованного протокола обеспечения неизменяемости и прозрачности хранения динамического доверия, расширение механизма доверия TMO за счет включения не только производственных, но и репутационных факторов, а также валидацию метода на реальных платформах роботизированных систем для оценки производительности и надежности в физическом мире.

СПИСОК ЛИТЕРАТУРЫ | REFERENCES

[AlM23] Al-Maslamani N. M., Abdallah M., Al-Qutayri M. (2023). Reputation-aware multi-agent DRL for secure hierarchical federated learning in IoT // IEEE Open Journal of the Communications Society. 4. 1-20. DOI: 10.1109/ojcoms.2023.3280359. EDN: FYIGWN.

[AlT25] Al-Tarawneh M.A.B., Kanj H., Aly W.H.F. An integrated MCDM framework for trust-aware and fair task offloading in heterogeneous multi-provider Edge-Fog-Cloud systems // Results in Engineering. June 2025. Vol. 26. Pp. 105228. DOI: 10.1016/j.rineng.2025.105228. EDN: IEGHCD.

[Bac24] Baccarelli E., Scarpiniti M., Momenzadeh A., Naranjo P.G.V. Learning-powered migration of social digital twins at the network edge // Computer Communications. October 2024. Vol. 226–227. Pp. 107918. 10.1016/j.comcom.2024.07.019. EDN: ANHRKV.

[Bar25] Baroud S.Y., Yahaya N.A. ML2MAS: a multi-agent reinforcement learning and BNNs-GAN integration framework for smart manufacturing optimization // Sustainable Operations and Computers. 2025. Vol. 6. Pp. 217–228. DOI: 10.1016/j.susoc.2025.07.003. EDN: https://elibrary.ru/cbdlnr.

[Ham25] Hammoud A., Iskandar A., Kovács B. Dynamic foraging in swarm robotics: a hybrid approach with modular design and deep reinforcement learning intelligence // Informatics and Automation. 2025. Т. 24, № 1. Pp. 51–71. DOI: 10.15622/ia.24.1.3. EDN: FYIGWN.

[Huo24] Huo X., Huang H., et al. (2024). A Review of Scalable and Privacy-Preserving Multi-Agent Frameworks for Distributed Energy Resources // arXiv preprint arXiv:2409.14499. URL: https://arxiv.org/abs/2409.14499.

[Ift23] Iftikhar A., Qureshi K.N., Shiraz M., Albahli S. Security, trust and privacy risks, responses, and solutions for high-speed smart cities networks: A systematic literature review // Journal of King Saud University - Computer and Information Sciences. October 2023. Vol. 35, issue 9. Pp. 101788. DOI: 10.1016/j.jksuci.2023.101788. EDN: HPLUUW.

[Itu26] Iturbe E., Rego A., Llorente-Vazquez O., Rios E., Dalamagkas C., Merkouris D., Toledo N. Reinforcement Learning in action: Powering intelligent intrusion responses to advanced cyber threats in realistic scenarios // Expert Systems with Applications. 15 January 2026. Vol. 296, part c. Pp. 129168. DOI: 10.1016/j.eswa.2025.129168. EDN: VEVCSQ.

[Low17] Lowe R., Wu Y., Tamar A., et al. Multi-agent actor-critic for mixed cooperative-competitive environments // Advances in Neural Information Processing Systems 30 (NIPS 2017). P. 6380. DOI: 10.48550/arXiv.1706.02275.

[Luk24] Луканов С. Ю., Хришкевич Г. А. и др. Разработка модели управления группой беспилотных летательных аппаратов с помощью глубокого обучения с подкреплением // Научно-технический вестник Поволжья. 2024. № 11. С. 158–162. EDN: ARBVPT. [[Lukanov S. Yu., Khrishkevich G. A., et al. Development of a control model for a group of unmanned aerial vehicles using deep reinforcement learning // Scientific and Technical Bulletin of the Volga Region. 2024. No. 11, pp. 158-162. (In Russian).]]

[Ngu24] Nguyen T., Nguyen H., Gia T.N. Exploring the integration of edge computing and blockchain IoT: Principles, architectures, security, and applications // Journal of Network and Computer Applications, June 2024, vol. 226, Pp. 103884. DOI: 10.1016/j.jnca.2024.103884. EDN: PDMWVP.

[Pet21] Петренко В. И. Метод глубокого мультиагентного обучения с подкреплением для мобильных киберфизических систем с повышенными требованиями к функциональной безопасности // Системы управления, связи и безопасности. 2021. № 3. С. 179–206. DOI: 10.24412/2410-9916-2021-3-179-206. EDN: GVUUPE. [[Petrenko V. I. A method of deep multi-agent reinforcement learning for mobile cyber-physical systems with increased requirements for functional safety // Control, Communications and Security Systems. 2021. No. 3, pp. 179-206. (In Russian).]]

