УДК 004.852

Метод глубокого мультиагентного обучения с подкреплением для мобильных киберфизических систем с повышенными требованиями к функциональной безопасности

Петренко В. И.

Постановка задачи: *увеличение сложности задач, решаемых мобильными киберфизическими системами (МКФС), актуализирует вопросы применения такой технологии искусственного интеллекта, как глубокое мультиагентное обучение с подкреплением (ГМОП). Для применения методов ГМОП на практике необходимо повышение обеспечиваемой ими функциональной безопасности.* ***Целью работы*** *является повышение функциональной безопасности МКФС, обученных с помощью метода ГМОП на основе архитектуры «Эктор» – «Критик». Предлагается более тщательно выполнять обучение в состояниях, вызывающих некорректное поведение МКФС, путем повышения доли подобных состояний в обучающей выборке.* ***Используемые методы:*** *ГМОП осуществляется на основе метода MADDPG (multi-agent deep deterministic policy gradient). Для генерации обучающей выборки с необходимой плотностью вероятности на основе генератора случайных чисел с равномерной плотностью вероятности используется отдельная искусственная нейронная сеть (ИНС) «Тренер». ИНС «Тренер» также обучается в процессе ГМОП для повышения вероятности включения в обучающую выборку состояний, вызывающих некорректные поведение МКФС, и уменьшения вероятности включения в обучающую выборку ситуаций с корректным поведением МКФС.* ***Новизна:*** *элементами новизны представленного метода являются: 1) использование обучающей выборки с неравномерной плотностью вероятности состояний; 2) использование отдельной ИНС для генерации обучающей выборки с необходимой плотностью вероятности.* ***Результат:*** *использование предложенного метода позволило снизить по сравнению с аналогом вероятность возникновения опасных состояний в задаче кооперативной навигации с 19,1% до 0,02% при том же количестве шагов обучения.* ***Практическая значимость:*** *предложенный метод может быть использован для обучения или предобучения МКФС в симуляционных средах. Ожидается, что предложенный метод расширит применимость ГМОП в реальных МКФС.*

Ключевые слова: *глубокое мультиагентное обучение с подкреплением, искусственный интеллект, мобильные киберфизические системы, функциональная безопасность.*

Актуальность

Киберфизические системы (КФС) представляют собой системы, в которых интегрируются вычислительные, коммуникационные и физические процессы [1]. КФС находят применение в таких областях [2] как транспортировка, умные дома, робототехника, авиация, объекты инфраструктуры, медицина и др., а также выступают в качестве центральных компонентов критических информационных инфраструктур. Зачастую КФС включают в себя множество разнородных подсистем с высокой сложностью динамики и неоднородностью, поэтому решение оптимизационных задач в данной области с помощью тради-

Библиографическая ссылка на статью:
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

Reference for citation:
Petrenko V. I. Multi-agent Deep Reinforcement Learning Method for Mobile Cyber-Physical Systems with Increased Functional Safety Requirements. *Systems of Control, Communication and Security*, 2021, no. 3, pp. 179-206 (in Russian). DOI: 10.24412/2410-9916-2021-3-179-206

ционных алгоритмов является трудоёмким и дорогостоящим [3]. Перспективным направлением является использование такого метода искусственного интеллекта, как машинное обучение. В данной статье рассматривается разновидность методов машинного обучения – обучение с подкреплением (ОП, англ. reinforcement learning). Методы ОП нашли своё применение в КФС, используемых в умных фабриках [4], транспортировке [5], электрических сетях [6], коммуникационных сетях [7], военных операциях [8–10] и др.

Мобильные КФС (МКФС), такие как группа беспилотных летательных аппаратов (БПЛА) или мобильных роботов, могут рассматриваться как разновидность мультиагентных систем (МАС). Объектом данной статьи являются методы глубокого мультиагентного обучения с подкреплением (ГМОП), являющиеся разновидностью ОП для работы в МАС. Интерес к МАС обусловлен следующими причинами: применение МАС из более простых агентов вместо одного более сложного агента является экономически более эффективным [11]; децентрализованное решение задач с помощью МАС характеризуется более высокой эффективностью по сравнению с аналогичными централизованными методами [12]; за счет более высокой устойчивости функционирования МАС по сравнению с единичным агентом повышается вероятность выполнения целевой задачи. Важными вопросами исследования МАС являются вопросы управления поведением [13, 14], коллективного принятия решений [11], распределения ресурсов [15] и обеспечения безопасности [16]. Мощным и универсальным средством решения интеллектуальных задач является использование искусственных нейронных сетей (ИНС) и глубокого одноагентного обучения с подкреплением [17, 18].

Во многих случаях КФС являются критическими системами, т. е. обладающими повышенными требованиями к надежности и безопасности [19–22] в том числе к функциональной безопасности (ФБ) [23–25]. Область применения ИНС для управления техническими системами на данный момент слабо стандартизирована, поэтому специфический стандарт по ФБ для систем на основе ИНС отсутствует. Согласно [24], наиболее подходящим стандартом для систем на основе ИНС является стандарт ИСО 26262-1:2011 «Дорожные транспортные средства. Функциональная безопасность». В его российском аналоге [26] дано следующее определение: функциональная безопасность (англ. functional safety) – это отсутствие неоправданного риска вследствие опасностей, вызванных неправильным поведением электрических и/или электронных систем. Данное определение может быть применено к любой области применения ИНС для решения задач управления.

Обозначим поведение агентов МКФС, приводящее к неоправданным рискам при функционировании в штатных ситуациях как небезопасное с точки зрения ФБ, далее просто «небезопасное». Для описания нежелательного поведения агентов под управлением ИНС может также использоваться критерий «функциональность», который согласно ГОСТ Р ИСО/МЭК 25040-2014 «Информационные технологии (ИТ). Системная и программная инженерия. Требования и оценка качества систем и программного обеспечения (SQuaRE). Процесс оценки» «можно использовать, чтобы задать или оценить пригодность,

точность, интероперабельность, безопасность и нормосоответствие функции». Однако понятие ФБ по мнению автора является более подходящим. Фраза «повышение функциональности результатов ГМОП» может быть более вероятно неправильно интерпретирована, чем фраза «повышение функциональной безопасности результатов ГМОП».

С точки зрения ФБ применение ИНС и методов ГМОП для обеспечения функционирования МКФС обладает следующими особенностями:

1) решения на основе ИНС, в отличие от программируемых алгоритмов, устанавливающих связь между входными и выходными данными в виде спецификации, обладают слабой интерпретируемостью. В случае управления агентом МКФС на основе запрограммированного без ошибок алгоритма, поведение агента в опасных ситуациях при известных входных данных может быть предсказано, а устойчивость используемого решения оценена. В случае ИНС её спецификацией являются веса связей, поэтому поведение агента в опасной ситуации не может быть предсказано, а может быть только оценено путем непосредственного вычисления выходных сигналов ИНС. Вследствие возможности проблемы переобучения адекватное поведение агента при одних входных данных не гарантирует приемлемого решения при незначительном изменении этих входных данных. ФБ поведения агентов под управлением ИНС может быть оценена только статистически, на основе многократных физических или симуляционных испытаний;
2) классические методы ГМОП ориентированы на максимизацию среднего вознаграждения, получаемого агентами МКФС. Недопустимость опасных состояний учитывается лишь косвенно, за счет введения отрицательной награды (штрафа) за попадание агентов МКФС в эти состояния. Такой подход уменьшает, но не исключает вероятность попадания агентов в эти опасные состояния.

В совокупности данные особенности привели к появлению класса методов «безопасного» ОП. Обзор данной области исследований выполнен в работе [27]. Безопасное ОП может быть определено как процесс оптимизации политики принятия решений, который максимизирует среднее значение вознаграждения в задачах, в которых важно обеспечение разумной производительности системы и/или соблюдение ограничений безопасности во время процессов обучения и/или функционирования. Рассматриваемая в работе функциональная безопасность результатов ГМОП может рассматриваться как разновидность безопасного обучения, нацеленная на соблюдение ограничений безопасности в процессе функционирования обученной системы. Классификация методов безопасного ОП, предложенная в работе [27], представлена на рис. 1. В работе [27] приведено достаточно большое множество методов безопасного одноагентного обучения с подкреплением, в то время как безопасное ГМОП остаётся слабо исследованной областью [25].

В рамках ГМОП наибольшее распространение получили методы на основе модификации процесса исследования пространства состояний среды [28, 29]. Предложенный в работе [29] метод Shielding нацелен на «безопасное обуче-

ние», т. е. исключение возникновения опасных ситуаций на стадиях обучения и функционирования. Безопасность обучения является обязательной при проведении обучения в физической среде, при этом не является обязательной для обучения в симуляционной среде.

Рис. 1. Классификация методов безопасного обучения с подкреплением

Суть метода Shielding заключается в следующем:

1) разрабатывается спецификация безопасности, т. е. описание недопустимого поведения агентов, с помощью LTL – языка спецификации для критических систем [30, 31];
2) на основе спецификации безопасности формируется дополнительный блок защиты (англ. shield) между агентами МКФС и средой, осуществляющий коррекцию действий, которые собираются предпринять агенты, если эти действия могут привести к опасной ситуации (рис. 2).

Используемый в методе Shielding подход обладает следующими недостатками:

1) необходимо создание сложной спецификации безопасного поведения агентов. Для каждого сочетания среда/МКФС необходима разработка новой спецификации безопасности;
2) формирование блока защиты является трудоёмкой процедурой. При высокой сложности и динамике МКФС блок защиты обладает высокой сложностью, может ложно срабатывать при недостаточной детализации;

3) метод не позволяет агентам в процессе обучения попасть из околоопасного состояния в опасное за счет использования блока защиты, но при этом не стимулирует агентов избегать околоопасные состояния;
4) для функционирования метода требуется предсказательная модель среды. В случае симуляционной среды на этапе обучения дальнейшее состояние системы может быть предсказано путем проведения шага симуляции. Однако при дальнейшем функционировании в физической среде необходима точная математическая модель среды.

