

Научная статья
УДК 004.89
MSC 94-10
doi: 10.17586/2713-1874-2024-4-50-59

ОБЪЯСНИМОСТЬ ПОВЕДЕНИЯ АГЕНТОВ В СИСТЕМАХ ПОДДЕРЖКИ ПРИНЯТИЯ КЛИНИЧЕСКИХ РЕШЕНИЙ

*Артем Олегович Исаков¹, Наталия Федоровна Гусарова²,
Дмитрий Александрович Добренко³, Александр Андреевич Голубев⁴*

^{1,2,3,4}Университет ИТМО, Санкт-Петербург, Россия
¹aoisakov@itmo.ru, <https://orcid.org/0000-0002-2938-0575>
²nfngusarova@itmo.ru, <http://orcid.org/0000-0002-1361-6037>
³enotpalaskun@gmail.com, <https://orcid.org/0009-0006-1485-1166>
⁴9459539@gmail.com, <https://orcid.org/0000-0001-7417-6947>
Язык статьи – русский

Аннотация: В статье рассматривается задача улучшения объяснимости поведения агентов в системах поддержки клинических решений с использованием многоагентного обучения с подкреплением, учитывая поведенческие особенности отдельных агентов. Особое внимание уделяется применению алгоритма проксимальной оптимизации политики (PPO), который используется для моделирования взаимодействия агентов искусственного и естественного интеллекта. Кроме того, рассматривается важность учета поведенческих особенностей пациентов, что достигается через использование разработанного фреймворка «убеждения-желания-намерения» (BDI). Фреймворк позволяет моделировать поведение пациентов более точно, учитывая их индивидуальные предпочтения и потребности, что способствует улучшению качества медицинского обслуживания. Для повышения объяснимости поведения агентов и обеспечения прозрачности принимаемых решений предложено использование большой языковой модели (LLM). Для оценки качества фреймворка как средства обеспечения требований ХАИ (объяснимого искусственного интеллекта) разработан набор критериев, продемонстрирована его репрезентативность. Оценка проводилась путем сравнения объяснений, сформированных четырьмя популярными языковыми моделями по семи запускам игрового процесса с различными исходными данными. Предложенный подход к обеспечению объяснимости в многоагентных симуляциях открывает возможности для дальнейшего развития и внедрения таких систем в клиническую практику.

Ключевые слова: многоагентные системы, обучение с подкреплением, объяснимость, принятие решений, системы реального времени, составление расписаний

Ссылка для цитирования: Исаков А. О., Гусарова Н. Ф., Добренко Д. А., Голубев А. А. Объяснимость поведения агентов в системах поддержки принятия клинических решений // Экономика. Право. Инновации. 2024. № 4. С. 50–59. <http://dx.doi.org/10.17586/2713-1874-2024-4-50-59>.

EXPLAINABILITY OF AGENT BEHAVIOR IN CLINICAL DECISION SUPPORT SYSTEMS

Artem O. Isakov¹, Natalia F. Gusarova², Dmitry A. Dobrenko³, Aleksandr A. Golubev⁴

^{1,2,3,4}ITMO University, Saint Petersburg, Russia
¹aoisakov@itmo.ru, <https://orcid.org/0000-0002-2938-0575>
²nfngusarova@itmo.ru, <http://orcid.org/0000-0002-1361-6037>
³enotpalaskun@gmail.com, <https://orcid.org/0009-0006-1485-1166>
⁴9459539@gmail.com, <https://orcid.org/0000-0001-7417-6947>
Article in Russian