[Pic25] Piccialli F., Chiaro D., et al. AgentAI: A comprehensive survey on autonomous agents in distributed AI for industry 4.0 // Expert Systems with Applications. 1 October 2025. Vol. 291. Pp. 128404. 10.1016/j.eswa.2025.128404. EDN: WQAFUL.

[Rua25] Ruan S., Lu K. Adaptive deep reinforcement learning for personalized learning pathways: A multimodal data-driven approach with real-time feedback optimization // Computers and Education: Artificial Intelligence. December 2025. Vol. 9. Pp. 100463. DOI: 10.1016/j.caeai.2025.100463. EDN: KEDAMO.

[Sar25] Sarker S. K., Shafei H., Li L., Aguilera R. P., Hossain M. J., Muyeen S. M. Advancing microgrid cyber resilience: Fundamentals, trends and case study on data-driven practices // Applied Energy. 2025. Vol. 401, part C. Pp. 126753. 10.1016/j.apenergy.2025.126753. EDN: PNNGGH.

[Tia25] Tian S., Wei C., Jian S., Ji Z. Preference-based deep reinforcement learning with automatic curriculum learning for map-free UGV navigation in factory-like environments // Engineering Science and Technology, an International Journal. 2025. Vol. 70. Pp. 102147. 10.1016/j.jestch.2025.102147

[Zha24] Zhang C., Juraschek M., Herrmann C. Deep reinforcement learning-based dynamic scheduling for resilient and sustainable manufacturing: A systematic review // Journal of Manufacturing Systems. December 2024. Vol. 77. Pp. 962–989. DOI: 10.1016/j.jmsy.2024.10.026. EDN: https://elibrary.ru/riyvwn.

[Zhu25] Zhu C., Zhu X., Qin T. Joint trajectory and incentive optimization for privacy-preserving UAV crowdsensing via multi-agent federated reinforcement learning // Internet of Things. 2025. Vol. 33. Pp. 101689. DOI: 10.1016/j.iot.2025.101689. EDN: CUXJER.

[Гур24] Гурчинский М. М., Тебуева Ф. Б. Обнаружение нарушителя агентами роевых робототехнических систем в условиях недетерминированной среды функционирования // СИИТ. 2024. Т. 6, № 3(18). С. 71–82. DOI: 10.54708/2658-5014-SIIT-2024-no3-p71. EDN: AUVYOX. [[Gurchinsky M. M., Tebueva F. B. Detection of an intruder by agents of swarm robotic systems in a non-deterministic operating environment // SIIT.2024. Vol. 6, No. 3(18). P. 71-82. (In Russian).]]

[Мир25] Миронов К. В. Transport-by-Throwing - робототехнический переброс: эксперименты и реализация // СИИТ. 2025. Т. 7, № 5(24). С. 40–56. DOI: 10.54708/2658-5014-SIIT-2025-no5-p40. EDN: UDALGS. [[Transport-by-Throwing - robotic transfer: experiments and implementation // SIIT.2025. Vol. 7, No. 5(24). P. 40-56. (In Russian).]]

[Мус24] Муслимов Т. З. Методы и алгоритмы группового управления беспилотными летательными аппаратами самолетного типа // СИИТ. 2024. Т. 6, № 1(16). С. 3–15. DOI: 10.54708/2658-5014-SIIT-2024-no1-p3. EDN: HOTUZU. [[Muslimov T. Z. Methods and algorithms for group control of unmanned aerial vehicles of the aircraft type // SIIT. 2024. Vol. 6, No. 1(16). P. 3-15. (In Russian).]]

[Пет21] Петренко В. И., Тебуева Ф. Б. и др. Алгоритм машинного обучения системы управления антропоморфными манипуляторами // СИИТ. 2021. Т. 3, № 2(6). С. 35–43. DOI: 10.54708/26585014_2021_32635. EDN: USZJSM. [[Petrenko V. I., Tebueva F. B., et al. Machine learning algorithm for the control system of anthropomorphic manipulators // SIIT. 2021. Vol. 3, No. 2(6). P. 35-43. (In Russian).]]

[При25] Приходько В. Е., Тепляшин и др. Практическая реализация коммуникационной системы мобильной группы на основе нейронных сетей // СИИТ. 2025. Т. 7, № 1(20). С. 96–104. DOI: 10.54708/2658-5014-SIIT-2025-no1-p96. EDN: UYDDVC. [[Prikhodko V. E., Teplyashin et al. Practical implementation of a mobile group communication system based on neural networks // SIIT. 2025. Vol. 7, No. 1(20). P. 96-104. (In Russian).]]