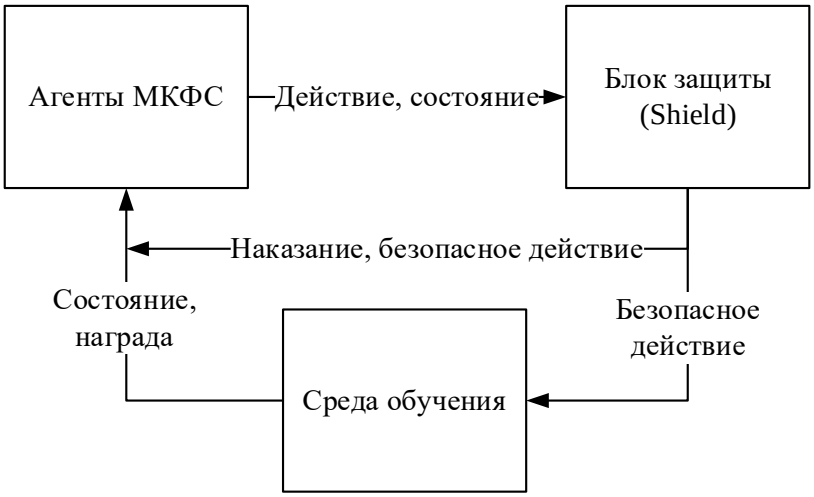

Рис. 2. Безопасное обучение на основе метода Shielding

Перспективным является использование для безопасного ГМОП другого подхода безопасного ОП – исследование пространства состояний, направляемое риском (рис. 1). Преимуществами данного подхода является исключение из процесса обучения внешних знаний, что делает его универсальным для решения различных задач. Подход не требует наличия математической модели среды и основывается исключительно на информации, полученной агентами в процессе взаимодействия со средой. Отсутствие аналитически сгенерированного блока защиты позволяет обученной системе адаптироваться к новым опасным факторам, возникающим в опасных средах. С точки зрения автономии обучающего алгоритма, такой подход является более перспективным для построения сильного интеллекта. Целью данной работы является повышение ФБ результатов ГМОП путем разработки соответствующего метода на основе исследования пространства состояний, направляемого риском.

Модель процесса ГМОП

Процесс ГМОП представляет собой взаимодействие глубокой мультиагентной задачи (ГМАЗ) T и метода ГМОП M (рис. 3). На рис. 3 прямоугольниками обозначены элементы данного процесса, преобразующие входные переменные в выходные, стрелками обозначены передаваемые между элементами переменные. Описание элементов и переменных излагается далее.

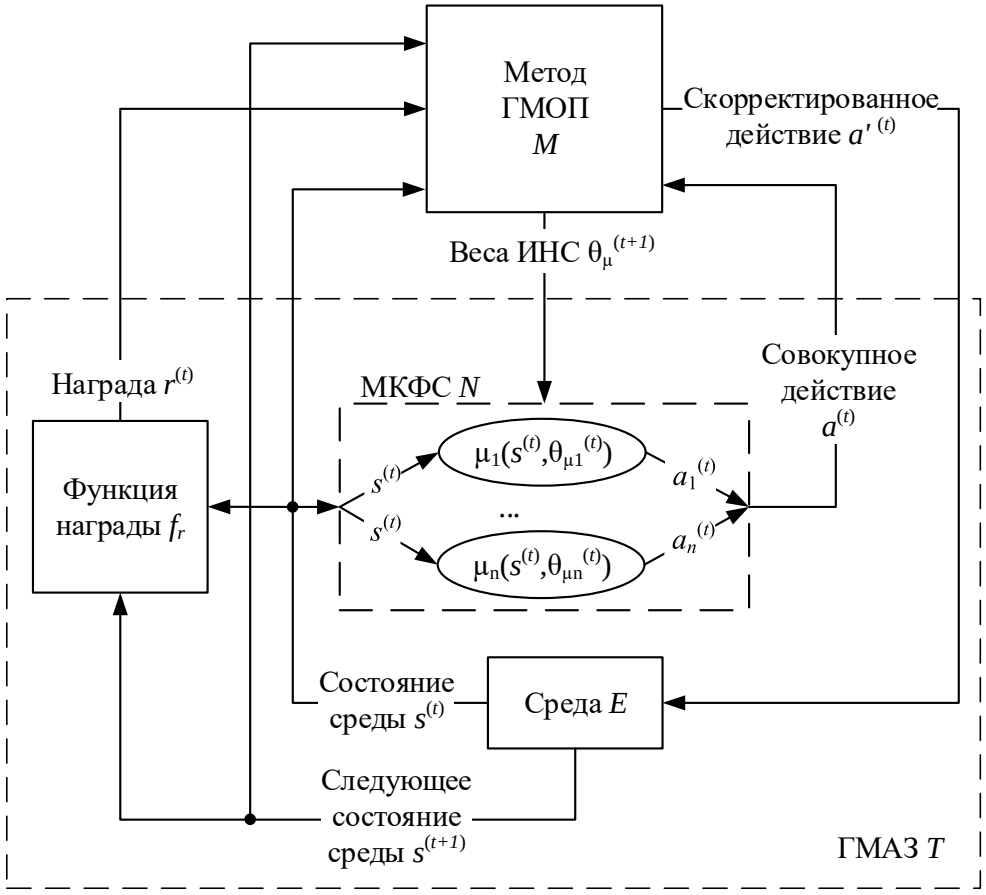

Рис. 3. Модель процесса ГМОП

В ГМАЗ T (нижняя часть на рис. 3) входят следующие элементы:

– МКФС N;
– среда E, в которой функционирует МКФС N;
– функция награды f_r, отражающая назначение МКФС N.

Математически ГМАЗ T может быть описана как мультиагентное расширение Марковского процесса принятия решений (МППР), описываемое кортежем [28]:

$$T = \left(E, S, N, A, f_s, f_r, \gamma\right), \tag{1}$$

где S – множество возможных значений состояния s среды E, в состояние s входят как переменные, описывающие непосредственно среду, так переменные физического состояния агентов МКФС N; A – множество возможных значений совокупного действия a всех агентов МКФС N; $f_s : S \times A \to S$ функция перехода, принимающая в качестве аргументов текущее состояние ГМАЗ $s^{(t)} \in S$ и совокупное действие $a^{(t)} \in A$ в момент времени t, возвращающая состояние среды $s^{(t+1)}$ в следующий момент времени $\left(t+1\right)$; $f_r : S \times A \times S \to \mathbb{R}^n$ – векторная функция награды (англ. reward), принимающая в качестве аргументов $s^{(t)}$, $a^{(t)}$ и $s^{(t+1)}$, возвращающая кортеж вознаграждений $r^{(t)} = (r_i^{(t)} \mid i = \overline{1,n})$, отражающий полезность действий агентов с точки зрения достижения цели

ГМАЗ в момент времени t; $\gamma \in [0,1]$ – фактор дисконтирования, отражающий важность получения текущей награды $r^{(t)}$ по сравнению с будущими.

В статье значение верхнего индекса, взятое в скобки, используется для обозначения момента времени, например: $s^{(t)}$, $a^{(t)}$ и т. п. Верхний индекс без скобок используется для указания степени, например: $\mathbb{R}^n$, γ^t.

Термин «глубокая» в аббревиатуре ГМАЗ означает, что для управления каждым агентом МКФС N используется отдельная политика принятия решения, аппроксимируемая ИНС. Обозначим такую ИНС, управляющую i-м агентом МКФС N, как ИНС μ_i «Эктор» (рис. 3). ИНС μ_i «Эктор» параметризуется весами связей θ_{μ_i} и осуществляет преобразование наблюдаемого i-м агентом состояния среды $s^{(t)}$ в предпринимаемое действие $a_i^{(t)}$:

$$a_i^{(t)} = \mu_i\left(s^{(t)}, \theta_{\mu_i}\right). \quad (2)$$

Совокупность действий $a_i^{(t)}$ агентов МКФС N эквивалентна совокупному действию агентов $a^{(t)}$:

$$a^{(t)} = \left(a_i^{(t)} | i = \overline{1,n}\right), \quad (3)$$

где n – количество агентов в МКФС N.

Обозначим сумму наград, получаемых агентами МКФС N в течение некоторого интервала времени как вознаграждение R (англ. return):

$$R = \sum_{t=0}^{t_f} \gamma^t r^{(t)}, \quad (4)$$

где t – момент времени; t_f – длительность эпизода для эпизодических ГМАЗ (ограниченных по времени) и ширина временного окна для периодических ГМАЗ (не ограниченных по времени).

Назначением применения метода ГМОП M к ГМАЗ является оптимизация весов θ_{μ_i} соответствующих ИНС μ_i «Эктор» с целью максимизации по критерию результативности q_r, равного среднему значению вознаграждения R МКФС N:

$$q_r = \bar{R}, \quad (5)$$

где $\bar{R}$ – среднее значение вознаграждения R.

В качестве входных данных (рис. 3) метод ГМОП M использует переменные $s^{(t)}, a^{(t)}, s^{(t+1)}$, выходными данными являются веса $\theta_\mu = \{\theta_{\mu_i} | i = \overline{1,n}\}$ группы $\mu = \{\mu_i | i = \overline{1,n}\}$ ИНС «Эктор» и скорректированное действие $a'^{(t)}$ (описывается далее).

В данной работе за основу для улучшений взят метод ГМОП MADDPG. Обобщённый алгоритм метода MADDPG [32], приведен на рис. 4. Блоки, подвергшиеся изменению в рамках предлагаемого метода, выделены на рис. 4 цветом.