Abstract: The article considers the task of improving the explainability of agent behavior in clinical decision support systems using multi-agent reinforcement learning, taking into account the behavioral characteristics of individual agents. Special attention is paid to the application of the proximal policy optimization (PPO) algorithm, which is used to simulate the interaction of artificial and natural intelligence agents. In addition, the importance of taking into account the behavioral characteristics of patients is considered, which is achieved through the use of the developed framework «beliefs-desires-intentions» (BDI). The framework allows you to model patient behavior more accurately, taking into account their individual preferences and needs, which contributes to improving the quality of medical care. To increase the explainability

of agent behavior and ensure transparency of decisions, the use of a large language model (LLM) is proposed. To assess the quality of the framework as a means of ensuring the requirements of XAI (explicable artificial intelligence), a set of criteria has been developed and its representativeness has been demonstrated. The assessment was carried out by comparing the explanations generated by four popular language models for seven gameplay launches with different initial data. The proposed approach to ensuring explainability in multi-agent simulations opens up opportunities for further development and implementation of such systems in clinical practice.

Keywords: decision making, explainability, multi-agent systems, real time systems, reinforcement learning, scheduling

For citation: Isakov A. O., Gusarova N. F., Dobrenko D. A., Golubev A. A. Explainability of Agent Behavior in Clinical Decision Support Systems. *Ekonomika. Pravo. Innovacii*. 2024. No. 4. pp. 50–59. (In Russ.). <http://dx.doi.org/10.17586/2713-1874-2024-4-50-59>.

Введение. Постановка задачи. Системы поддержки принятия клинических решений (СПКР) широко внедряются в практику работы учреждений здравоохранения. Спектр их задач включает помощь врачам в постановке диагноза, организацию общения в системе «врач – пациент», контроль выполнения клинических протоколов, а также комплексное планирование деятельности медицинской организации на всех уровнях. Особенно остро вопрос разработки СПКР стоит для клиник высокотехнологичной медицинской помощи, где сосредоточены высококвалифицированные кадры и дорогостоящее оборудование. Здесь большое значение приобретает распределение свободных операционных в целях максимально эффективного использования ресурсов клиник при увеличении входящего потока пациентов.

Оптимизация использования ресурсов в клинической практике и, в частности, оптимизация использования операционных (operation rooms) является востребованной задачей в рамках создания СПКР [1]. Известны различные подходы к ее решению, в том числе подходы на основе теории расписаний с очередями и аналогичные, а также подходы на основе искусственного интеллекта (см., например, [2–6]). Однако следует подчеркнуть, что предметом рассмотрения в настоящей статье является не создание таких СПКР и не организация их сравнения, а разработка методики объяснимости их решений для конечного пользователя, что отражено и в названии статьи.

Актуальность темы статьи непосредственно вытекает из требования обеспечения доверия к системам искусственного интеллекта (ИИ), заявленного в [7], а также в многочисленных зарубежных источниках (например, [8]). Из набора характеристик,

составляющих доверительный ИИ, в настоящей статье авторы обратились к объяснимости (интерпретируемости) систем ИИ [9, 10], отсутствие которой, как показала практика последних лет, остается «камнем преткновения» для внедрения систем ИИ и, в частности, СПКР на их основе, в реальную практику работы медицинских учреждений.

В качестве **объекта рассмотрения** в статье выбрана СПКР, построенная на основе ИИ, а именно на основе многоагентного обучения с подкреплением, что соответствует передовым тенденциям в идеологии СППР в целом.

Преимуществами такого решения для СПКР являются возможность обработки нестационарности, большая гибкость в учете особенностей пациентов, возможность текущей корректировки решений на основе обратной связи. СПКР использует многоагентный симуляционный подход, в рамках которого каждому агенту естественного интеллекта (ЕИ) ставится в соответствие агент искусственного интеллекта (ИИ) с некоторым набором внутренних параметров, унаследованных от агента ЕИ. Агенты вступают между собой в игровое взаимодействие, организуемое на основе алгоритма проксимальной оптимизации политики (Proximal Policy Optimization, PPO) [11], который принадлежит к классу алгоритмов обучения с подкреплением (Reinforcement Learning, RL) [12]. Каждый агент преследует свои локальные цели, однако, помимо этого, все агенты стремятся к некоторой общей глобальной цели, в нашем случае – к снижению среднего времени обслуживания. При этом открывается возможность оценить реакцию на возникновение эмерджентного события, в данном случае – такого агента, который отказывается идти на компромиссы и занимает чисто

кооперативную (альтруист) или чисто состязательную (эгоцентрик) позицию.