ОБ АВТОРАХ | ABOUT THE AUTHORS

ПЕТРЕНКО Вячеслав Иванович
Северо-Кавказский федеральный университет, Россия.
vipetrenko@ncfu.ru ORCID: 0000-0003-4293-7013.
Зав. кафедрой, канд. техн. наук, доцент. Исс. в обл. кибербезопасности роботизированных агентов.

ТЕБУЕВА Фариза Биляловна
Северо-Кавказский федеральный университет, Россия.
ftebueva@ncfu.ru ORCID: 0000-0002-7373-4692.
Проф. кафедры, д-р физ.-мат. наук, доцент. Исс. в обл. кибербезопасности роботизированных агентов.

СОБОЛЕВА Полина Александровна
Северо-Кавказский федеральный университет, Россия.
polina_soboleva_2019@mail.ru
Магистрант по напр. «Информационная безопасность».

PETRENKO Vyacheslav Ivanovich
North-Caucasus Federal University, Russia.
vipetrenko@ncfu.ru ORCID: 0000-0003-4293-7013.
Head of Dept. PhD in Engineering, Assoc. Prof. Research in the field of cybersecurity of robotic agents.

TEBUEVA Fariza Bilyalovna
North-Caucasus Federal University, Russia.
ftebueva@ncfu.ru ORCID: 0000-0002-7373-4692.
Prof. of the Dept., Dr. of Physical and Math. Sci., Assoc. Prof. Research in the field of cybersecurity of robotic agents.

SOBOLEVA Polina Alexandrovna
North-Caucasus Federal University, Russia.
polina_soboleva_2019@mail.ru
Master's student in Information Security.

МЕТАДАННЫЕ | METADATA

Заглавие: Метод доверенной оркестрации роботизированных агентов в децентрализованных средах на основе глубокого обучения с подкреплением.

Авторы: Петренко В. И., Тебуева Ф. Б., Соболева П. А.

Аннотация: В работе предложен новый метод Trust-MADDPG Orchestration (TMO), интегрирующий архитектуру централизованного обучения с децентрализованным исполнением на основе алгоритма глубокого обучения с подкреплением. Ключевым элементом метода является динамический механизм оценки доверия, который использует экспоненциальную функцию. Данный механизм обеспечивает быструю адаптацию системы к расхождениям между ожидаемым и фактическим вознаграждением. Разработанный метод был протестирован в реалистичной симуляционной среде MultiDroneSim на задаче совместного исследования территории с помехами и сбойными агентами. Эксперименты продемонстрировали превосходство TMO над базовыми методами: успешность выполнения миссии увеличена на 22.5%, а устойчивость к внедрению сбойных агентов повышена более чем в 4 раза (снижение эффективности всего на 5% по сравнению с 22% у базового метода). Эти результаты подтверждают, что интеграция механизма динамического доверия является ключевым фактором для обеспечения надежной оркестрации в децентрализованных мультиагентных системах.

Ключевые слова: Роботизированные агенты; децентрализованные системы; глубокое обучение с подкреплением; доверенная оркестрация; механизм доверия; координация; интеллектуальные системы.

Язык: Русский.

Статья поступила в редакцию 26 января 2026 г.

Title: A method for trusted orchestration of robotic agents in decentralized environments based on deep reinforcement learning.

Authors: Petrenko V. I., Tebueva F. B., Soboleva P. A.

Abstract: This paper proposes a new method, Trust-MADDPG Orchestration (TMO), which integrates a centralized learning architecture with decentralized execution based on a deep reinforcement learning algorithm. The key element of the method is a dynamic trust evaluation mechanism that uses an exponential function. This mechanism ensures rapid adaptation of the system to discrepancies between expected and actual rewards. The developed method was tested in the realistic MultiDroneSim simulation environment on the task of jointly exploring an area with interference and faulty agents. Experiments demonstrated the superiority of TMO over baseline methods: mission success increased by 22.5%, and resilience to faulty agent injection increased by more than four times (with a reduction in efficiency of only 5% compared to 22% for the baseline method). These results confirm that the integration of a dynamic trust mechanism is a key factor for ensuring reliable orchestration in decentralized multi-agent systems.

Key words: robotic agents; decentralized systems; deep reinforcement learning; trusted orchestration; trust mechanism; coordination; intelligent systems

Language: Russian.

The article was received by the editors on 26 January 2026.