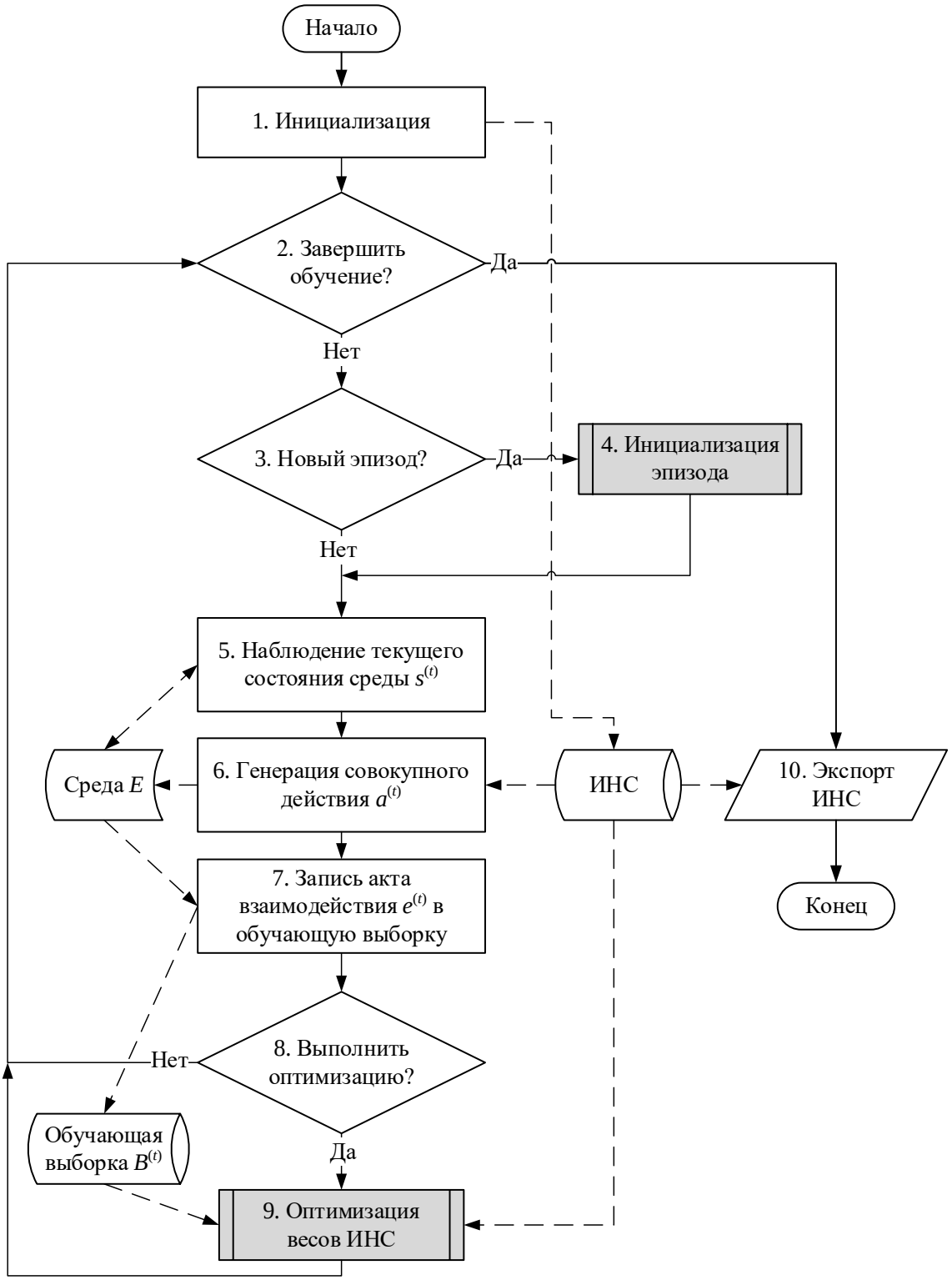

Рис. 4. Обобщённый алгоритм предлагаемого метода

Обобщённый алгоритм метода ГМОП MADDPG включает в себя следующую последовательность действий.

Шаг 1. Инициализируются необходимые переменные, генерируются группа ИНС $\mu = \{\mu_i \mid i = \overline{1,n}\}$ «Эктор», ИНС Q «Критик», ИНС τ «Тренер». Веса $\theta_\mu^{(0)}, \theta_Q^{(0)}, \theta_\tau^{(0)}$ в момент времени $t = 0$ перечисленных ИНС инициализируются

случайными значениями. Генерируются целевые ИНС μ' «Эктор» и Q' «Критик». Под «целевыми» понимаются те экземпляры ИНС, которые будут в дальнейшем использоваться на стадии функционирования МКФС N. Значения весов ИНС $\theta_{\mu}^{(0)}$, $\theta_{Q}^{(0)}$ присваиваются весам соответствующих целевых ИНС:

$$\theta_{\mu'}^{(0)} \leftarrow \theta_{\mu}^{(0)}, \tag{6}$$

$$\theta_{Q'}^{(0)} \leftarrow \theta_{Q}^{(0)}, \tag{7}$$

Создаётся обучающая выборка, представляющая пустое множество:

$$B^{(t)} \leftarrow \varnothing. \tag{8}$$

Шаг 2. Если не выполнен критерий завершенности обучения, выполняется переход на шаг 3. Если выполнен, то выполняется переход на шаг 10. В качестве критерия завершенности обучения обычно используется превышение заданного количества шагов обучения или достижение порогового значения по критерию q_r.

Шаг 3. Если необходимо инициализировать новый эпизод, выполняется переход на шаг 4. Если необходимости нет, выполняется переход на шаг 5. Под эпизодом понимается последовательность шагов обучения, между которыми сохраняется состояние среды E. В начале эпизода выполняется его инициализация начальным состоянием $s^{(0)}$. Эпизод длится пока не будет выполнено условие завершения эпизода, например превышение заданного количества шагов или достижение агентами МКФС N цели.

Шаг 4. Инициализация эпизода. Осуществляется выборка начального состояния $s^{(0)}$ среды E. В методе MADDPG выборка начального состояния $s^{(0)}$ среды E осуществляется с равномерным распределением из множества возможных состояний среды S:

$$s^{(0)} \underset{\rho_c}{\Leftarrow} S, \tag{9}$$

где $x \underset{\rho_c}{\Leftarrow} X$ означает выборку случайной величины x из множества X с равномерным распределением, имеющим постоянную плотность вероятности ρ_c (англ. constant).

Шаг 5. Агентами МКФС N выполняется наблюдение текущего состояния $s^{(t)}$ среды E.

Шаг 6. Выполняется генерация совокупного действия $a^{(t)}$ агентов МКФС N на основе политик принятия решения $\mu_i^{(t)}$:

$$a^{(t)} = \left\{ \mu_i^{(t)} \left(s^{(t)}, \theta_{\mu_i}^{(t)} \right) | i = \overline{1,n} \right\}. \tag{10}$$

Вычисляется скорректированное действие $a'^{(t)}$ согласно формуле:

$$a'^{(t)} = a^{(t)} + \Delta a, \Delta a \underset{\rho_n}{\Leftarrow} A, \tag{11}$$

где Δa – случайное отклонение, подчиняющееся нормальному распределению с плотностью вероятности ρ_n.

Случайное отклонение Δa вводится с целью приобретения агентами нового опыта в ходе случайных отклонений от действия $a^{(t)}$, генерируемого группой ИНС μ «Эктор».

Шаг 7. На основе скорректированного действия $a'^{(t)}$ моделируется следующее состояние среды $s^{(t+1)}$ и определяется награда $r^{(t)}$. Акт опыта взаимодействия $e^{(t)}$ (англ. experience) в виде кортежа:

$$e^{(t)} = \left(s^{(t)},\ a'^{(t)},\ s^{(t+1)}, r^{(t)}\right), \tag{12}$$

записывается в обучающую выборку $B^{(t)}$:

$$B^{(t)} \leftarrow \left\{B^{(t)}; e^{(t)}\right\}, \tag{13}$$

где символ «$\leftarrow$» означает присваивание левой части значения правой.

Шаг 8. При необходимости выполнения оптимизации весов ИНС выполняется переход на шаг 9, в противном случае выполняется переход на шаг 2. Периодичность выполнения оптимизации весов ИНС определяется разработчиком.

Шаг 9. Выполняется оптимизация ИНС согласно алгоритму, излагаемому далее.

Шаг 10. Выполняется сохранение обученных ИНС.

Оптимизация весов ИНС на шаге 9 обобщённого алгоритма метода ГМОП MADDPG включает в себя следующую последовательность действий (рис. 5).

Шаг 1. Из обучающей выборки $B^{(t)}$ извлекается минибатч $b^{(t)}$ размерности d:

$$b^{(t)} \sim_d B^{(t)}, \tag{14}$$

где «$\sim_d$» обозначает операцию извлечения случайной выборки размерности d.

Шаг 2. Для каждого элемента $b_j^{(t)}$ минибатча $b^{(t)}$ выполняются следующие действия.

Шаг 2.1. Вычисляется уточненное значение y_j для Q-функции на основе уравнения Беллмана с помощью целевой ИНС Q' «Критик» и группы ИНС μ′ «Эктор»:

$$\begin{aligned} y_j &= r_j^{(t)} + \gamma Q'\left(s_j^{(t+1)}, a^{(t+1)}\right) = \\ &= r_j^{(t)} + \gamma Q'\left(s_j^{(t+1)}, \mu_i'^{(t)}\left(s_j^{(t+1)}\right)\right). \end{aligned} \tag{15}$$

Шаг 3. Вычисляется функция общей ошибки Q-функции для всего минибатча $b^{(t)}$:

$$\mathcal{L} = \frac{1}{d}\sum_j \left(y_j - Q(s_j^{(t)}, a_j^{(t)}\right)^2. \tag{16}$$

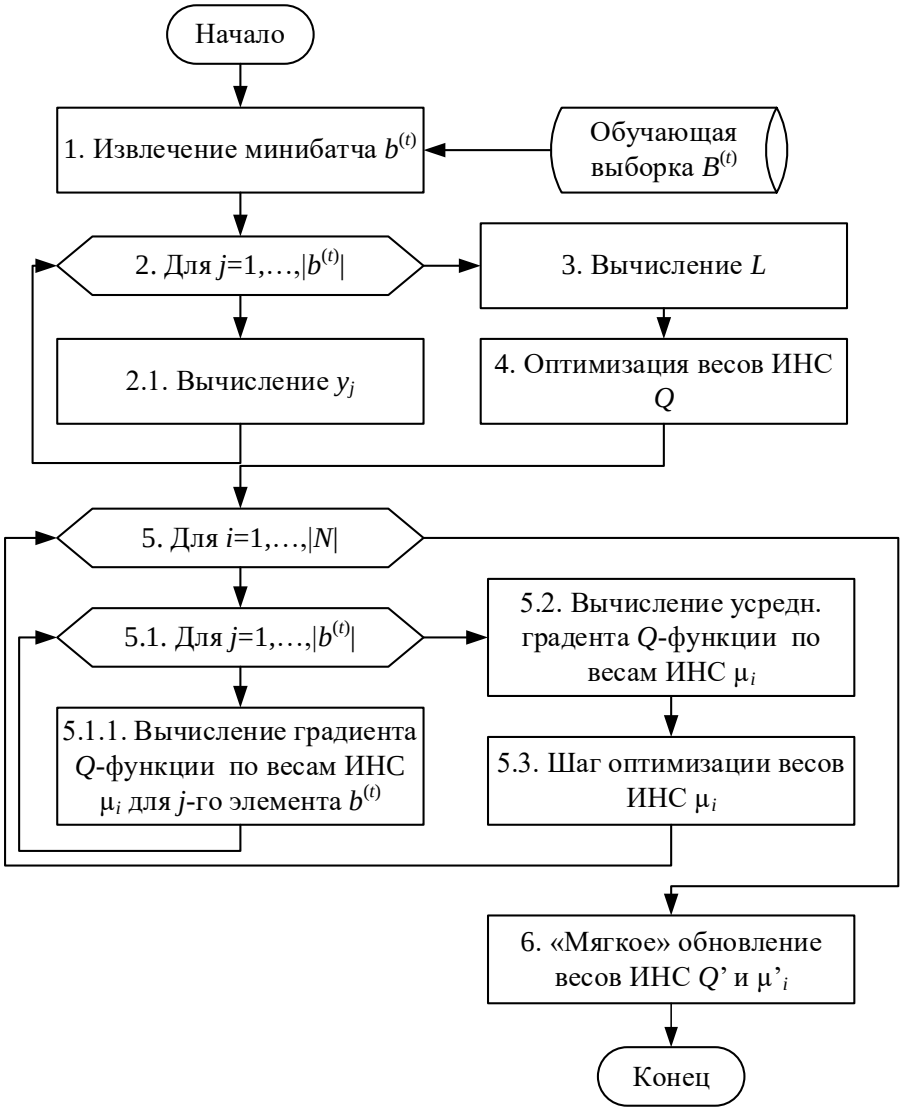

Рис. 5. Алгоритм оптимизации весов ИНС

Шаг 4. Выполняется оптимизация весов ИНС Q «Критик» с целью минимизации функции ошибки $\mathcal{L}$ методом градиентного спуска.