Специфику построения СПКР рассматриваемого типа определяют два фактора. Во-первых, в период ожидания состояние пациента может ухудшиться, и в этом случае он будет нуждаться в оказании медицинской помощи в первую очередь/вне очереди, вплоть до проведения оперативного вмешательства по экстренным показаниям. Это может вызвать недовольство других пациентов, уже внесенных в расписание. Во-вторых, пациенты, ожидающие своей очереди на оказание медицинской помощи и лечение в клинике, могут, в силу личных/различных обстоятельств, изменять время/сроки проведения манипуляции/операции и т.п., что также может негативно сказаться на других пациентах.

Таким образом, при описании интеллектуальных агентов, репрезентирующих пациентов, необходимо учитывать их поведенческие особенности. С этой целью в рассматриваемой СПКР использован когнитивный фреймворк «убеждения–желания–намерения» (Belief–Desire–Intention, BDI) [13]. Убеждения интеллектуального агента отражают его представления о себе и внешней по отношению к нему среде, включая других агентов. Представления агента не обязательно являются правдой и могут измениться. Желания – это представления агента о состоянии внешней среды, к которому он стремится. Намерения – это активные решения агента, приближающие его к желаемому состоянию внешней среды.

В ходе проектирования игры параметры фреймворка BDI были включены в модель построения суждений агента и принятую систему ролей. Для управления процессом игры крайне желательно отслеживать в явном виде, в какой форме и в какой мере конкретный агент следует схеме BDI. Однако такое отслеживание представляет трудности для менеджера клиники, так как информация о процессе игры записывается только в лог игры.

Методы исследования. В настоящей работе рассматривается проблема повышения объяснимости поведения агентов в СПКР, построенной на основе многоагентного симуляционного подхода. В качестве средства ее решения предлагается использовать большую языковую модель (Large Linguistic Model,

LLM). Разработана методика ее применения в составе системы объяснимого ИИ на основе промптов, проведено сравнение результатов ее применения на разных типах LLM, которое подтвердило эффективность методики.

Анализ способов реализации объяснимости поведения агентов в СПКР (обзор литературы). Объяснимость искусственного интеллекта (eXplainability of Artificial Intelligence, XAI) – важный аспект доверия к системам искусственного интеллекта, которое, в свою очередь, является одним важнейших потребительских свойств и лимитируется законодательно [7]. Объяснимость систем ИИ и их компонентов является важным аспектом в развитии доверия. Пользователю системы важно понимать, как система ИИ принимает решения и какие ее компоненты были наиболее сильно задействованы в каждом отдельном акте процесса принятия решения.

Рассматривая СПКР как многоагентную систему, можно выделить в качестве наиболее близких подходов работы [14–16]. В работе [14] авторы рассматривают архитектуру программного обеспечения для разработки приложений Web.3, интегрирующих интеллектуальные MAC, работающие через децентрализованную автономную организацию путем формирования консенсуса на основе блокчейна. В качестве механизма объяснимости авторы используют широковебательную передачу информации от отдельного агента всем заинтересованным агентам и технологию публикации-подписки в реальном времени. Специфика поведения отдельных агентов, в том числе BDI-свойства, в таком подходе принципиально не учитывается.

В работе [15] в качестве подхода к объяснимости предлагается использовать язык i^* в качестве языка моделирования для графического представления действий агентов. Модель i^* представляет собой граф, вершинами которого являются элементы различных видов, таких как цели, задачи, ресурсы, а ребра задают отношения различных видов между элементами. Такая модель представляется эффективной для целей планирования коммуникации большого числа агентов в контексте интернета вещей (Internet of Things), но не для целей оперативного отслеживания поведения агентов в СПКР.