Шаг 5. Для каждого i -го агента выполняются следующие действия.

Шаг 5.1. Для каждого j -го элемента $b_j^{(t)}$ минибатча $b^{(t)}$ выполняются следующие действия.

Шаг 5.1.1. Вычисляется градиент Q -функции по весам $\theta_{\mu_i}^{(t)}$ ИНС μ_i «Эктор»:

$$\nabla_{\theta_{\mu_i}} Q_j = \nabla_{a\,j,i} Q\left(o_j, a_j | \theta_Q^{(t)}\right) \nabla_{\theta_{\mu_i}} \mu_i\left(o_j | \theta_{\mu_i}^{(t)}\right), \tag{17}$$

Шаг 5.2. Вычисляется усредненный по минибатчу градиент Q -функции по весам $\theta_{\mu_i}^{(t)}$ ИНС «Эктор» μ_i:

$$\nabla_{\theta_{\mu_i}} Q = \frac{1}{d} \sum_j \nabla_{\theta_{\mu_i}} Q_j, \tag{18}$$

Шаг 5.3. Выполняется оптимизация вектора параметров θ_{μ_i} с помощью метода градиентного подъёма по формуле:

$$\theta_{\mu_i}^{(t+1)} = \theta_{\mu_i}^{(t)} + \alpha \nabla_{\theta_{\mu_i}} Q, \quad (19)$$

где α – шаг оптимизации.

Шаг 6. При необходимости выполняется «мягкое» обновление весов целевых ИНС μ' «Эктор» и Q' «Критик»:

$$\theta_{\mu'} \leftarrow \rho\theta_{\mu'} + (1-\rho)\theta_{\mu}, \quad (20)$$

$$\theta_{Q'} \leftarrow \rho\theta_{Q'} + (1-\rho)\theta_{Q}, \quad (21)$$

где $\rho \in [0;1]$ – коэффициент обновления.

Постановка задачи

Согласно [33] в данной работе рассматриваются МКФС, которые могут быть классифицированы следующим образом:

- по свойствам агентов как МАС с искусственными виртуальными мобильными и интеллектуальными агентами;
- по виду взаимодействия между агентами как кооперативные МАС с простым сотрудничеством;
- по свойствам организации как МАС в виде гетерархических самоорганизующихся сообществ.

В работе используется понятие опасного состояния – состояния объекта, в котором возникает недопустимый риск причинения вреда людям, или окружающей среде, или существенных материальных потерь, или других неприемлемых последствий [34]. Примерами таких опасных ситуаций является крушение агентов группы БПЛА, столкновение беспилотного транспорта, повреждение или выход из строя объекта управления и др. Обозначим непрерывное множество таких объективно опасных состояний среды E как S_d (англ. danger). Обозначим множество состояний s среды E, из которых осуществляется переход в опасные состояния $\{s \mid s \in S_d\}$ при управлении МКФС N согласно текущей политики принятия решений μ вследствие недостаточного уровня ФБ, как S_w (англ. wrong). Обозначим совокупность S_u множества опасных состояний S_d и множества ошибочных состояний S_w как множество потенциально опасных состояний (англ. potential):

$$S_p = S_d \cup S_w. \quad (22)$$

На основе введённых определений в качестве критерия q_s ФБ (англ. safety) может использоваться вероятность p_d наступления в какой-либо момент времени t опасного состояния $s^{(t)} \in S_d$:

$$q_s = p_d\left(s^{(t)} \in S_d\right). \quad (23)$$

Тогда вербальная постановка научной задачи может быть сформулирована следующим образом: необходимо разработать метод M_1 ГМОП для повышения ФБ по критерию q_s в диапазоне значений входных и выходных переменных (S, A) МКФС N, за счет варьирования значений весов θ_{μ} группы ис-

пользуемых ИНС μ «Эктор» при ограничении на минимальное значение критерия результативности q_r и функционировании в среде E.

Формальная постановка научной задачи имеет следующий вид. Необходимо найти метод M_1 такой, что:

$$M_1 : N, S, A, \theta, E, Q \to \Delta q_i \geq 0, q_i \in Q, \quad (24)$$

где $Q = \{q_r, q_s\}$ – множество рассматриваемых критериев оценки функционирования МКФС N; $\Delta q_i = q_i^{\text{п}} - q_i^{\text{д}}$, где индекс «д» значит «до использования метода» M_1, индекс «п» – «после использования метода» M_1;
при ограничениях на варьируемые переменные: $\theta_\mu \in \theta_{\text{доп}}$, где $\theta_{\text{доп}}$ – множество допустимых значений θ_μ; ограничения на неварьируемые переменные: $s \in S \subseteq S_{\text{доп}}$, где $S_{\text{доп}}$ – множество допустимых значений S; $a \in A \subseteq A_{\text{доп}}$, где $A_{\text{доп}}$ – множество допустимых значений A.

Научная идея

Метод ГМОП MADDPG обеспечивает недопустимо низкую ФБ в состояниях задачи, близких к опасным, приводящей к переходу задачи в опасные состояния. Как следует из экспериментальной оценки, приведенной в разделе «Результаты», после завершения обучения с помощью метода MADDPG опасные ситуации возникают в 19% экспериментальных запусков. Причиной данной проблемы являются следующие факторы:

1) генерация обучающей выборки B и извлечение из неё актов взаимодействия МКФС N со средой E выполняется случайным образом с равномерным распределением. Т. к. в большинстве случаев опасные состояния составляют малую долю от общего количества возможных состояний среды E, доля актов взаимодействия МКФС N со средой E в потенциально опасных состояниях мала. Поэтому акты обучения МКФС N поведению в потенциально опасных состояниях происходят достаточно редко;
2) в качестве критерия завершенности обучения используется условное постоянство критерия q_r. Вследствие первого фактора значение критерия q_r перестаёт возрастать после обучения агентов эффективному поведению в большинстве неопасных состояний s, несмотря на неэффективное поведение в потенциально опасных состояниях. Данный критерий скорее отражает невозможность достижения лучших результатов с помощью используемого метода, чем достижение цели обучения.

Научная идея предлагаемого решения заключается в следующих пунктах:

1) предлагается повысить долю актов обучения поведению в потенциально опасных состояниях в общей совокупности актов обучения. Предлагается использовать обучающую выборку с неравномерным распределением состояний среды. Плотность вероятности состояний в обучающей выборке предлагается построить на основе опасности состоя-

ний. В качестве косвенной меры опасности состояния предлагается использовать величину Q-функции данного состояния;

2) для генерации случайных состояний с заданной плотностью вероятности предлагается использовать дополнительно введенную ИНС τ «Тренер».

Научная гипотеза исследования заключается в том, что предложенные решения позволят повысить ФБ обученной МКФС по сравнению с существующим методом ГМОП MADDPG.

Предлагаемый метод

Предлагаемый метод вносит изменения в реализацию шагов 4 и 9 обобщённого алгоритма ГМОП метода MADDPG (рис. 4).

В предлагаемом методе на шаге 4 выборка начального состояния $s^{(0)}$ осуществляется с некоторой неравномерной (англ. uneven) плотностью вероятности ρ_u:

$$s^{(0)} \underset{\rho_u}{\Leftarrow} S. \tag{25}$$

На плотность вероятности ρ_u для состояния $s \in S$ накладывается следующее ограничение. Плотность вероятности $\rho_u(s)$ должна быть тем выше, чем ниже $Q(s,a)$, и наоборот. Существующие методы генерации случайных чисел предназначены в основном для генерации чисел с равномерным распределением. На основе генератора случайных чисел с равномерным распределением может быть получено необходимое заданное аналитически распределение на основе использования обратной функции распределения. Сложность заключается в том, что аналитические формы функции $Q(s,a)$, и необходимого распределения вероятности неизвестны. Функция $Q(s,a)$ аппроксимирована ИНС Q «Критик» и может быть вычислена для любых входных значений, однако анализ её аналитической формы затруднителен вследствие сложности структуры ИНС. Для решения данной проблемы предлагается осуществлять генерацию начального состояния $s^{(0)}$ с помощью дополнительной ИНС τ «Тренер», общей для всех агентов (рис. 6).

На вход ИНС τ подаётся входное состояние s_i (англ. input), сгенерированное с помощью генератора случайных чисел согласно равномерному распределению с постоянной плотностью вероятности ρ_c. На выходе ИНС τ формируется выходное состояние s_o (англ. output), подчиняющееся распределению с плотностью вероятности ρ_u:

$$s_o = \tau(s_i, \theta_\tau),\ s_o \sim \rho_u, s_i \sim \rho_c, \tag{26}$$

где θ_τ – веса ИНС τ «Тренер»; выражение $x \sim \rho$ означает, что случайная величина x подчиняется распределению с плотностью вероятности ρ.

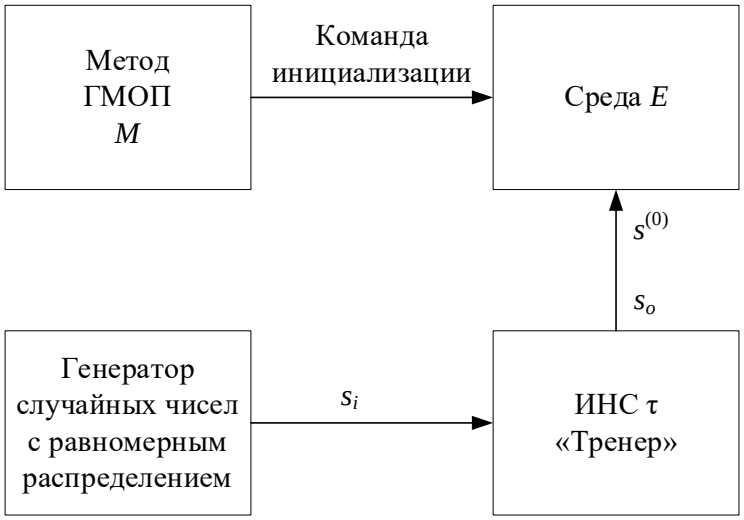

Рис. 6. Схема генерации начального состояния среды E

Сгенерированное состояние s_o используется для инициализации начального состояния $s^{(0)}$ среды E:

$$s^{(0)} \leftarrow s_o. \tag{27}$$

Плотность вероятности ρ_u распределения выходного состояния s_o должна соответствовать условию:

$$\forall s, s' \in S \quad \rho_u(s) > \rho_u(s') \mid Q(s, \mu(s)) < Q(s', \mu(s')). \tag{28}$$

Преобразование τ из s_i в s_o предлагается построить на основе изменения плотности вероятности. Иллюстрация для случая с одномерными состояниями s_i и s_o приведена на рис. 7.

На рис. 7 по горизонтальной оси отложены значения входного состояния s_i, имеющего постоянную плотность вероятности ρ_c. График плотности вероятности $\rho_c(s_i)$ изображен штрих-пунктирной линией вдоль оси s_i.

Для плотности вероятности ρ_u справедливо следующее ограничение:

$$\int \rho_c ds_i = \int \rho_u ds_o. \tag{29}$$

В таком случае справедливы следующие уравнения:

$$\int \rho_c ds_i = \int \rho_u d\left(\tau(s_i)\right) = \int \rho_u \frac{d\tau}{ds_i} ds_i, \tag{30}$$

$$\rho_c = \rho_u \frac{d\tau}{ds_i}, \tag{31}$$

$$\frac{d\tau}{ds_i} = \frac{\rho_c}{\rho_u}. \tag{32}$$

Таким образом, получаемое значение ρ_u в результате преобразования τ зависит от угла наклона τ. В точке P с единичной производной наблюдается равенство плотности вероятности (рис. 7):

$$\rho_c(s_{i,P}) = \rho_u(s_{o,P}), \tag{33}$$

где $s_{i,P}, s_{o,P}$ – проекция точки P на оси s_i, s_o.

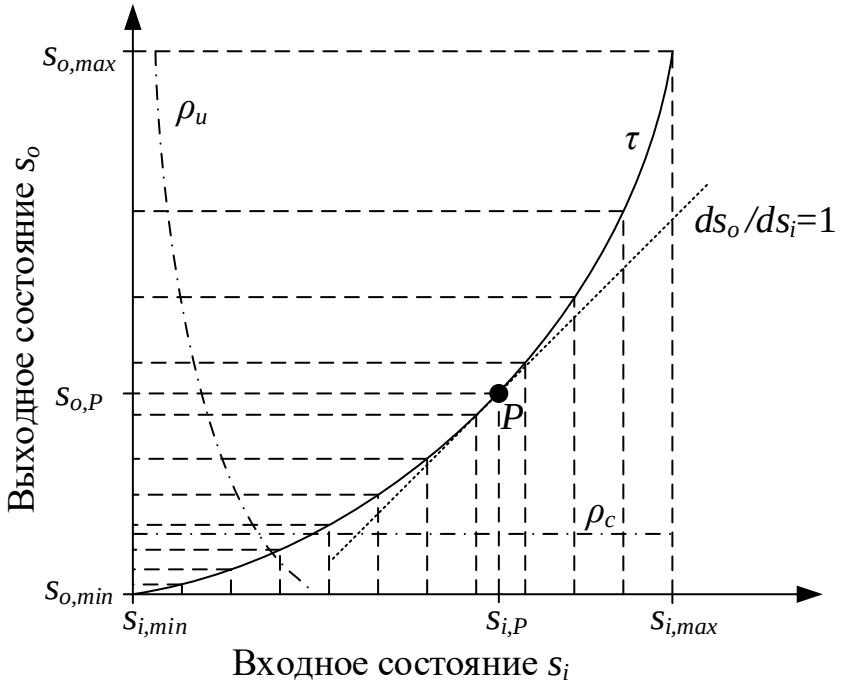

Рис. 7. Преобразование плотности состояний

На начальном участке графика на рис. 7 до точки равной плотности P, величина плотности вероятности ρ_u превышает значение равномерной плотности вероятности ρ_c. И наоборот, на участке графика на рис. 7 после точки P, величина плотности вероятности ρ_u меньше значения равномерной плотности вероятности ρ_c. Качественный вид графика ρ_u приведен на рис. 7 штрих-пунктирной линией вдоль оси s_o.

При переходе от одномерных состояний s_i и s_o к многомерным, отношение плотностей вероятности (32) может быть найдено как отношение многомерных элементарных параллелепипедов, т. е. как якобиан (определитель матрицы Якоби):

$$J_\tau(s_i) = \frac{Ds_o}{Ds_i} = \frac{\rho_c}{\rho_u}. \tag{34}$$

Так как $\rho_c = const$, справедливо соотношение

$$\rho_u \sim \frac{1}{J_\tau(s_i)}. \tag{35}$$

Для удовлетворения условия (28) наложим следующее ограничение: пусть ИНС τ «Тренер» должна обеспечивать пропорциональность:

$$J_\tau(s_i) \sim Q(s_o, \mu(s_o)). \tag{36}$$

Пусть:

$$Q(s, \mu(s)) < Q(s', \mu(s')), \tag{37}$$

тогда:

$$J_\tau(s_i)\big|_{\tau(s_i)=s} \left\langle J_\tau(s_i')\big|_{\tau(s_i')=s'} \Rightarrow \rho_u(s) \right\rangle \rho_u(s'). \tag{38}$$

И наоборот, если

$$Q(s,\mu(s)) > Q(s',\mu(s')), \tag{39}$$

тогда:

$$J_\tau(s_i)\big|_{\tau(s_i)=s} > J_\tau(s_i')\big|_{\tau(s_i')=s'} \Rightarrow \rho_u(s) < \rho_u(s'). \tag{40}$$

Таким образом, ограничение (36) является необходимым и достаточным для выполнения условия (28).

Для выполнения условия (28) предлагается выполнять оптимизацию весов ИНС τ «Тренер» следующим образом. Пусть инициализация ИНС τ «Тренер» выполняется случайными весами $\theta_\tau^{(0)}$. Для их последующей оптимизации в качестве функции ошибки $\mathcal{L}_\tau$ для минибатча b_τ входных состояний s_i может использоваться среднеквадратичное отклонение регрессии нормализованного значения якобиана $\hat{J}_\tau$ и нормализованного $\hat{Q}$ значения Q-функции от линейной регрессии:

$$b_\tau = \left\{ s_{i,j} | j = \overline{1, d_\tau} \right\}, \tag{41}$$

$$\hat{J}_{\tau,j} = \frac{J_{\tau,j}(s_{i,j}) - \min\limits_j J_{\tau,j}(s_{i,j})}{\max\limits_j J_{\tau,j}(s_{i,j}) - \min\limits_j J_{\tau,j}(s_{i,j})}, \tag{42}$$

$$s_{o,j} = \tau(s_{i,j}), \tag{43}$$

$$\hat{Q}_j = \frac{Q(s_{o,j},\mu(s_{o,j})) - \min\limits_j Q(s_{o,j},\mu(s_{o,j}))}{\max\limits_j Q(s_{o,j},\mu(s_{o,j})) - \min\limits_j Q(s_{o,j},\mu(s_{o,j}))}, \tag{44}$$

$$\mathcal{L}_\tau = \frac{1}{d_\tau} \sum_j \left(\hat{J}_{\tau,j} - \hat{Q}_j \right)^2, \tag{45}$$

где d_τ – размерность минибатча b_τ; $\hat{J}_{\tau,j}$ – нормализованное значение якобиана J_τ для j-го элемента минибатча b_τ; $\hat{Q}_j$ – нормализованной значение Q-функции для j-го элемента минибатча b_τ.

Для модификации шага 9 обобщённого алгоритма метода ГМОП MADDPG (рис. 4), предлагается добавить в алгоритм оптимизации весов ИНС (рис. 5) следующие шаги (рис. 8).

Шаг 7. На основе множества S возможных состояний среды E генерируется минибатч $b_\tau^{(t)}$ размерности d_τ входных состояний s_i для ИНС τ «Тренер»:

$$b_\tau^{(t)} = \left\{ s_{i,j} \mid j = \overline{1, d_\tau}, s_{i,j} \underset{\rho_c}{\Leftarrow} S \right\}, \tag{46}$$

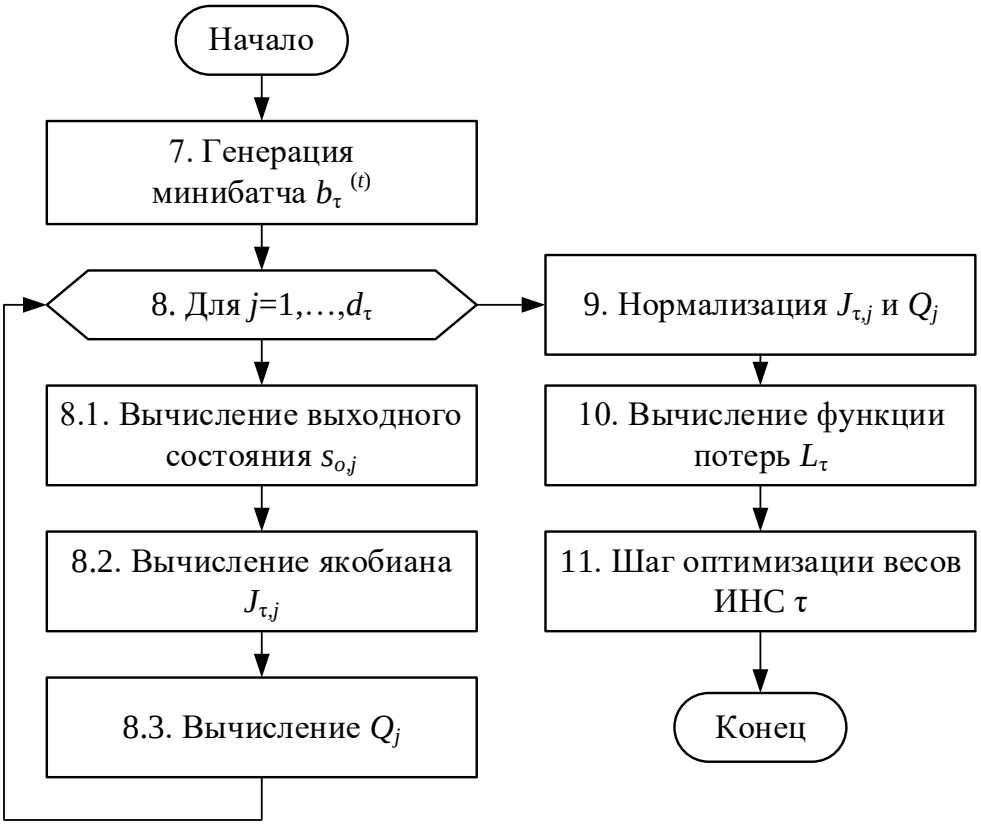

Рис. 8. Дополнительные шаги оптимизации весов ИНС

Шаг 8. Для каждого j-го элемента $s_{i,j}$ минибатча $b_\tau^{(t)}$ выполняются следующие действия.