В работе [16] подчеркивается необходимость объяснимости поведения агентов МАС в реальном времени и предлагается использовать для этого фреймворк Real-Time Beliefs Desires Intentions (RT-BDI). Однако указанный фреймворк описан в работе только на концептуальном уровне, более того, явно выделены проблемы его программной реализации существующими средствами, в частности, с помощью JASON-подобных технологий.

Переходя к реализации ХАИ в системах, реализующих концепцию RL, следует выделить обзорные работы [17, 18], которые рассматривают не только классификацию подходов, но что особенно важно, классификацию метрик для измерения уровня ХАИ. В работе [18] предложена следующая классификация:

1) Точность (Fidelity) предполагает сравнение между решением, предлагаемым моделью с ХАИ, и неинтерпретируемой моделью.

2) Производительность (Performance) является стандартной метрикой для любой задачи и показывает, насколько есть уверенность в том, что модель будет работать ожидаемым образом при развертывании в целевой среде.

3) Понятность (Comprehensibility) относится к способности целевой аудитории понимать объяснение. Разработка тестов для оценки этой способности требует тщательного рассмотрения и интерпретации неопределенного термина «понимание». Оценка этого показателя может включать в себя отчет экспертов.

4) Предпочтительность (Preferability) может быть выражена в явной или неявной (скрытой) форме. Она является одним из важнейших показателей, так как более предпочтительные объяснения в идеале должны иметь высокую положительную корреляцию с их интегральной полезностью для пользователя при решении целевой задачи.

5) Действенность (Actionability) соотносится с тем, позволяют ли объяснения пользователям предпринимать осмысленные действия на основе их понимания объяснения.

6) Когнитивная нагрузка (Cognitive Load) отражает уровень когнитивных ресурсов, которые необходимо затратить пользователю для восприятия и реализации предлагаемого объяснения.

7) Визуализация (Visualizations) рассматривается авторами статьи как псевдометрика. Она обычно используется вместо количественной метрики для изучения интерпретируемости метода объяснения. Важно, что визуализации не обязательно должны быть изображениями: они могут включать описания на естественном языке, логические формулы и многое другое.

Обзор [17] отмечает сложность задачи формирования объяснимости в игровых моделях и считает, что здесь требуется объяснять не только действия игрока, но и его политику в целом. Поведенческие особенности игрока авторы обзора не выделяют как самостоятельный фактор. Говоря о классификации метрик объяснимости, авторы [17] выделяют ряд усложняющих факторов, таких как разнообразие трактовок объяснимости в целом и у различных типов аудитории, и в целом полагаются только на экспертные оценки, разделяя их на три уровня: уровень приложений, уровень человека и уровень функциональности [19].

В обзоре [20] предложена целевая классификация исследований в области ХАИ: объяснение сотрудничества агентов; визуализация; обобщение политики; объяснения на основе запросов; валидация. Сильной стороной обзора применительно к проблематике настоящей работы является фреймворк для составления запросов, направленных на объяснение поведения агентов. Фреймворк включает не только собственно формулировку запроса, но и задание релевантных ему состояний и их атрибутов, а также формирование ответа на запрос на естественном языке (хотя и из имеющегося набора предикатов, что затрудняет понимание ответа конечным пользователем).

Широкое внедрение LLM [21, 22] в практику разработки интеллектуальных систем дало новый импульс в области МАС: LLM предлагается использовать не только как самостоятельных агентов МАС [23], но и как средство объяснимости поведения других агентов [24, 25].

В работе [24] каждому автономному агенту придается своя LLM, которая создает текстовое описание хода игры; очевидно, что эти тексты во многом различаются. Для формирования единого объяснения предлагается проверять каждый такой текст единым сред-

ством ХАИ (типа Shap, LIME и др.), генерирующим набор соответствующих лемм. Множество таких наборов лемм вместе с предсказаниями LLM подается в алгоритм CART. Результатом работы CART является логическая структура лемм, имитирующая объяснение «обобщенной» LLM.