Шаг 8.1. Вычисляется выходное состояние $s_{o,j}$ с помощью ИНС τ «Тренер»:

$$s_{o,j} = \tau^{(t)}\left(s_{i,j}\right). \tag{47}$$

Шаг 8.2. Для ИНС τ «Тренер» вычисляется якобиан (определитель матрицы Якоби) $J_\tau\left(s_i\right)$ частных производных элементов вектора s_o от вектора s_i:

$$J_{\tau,i}(s_i) = \begin{vmatrix} \frac{\partial s_{o,1}}{\partial s_{i,1}}(s_i) & \frac{\partial s_{o,1}}{\partial s_{i,2}}(s_i) & \cdots & \frac{\partial s_{o,1}}{\partial s_{i,|S|}}(s_i) \\ \frac{\partial s_{o,2}}{\partial s_{i,1}}(s_i) & \frac{\partial s_{o,2}}{\partial s_{i,2}}(s_i) & \cdots & \frac{\partial s_{o,2}}{\partial s_{i,|S|}}(s_i) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial s_{o,|S|}}{\partial s_{i,1}}(s_i) & \frac{\partial s_{o,|S|}}{\partial s_{i,2}}(s_i) & \cdots & \frac{\partial s_{o,|S|}}{\partial s_{i,|S|}}(s_i) \end{vmatrix}. \tag{48}$$

Шаг 8.3. Вычисляется значение Q-функции $s_j^{(t)}$ для выходного состояния $s_{o,j}$ при функционировании согласно набору политик $\mu^{(t)}$:

$$Q_j = Q\left(s_{o,j}, \mu^{(t)}\left(s_{o,j}\right)\right). \tag{49}$$

Шаг 9. Вычисляются нормализованные значения $\hat{J}_{\tau,j}$ и $\hat{Q}_j$ для переменных $J_{\tau,j}$ и Q_j, соответственно:

$$\hat{J}_{\tau,j} = \frac{J_{\tau,j} - \min\limits_j J_{\tau,j}}{\max\limits_j J_{\tau,j} - \min\limits_j J_{\tau,j}}, \quad (50)$$

$$\hat{Q}_j = \frac{Q_j - \min\limits_j Q_j}{\max\limits_j Q_j - \min\limits_j Q_j}. \quad (51)$$

Шаг 10. Вычисляется суммарная функция потерь:

$$\mathcal{L}_\tau = \frac{1}{d_\tau}\sum_j \left(\hat{J}_{\tau,j} - \hat{Q}_j\right)^2. \quad (52)$$

Шаг 11. Выполняется шаг оптимизации весов θ_τ ИНС τ «тренер» с целью минимизации функции потерь $\mathcal{L}_\tau$.

Результаты

Для проверки гипотезы о повышении ФБ за счет предложенных решений было выполнено экспериментальное исследование на примере обучения МКФС из n агентов решению задачи кооперативной навигации [32]. Задача кооперативной навигации (рис. 9) заключается в следующем. На двумерной плоскости размещаются n агентов и n целевых позиций. На рис. 9 агенты обозначены синими кружками, а целевые позиции черными крестиками. Задача считается решенной, когда все целевые позиции заняты агентами.

Экспериментальное исследование включало в себя два этапа. На первом этапе было выполнено обучение МКФС с помощью метода MADDPG и предлагаемого метода. На втором этапе была выполнена оценка функциональной безопасности обученных МКФС по критерию q_s.

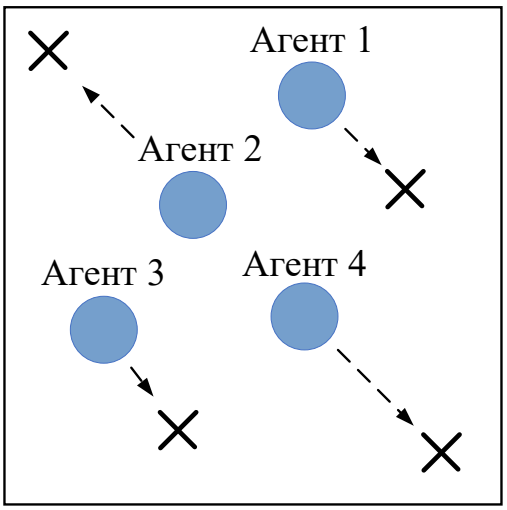

Рис. 9. Задача кооперативной навигации

На первом этапе при обучении использовалась следующая функция награды f_r:

$$f_r = -\sum_{i=1,n}\left(\min_{j\in 1,n}\left|c_{t,i} - c_{a,j}\right|\right) - 0{,}5\sum_{i=1,n}\sum_{i=1,n}\left(\left|c_{a,i} - c_{a,j}\right| < d_{min}\right), \quad (53)$$

где $c_{a,i}$ – координаты i-го агента; $c_{t,i}$ – координаты j-го агента; d_{min} – минимальное допустимое расстояние между агентами.

Первое слагаемое функции награды f_r представляет собой штраф, пропорциональный сумме расстояний от целевых позиций до ближайшего агента. Второе слагаемое представляет собой штраф за столкновение агентов между собой, т. е. наступление опасных состояний. Столкновение считается произошедшим, если расстояние между агентами меньше, чем некоторое пороговое значение d_{min}.

При обучении с помощью метода MADDPG и предлагаемого метода использовались значения гиперпараметров, приведенные в таблице 1.

На рис. 10 представлена зависимости критерия q_r от количества эпизодов обучения, построенные на основе экспериментальных данных первого этапа.

Таблица 1 – Значения гиперпараметров процесса обучения

Гиперпараметр	Значение
Размерность d минибатча b, ед.	1024
Коэффициент дисконтирования γ, ед.	0,95
Среднеквадратичное отклонение Δa_σ, ед.	0,2
Шаг оптимизации α, ед.	0,01
Размерность d_τ минибатча b_τ, ед.	1024
Минимальное допустимое расстояние между агентами d_{min}, ед.	0,1
Коэффициент обновления целевых ИНС ρ, ед.	0,001
Количество эпизодов обучения, ед.	25 000

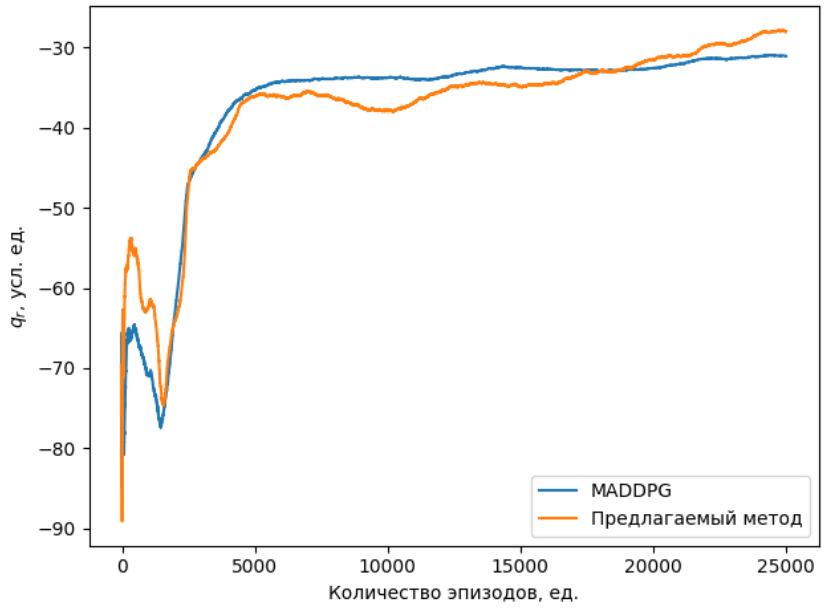

Рис. 10. Зависимость q_r от количества эпизодов обучения

Как следует из рис. 10, применение предлагаемого метода обеспечивает более высокое значение критерия q_r при одинаковом с методом-аналогом количестве эпизодов обучения.

На втором этапе экспериментального исследования была выполнена оценка значения критерия q_s, равного вероятности наступления опасных состояний, при выполнении обучения с помощью предлагаемого метода и метода MADDPG. Для каждого метода была выполнена симуляция 10 000 эпизодов при управлении МКФС с помощью группы ИНС μ «Эктор», обученных на первом этапе экспериментального исследования. Каждый эпизод был ограничен длительностью в 25 шагов. Если до наступления ограничения по количеству шагов наступало опасное состояние, симуляция эпизода завершалась. После завершения каждого эпизода инкрементировались счетчики эпизодов, завершившихся без/с наступлением опасного состояния. Путем деления значения счетчика эпизодов, завершившихся с наступлением опасного состояния, на общее количество эпизодов было вычислена точечная оценка критерия q_s. При доверительной вероятности 0,95 точечная оценка значения критерия q_s для МКФС, обученной с помощью метода MADDPG, составила 19,1% с доверительным интервалом [18,1%; 20,1%]. Для МКФС, обученной с помощью предлагаемого метода, точечная оценка значения критерия q_s составила 0,02% с доверительным интервалом [0,00%; 0,06%].

Выводы

Применение методов ГМОП в МКФС требует повышения их ФБ. Для решения данной проблемы в работе выдвинута гипотеза о возможности повышения ФБ обученной МКФС за счет повышения доли потенциально опасных состояний в обучающей выборке при осуществлении процесса ГМОП. Для проверки гипотезы был разработан метод повышения доли потенциально опасных состояний в обучающей выборке с помощью использования дополнительной ИНС «Тренер».

Согласно проведенному экспериментальному исследованию, предложенный метод позволил снизить вероятность наступления опасных состояний q_s с 19,1% до 0,02% при сохранении того же значения и даже небольшом повышении среднего значения вознаграждения q_r. Полученные результаты подтверждают выдвинутую гипотезу и обосновывают актуальность применения предложенных решений на практике.

Литература

1. Cyber-Physical Systems // The Ptolemy Project [Электронный ресурс]. 05.21.2021. – URL: https://ptolemy.berkeley.edu/projects/cps/ (дата обращения 05.21.2021).

2. Wang L., Törngren M., Onori M. Current Status and Advancement of Cyber-Physical Systems in Manufacturing // Journal of Manufacturing Systems. 2015. Vol. 37. No. 2. P. 517-527. doi: 10.1016/j.jmsy.2015.04.008.

3. Liu X., Xu H., Liao W., Yu W. Reinforcement Learning for Cyber-Physical Systems // 2019 IEEE International Conference on Industrial Internet (ICII). 2019. P. 318-327. doi: 10.1109/ICII.2019.00063.

4. Jiang Y. Fan J, Chai T., Lewis F. L. Dual-Rate Operational Optimal Control for Flotation Industrial Process with Unknown Operational Model // IEEE Transactions on Industrial Electronics. 2019. Vol. 66, No. 6. P. 4587-4599. doi: 10.1109/TIE.2018.2856198.

5. Ferdowsi A. Challita U., Saad W., Mandayam N. B. Robust Deep Reinforcement Learning for Security and Safety in Autonomous Vehicle Systems // 2018 21st International Conference on Intelligent Transportation Systems (ITSC). 2018. P. 307-312. doi: 10.1109/ITSC.2018.8569635.

6. Glavic M., Fonteneau R., Ernst D. Reinforcement Learning for Electric Power System Decision and Control: Past Considerations and Perspectives // IFAC-PapersOnLine. 2017. Vol. 50. No. 1. P. 6918-6927. doi: 10.1016/j.ifacol.2017.08.1217.

7 Luong N. C., Hoang D. T., Gong S., Niyato D., Wang P., Liang Y., Kim D. I. Applications of Deep Reinforcement Learning in Communications and Networking: A Survey // IEEE Communications Surveys and Tutorials. 2019. Vol. 21, No. 4. P. 3133-3174. doi: 10.1109/COMST.2019.2916583.

8. Васильченко А. С., Иванов М. С., Колмыков Г. Н. Формирование маршрутов полета беспилотных летательных аппаратов с учетом местоположения средств противовоздушной обороны и радиоэлектронного подавления // Системы управления, связи и безопасности. 2019. № 4. С. 403-420. doi: 10.24411/2410-9916-2019-10416

9. Васильченко А. С., Иванов М. С., Малышев В. А. Формирование полетных зон беспилотных летательных аппаратов по степени устойчивости управления ими в условиях применения средств противовоздушной обороны и радиоэлектронного подавления // Системы управления, связи и безопасности. 2019. № 4. С. 262-279. doi: 10.24411/2410-9916-2019-10410.

10. Юдинцев Б. С. Синтез нейросетевой системы планирования траекторий для группы мобильных роботов // Системы управления, связи и безопасности. 2019. № 4. С. 163-186. doi: 10.24411/2410-9916-2019-10406.

11. Petrenko V. I. Tebueva F. B., Ryabtsev S. S., Gurchinsky M. M, Struchkov I. V. Consensus Achievement Method for a Robotic Swarm About the Most Frequently Feature of an Environment // IOP Conference Series: Materials Science and Engineering. 2020. Vol. 919. doi: 10.1088/1757-899X/919/4/042025.

12. Kovács G., Yussupova N., Rizvanov D. Resource Management Simulation Using Multi-Agent Approach and Semantic Constraints // Pollack Periodica. 2017. Vol. 12. No. 1. P. 45-58. doi: 10.1556/606.2017.12.1.4.

13. Пшихопов В. Х., Медведев М. Ю. Групповое управление движением мобильных роботов в неопределенной среде с использованием неустойчивых режимов // Труды СПИИРАН. 2018. № 5 (60). С. 39–63. doi: 10.15622/sp.60.2.

14. Тугенгольд А. К., Лукьянов Е. А. Интеллектуальные функции и управление автономными технологическими мехатронными объектами. – Ростов-на-Дону: Донской государственный технический университет, 2013. – 203 с.

15. Даринцев О. В., Мигранов А. Б. Распределенная система управления группами мобильных роботов // Вестник Уфимского государственного авиационного технического университета. 2017. Т. 21. № 2 (76). С. 88-94.

16. Петренко В. И., Тебуева Ф.Б., Гурчинский М. М., Рябцев С. С. Анализ технологий обеспечения информационной безопасности мультиагентных робототехнических систем с роевым интеллектом // Наука и бизнес: пути развития. 2020. № 4 (106). С. 96-99.

17. Munasypov R. A., Masalimov K. A. Neural Network Models for Diagnostics of Complex Technical Objects State by Example of Electrochemical Treatment Process // Proceedings - 2017 2nd International Ural Conference on Measurements, UralCon 2017. 2017. P. 156–160. doi: 10.1109/URALCON.2017.8120703.

18. Mironov K. V., Pongratz M. U. Applying Neural Networks for Prediction of Flying Objects Trajectory // Вестник УГАТУ. 2013. Т. 17. № 6(59). С. 33-37.

19. Yusupova N., Rizvanov D., Andrushko D. Cyber-Physical Systems and Reliability Issues // Proceedings of the 8th Scientific Conference on Information Technologies for Intelligent Decision Making Support (ITIDS 2020). 2020. Vol. 174. P. 133-137. doi: 10.2991/aisr.k.201029.026.

20. Fabarisov T. Yusupova N., Ding K., Morozov A., Janschek K. Model-Based Stochastic Error Propagation Analysis for Cyber-physical Systems // Acta Polytechnica Hungarica. 2020. Vol. 17. No 8. P. 15-28. doi: 10.12700/APH.17.8.2020.8.2.

21. Valiev E. Yusupova N., Morozov A., Janschek K., Beyer M. Evaluation of the Impact of Random Computing Hardware Faults on the Performance of Convolutional Neural Networks // Proceedings of the 8th Scientific Conference on Information Technologies for Intelligent Decision Making Support (ITIDS 2020). 2020. Vol. 174. P. 307-312. doi: 10.2991/aisr.k.201029.058.

22. Beyer M. Morozov A., Ding K., Ding S., Janschek K. Quantification of the Impact of Random Hardware Faults on Safety-Critical AI Applications: CNN-Based Traffic Sign Recognition Case Study // 2019 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW). 2019. P. 118-119. doi: 10.1109/ISSREW.2019.00058.

23. Salay R., Queiroz R., Czarnecki K. An Analysis of ISO 26262: Using Machine Learning Safely in Automotive Software // Safety of the Intended Functionality (SAE). 2020. P. 13-25. doi: 10.4271/9780768002683.

24. Henriksson J., Borg M., Englund C. Automotive Safety and Machine Learning: Initial Results from a Study on How to Adapt the ISO 26262 Safety Standard // 2018 IEEE/ACM 1st International Workshop on Software Engineering for AI in Autonomous Systems (SEFAIAS). 2018. P. 47-49.

25. Martin H., Tschabuschnig K., Bridal O., Watzenig D. Functional Safety of Automated Driving Systems: Does ISO 26262 Meet the Challenges? // Automated Driving. 2017. P. 387-416. doi: 10.1007/978-3-319-31895-0_16.

26. ГОСТ Р ИСО 26262-1-2014 Дорожные транспортные средства. Функциональная безопасность. – М.: Стандартинформ, 2020. – 36 с.

27. García J., Fernández F. A Comprehensive Survey on Safe Reinforcement Learning // Journal of Machine Learning Research. 2015. Vol. 16. P. 1437-1480.

28. Zhang W., Bastani O., Kumar V. MAMPS: Safe Multi-Agent Reinforcement Learning via Model Predictive Shielding // arXiv.org [Электронный ресурс]. 21.05.2021. – URL: https://arxiv.org/pdf/1910.12639.pdf (дата обращения 21.05.2021).

29. Elsayed-Aly I., Bharadwaj S., Amato C., Ehlers R., Topcu U., Feng L. Safe Multi-Agent Reinforcement Learning via Shielding // arXiv.org [Электронный ресурс]. 21.05.2021. – URL: https://arxiv.org/pdf/2101.11196.pdf (дата обращения 21.05.2021).

30. Roy S., Das S. K. Principles of Cyber-Physical Systems: An Interdisciplinary Approach. – Cambridge: Cambridge University Press, 2020. – 400 p. doi: 10.1017/9781107588981.

31. Baier C., Katoen J.-P. Principles Of Model Checking. – MIT Press, 2008. 994 p.

32. Lowe R., Wu Y., Tamar A., Harb J., Abbeel P., Mordatch I. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments // 31st Conference on Neural Information Processing Systems (NIPS 2017). 2017. P. 6382-6393.

33. Тарасов В. Б. От многоагентных систем к интеллектуальным организациям: философия, психология, информатика. – М.: УРСС, 2002. – 352 с.

34. ГОСТ 27.002-2015 Надежность в технике (ССНТ). Термины и определения. – М.: Стандартинформ, 2016. – 30 с.

References

1. Cyber-physical systems. *The Ptolemy Project*, 05.21.2021. Available at: https://ptolemy.berkeley.edu/projects/cps/ (accessed 05 May 2021).

2. Wang L., Törngren M., Onori M. Current Status and Advancement of Cyber-Physical Systems in Manufacturing. *Journal of Manufacturing Systems*, 2015, vol. 37, no. 2, pp. 517-527. doi: 10.1016/j.jmsy.2015.04.008.

3. Liu X., Xu H., Liao W., Yu W. Reinforcement Learning for Cyber-Physical Systems. *2019 IEEE International Conference on Industrial Internet (ICII)*, Orlando, 2019, pp. 318-327. doi: 10.1109/ICII.2019.00063.

4. Jiang Y. Fan J, Chai T., Lewis F.L. Dual-Rate Operational Optimal Control for Flotation Industrial Process with Unknown Operational Model. *IEEE Transactions on Industrial Electronics*, 2019, vol. 66, no. 6, pp. 4587-4599. doi: 10.1109/TIE.2018.2856198.

5. Ferdowsi A. Challita U., Saad W., Mandayam N. B. Robust Deep Reinforcement Learning for Security and Safety in Autonomous Vehicle Systems. *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, Maui, HI, USA, 2018, pp. 307-312. doi: 10.1109/ITSC.2018.8569635.

6. Glavic M., Fonteneau R., Ernst D. Reinforcement Learning for Electric Power System Decision and Control: Past Considerations and Perspectives. *IFAC-PapersOnLine*, 2017, vol. 50, no. 1, pp. 6918-6927. doi: 10.1016/j.ifacol.2017.08.1217.

7. Luong N. C., Hoang D. T., Gong S., Niyato D., Wang P., Liang Y., Kim D. I. Applications of Deep Reinforcement Learning in Communications and Networking: A Survey. *IEEE Communications Surveys and Tutorials*, 2019, vol. 21, no. 4, pp. 3133-3174. doi: 10.1109/COMST.2019.2916583.

8. Vasilchenko A. S., Ivanov M. S., Kolmykov G. N. Unmanned Aerial Vehicles Flight Routes Formation, Taking into Account the Location of Air Defense and Electronic Warfare Means. *Systems of Control, Communication and Security*, 2019, no. 4. pp. 403-420. doi: 10.24411/2410-9916-2019-10416. (in Russian).

9. Vasilchenko A. S., Ivanov M. S., Malyshev V. A. Unmanned Aerial Vehicles Flight Zones Formation, Based on Their Control Stability Degree in Air Defense and Electronic Warfare Conditions. *Systems of Control, Communication and Security*, 2019, no. 4, pp. 262-279. doi: 10.24411/2410-9916-2019-10410. (in Russian).

10. Yudintsev B. S. A Path Planning System Synthesis for A Group of Mobile Robots Based on Neural Network. *Systems of Control, Communication and Security*, 2019, no. 4, pp. 163-186. doi: 10.24411/2410-9916-2019-10406. (in Russian).

11. Petrenko V. I. Tebueva F. B., Ryabtsev S. S., Gurchinsky M. M, Struchkov I. V. Consensus Achievement Method for A Robotic Swarm About the Most Frequently Feature of An Environment. *IOP Conference Series: Materials Science and Engineering*, 2020, vol. 919. doi: 10.1088/1757-899X/919/4/042025.

12. Kovács G., Yussupova N., Rizvanov D. Resource Management Simulation Using Multi-Agent Approach and Semantic Constraints. *Pollack Periodica*, 2017, vol. 12, no. 1, pp. 45-58. doi: 10.1556/606.2017.12.1.4.

13. Pshikhopov V. Kh., Medvedev M. Yu. Group Control of Autonomous Robots Motion in Uncertain Environment Via Unstable Modes. *SPIIRAS Proceedings*, 2018, no. 5 (60), pp. 39–63. doi: 10.15622/sp.60.2. (in Russian).

14. Tugengold A. K., Lukyanov E. A. *Intellektual'nye funkcii i upravlenie avtonomnymi tekhnologicheskimi mekhatronnymi ob"ektami* [Intelligent Functions and Control of Autonomous Technological Mechatronic Objects]. Rostov-on-Don, Don State Technical University, 2013. 203 p. (in Russian).

15. Darintsev O. V., Migranov A. B. Distributed Control System for Group of Mobile Robots. *Vestnik USATU*, 2017, vol. 21. no. 2 (76), pp. 88-94 (in Russian).

16. Petrenko V. I., Tebueva F. B., Gurchinsky M. M., Ryabtsev S. S. Analysis of Information Security Technologies for Multi-Agent Robotic Systems with Swarm Intelligence. *Science and business: ways of development*, 2020, no. 4 (106), pp. 96–99. (in Russian).

17. Munasypov R. A., Masalimov K. A. Neural Network Models for Diagnostics of Complex Technical Objects State by Example of Electrochemical Treatment Process. *Proceedings - 2017 2nd International Ural Conference on Measurements (UralCon)*, 2017, pp. 156–160. doi: 10.1109/URALCON.2017.8120703.

18. Mironov K. V., Pongratz M. U. Applying Neural Networks for Prediction of Flying Objects Trajectory. *Vestnik of the Ufa State Aviation Technical University*, 2013, vol. 17, no 6 (59), pp. 33-37.

19. Yusupova N., Rizvanov D., Andrushko D. Cyber-Physical Systems and Reliability Issues. *Proceedings of the 8th Scientific Conference on Information Technologies for Intelligent Decision Making Support (ITIDS)*, 2020, vol. 174, pp. 133-137. doi: 10.2991/aisr.k.201029.026.

20. Fabarisov T. Yusupova N., Ding K., Morozov A., Janschek K. Model-Based Stochastic Error Propagation Analysis for Cyber-Physical Systems. *Acta Polytechnica Hungarica*, 2020, vol. 17, no 8, pp. 15-28. doi: 10.12700/APH.17.8.2020.8.2.

21. Valiev E. Yusupova N., Morozov A., Janschek K., Beyer M. Evaluation of the Impact of Random Computing Hardware Faults on the Performance of Convolutional Neural Networks. *Proceedings of the 8th Scientific Conference on Information Technologies for Intelligent Decision Making Support (ITIDS)*, 2020, vol. 174, pp. 307-312. doi: 10.2991/aisr.k.201029.058.

22. Beyer M. Morozov A., Ding K., Ding S., Janschek K. Quantification of the Impact of Random Hardware Faults on Safety-Critical AI Applications: CNN-Based Traffic Sign Recognition Case Study. *2019 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)*, 2019, pp. 118-119. doi: 10.1109/ISSREW.2019.00058.

23. Salay R., Queiroz R., Czarnecki K. An Analysis of ISO 26262: Using Machine Learning Safely in Automotive Software. *Safety of the Intended Functionality (SAE)*, 2020, pp. 13-25. doi: 10.4271/9780768002683.

24. Henriksson J., Borg M., Englund C. Automotive Safety and Machine Learning: Initial Results from a Study on How to Adapt the ISO 26262 Safety Standard. *2018 IEEE/ACM 1st International Workshop on Software Engineering for AI in Autonomous Systems (SEFAIAS)*, 2018, pp. 47-49.

25. Martin H., Tschabuschnig K., Bridal O., Watzenig D. Functional Safety of Automated Driving Systems: Does ISO 26262 Meet the Challenges? *Automated Driving*, 2017, pp. 387-416. doi: 10.1007/978-3-319-31895-0_16.

26. State Standard ISO 26262-1-2014. Road vehicles. Functional safety. Part 1: Vocabulary. Moscow, Standartov Publ., 2020. 36 p. (in Russian).

27. García J., Fernández F. A Comprehensive Survey on Safe Reinforcement Learning. *Journal of Machine Learning Research*, 2015, vol. 16, pp. 1437-1480.

28. Zhang W., Bastani O., Kumar V. MAMPS: Safe Multi-Agent Reinforcement Learning Via Model Predictive Shielding. *arXiv.org*. Available at: https://arxiv.org/pdf/1910.12639.pdf (accessed 21 May 2021).

29. Elsayed-Aly I., Bharadwaj S., Amato C., Ehlers R., Topcu U., Feng L. Safe Multi-Agent Reinforcement Learning via Shielding. *arXiv.org*. Available at: https://arxiv.org/pdf/2101.11196.pdf (accessed 21 May 2021).

30. Roy S., Das S. K. *Principles of Cyber-Physical Systems: An Interdisciplinary Approach*. Cambridge, Cambridge University Press, 2020. 400 p. doi: 10.1017/9781107588981.

31. Baier C., Katoen J.-P. *Principles of Model Checking*. MIT Press, 2008. 994 p.

32. Lowe R., Wu Y., Tamar A., Harb J., Abbeel P., Mordatch I. Multi-Agent Actor-Critic For Mixed Cooperative-Competitive Environments. *31st Conference on Neural Information Processing Systems (NIPS)*, 2017, pp. 6382-6393.

33. Tarasov V. B. *Ot mnogoagentnykh sistem k intellektual'nym organizatsiiam: filosofiia, psikhologiia, informatika* [From multi-agent systems to intelligent organizations: philosophy, psychology, computer science]. Moscow, URSS Publ., 2002. 352 p. (in Russian).

34. State Standard 27.002-2015. Dependability in technics. Terms and definitions. Moscow, Standartov Publ., 2016. 30 p. (in Russian).

Статья поступила 21 мая 2021 г.

Информация об авторе

Петренко Вячеслав Иванович – кандидат технических наук, доцент. Заведующий кафедрой организации и технологии защиты информации. Северо-Кавказский федеральный университет. Область научных интересов: системы защиты информации, защита персональных данных, арифметические операции в конечных полях, синтез дискретных последовательностей, системы связи, методы искусственного интеллекта, мультиагентные системы, глубокое обучение с подкреплением. E–mail: vip.petrenko@gmail.com

Адрес: Россия, 355017, г. Ставрополь, ул. Пушкина, д. 1

Multi-agent Deep Reinforcement Learning Method for Mobile Cyber-Physical Systems with Increased Functional Safety Requirements

V. I. Petrenko

Purpose. *Increasing the complexity of tasks solved by mobile cyber-physical systems (MCPS), actualizes the application of such artificial intelligence technology as multi-agent deep reinforcement learning (MDRL). For the application of MDRL methods in practice, it is necessary to increase the functional safety provided by them. The aim of the work is to increase the functional safety of MCPS trained using the MDRL method based on the actor-critic architecture. It is proposed to perform training more thoroughly in states that cause the incorrect behavior of the MCPS, by increasing the fraction of such states in the replay buffer.* ***Methods.*** *MDRL is based on the MADDPG (multi-agent deep deterministic policy gradient) method. To generate a replay buffer with the required probability density based on a random number generator with a uniform probability density, a separate artificial neural network (ANN) "trainer" is used. ANN trainer is also trained in the MDRL process to increase the probability of including in the replay buffer of states that cause the incorrect behavior of the MCPS, and to reduce the probability of including situations with the correct behavior of the MCPS in the replay buffer.* ***Novelty.*** *The elements of novelty of the presented method are: 1) the use of a replay buffer with an uneven probability density of states; 2) the use of a separate ANN to generate a replay buffer with the required probability density.* ***Results.*** *The use of the proposed method made it possible to reduce, in comparison with the analogue, the probability of the occurrence of dangerous states in the problem of cooperative navigation from 19.1% to 0.02% with the same number of training steps.* ***Practical relevance.*** *The proposed method can be used for training or pre-training of MCPS in simulation environments. The proposed method is expected to expand the applicability of MDRL in real MCPS.*

Keywords: *multi-agent deep reinforcement learning, artificial intelligence, mobile cyber-physical systems, functional safety.*

Information about Author

Vyacheslav Ivanovich Petrenko – Ph.D. of Engineering Sciences, Associate Professor. Head of the Department of Organization and Technology of Information Security. North-Caucasian Federal University. Field of research: information security systems, personal data protection, arithmetic operations in finite fields, synthesis of discrete sequences, communication systems, artificial intelligence, multi-agent systems, deep reinforcement learning. E–mail: vip.petrenko@gmail.com

Address: Russia, 355017, Stavropol, Pushkina street 1.