В работе [25] для объяснения поведения агента предлагается трехшаговая процедура. На первом шаге политика агента преобразуется в дерево решений.

На втором шаге из дерева решений извлекается путь решения для заданного состояния, который служит локальным представлением поведения агента. На третьем шаге он преобразуется в текстовое представление и подается в предварительно обученную LLM, которая генерирует объяснение на естественном языке.

Обобщая проведенный анализ, можно констатировать следующее.

Многоагентное обучение с подкреплением относится к наиболее сложным для интерпретации системам ИИ, и конкретные решения по формированию систем ХАИ для них слабо представлены в литературе.

Существующие подходы в основном направлены на выявление логической структуры отдельного решения или политики, реализуемых отдельным агентом или усредненных по группе агентов. В большинстве работ

агенты рассматриваются как однотипные, поведенческие особенности агентов, в том числе BDI-свойства, как самостоятельный фактор не выделяются.

При использовании LLM в качестве инструмента для формирования объяснения обычно используется многостадийная процедура: вначале LLM для каждого агента или группы агентов формирует текстовое описание хода игры, из него извлекается логическая структура, и по ней создается текстовое описание, предъявляемое пользователю.

В ряде работ отмечается необходимость категоризации метрик для измерения уровня ХАИ, а также доведения методик измерения до уровня полноценных фреймворков.

Разработка фреймворка для реализации ХАИ в СПКР. СПКР, являющаяся предметом исследования в настоящей статье, реализуется в игровой среде многоагентной симуляции, которая представляет собой двухнедельный календарь по M свободных ресурсов на каждый день. Число одновременно действующих в среде агентов N задается в качестве гиперпараметра. Поведенческие особенности агентов задаются через когнитивный фреймворк Belief-Desire-Intention.

Фрагмент игровой среды показан на рисунке 1-а, соответствующий ему марковский процесс принятия решений изображен на рисунке 1-б.

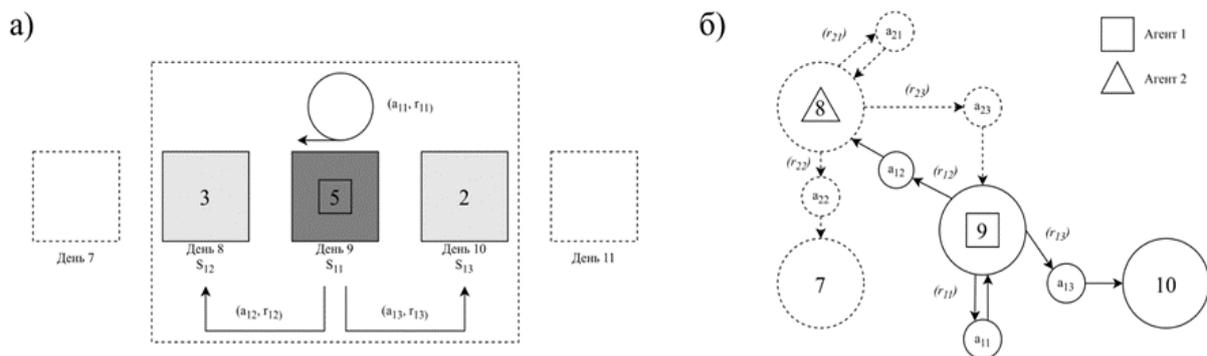


Рисунок 1 –
 а) – фрагмент игровой среды
 б) – соответствующий фрагмент процесса принятия решений

В процессе обучения модели собирается массив логов, которые в дальнейшем анализируются средствами LLM.

Эффективное использование LLM для реализации объяснимости ИИ зависит от умения правильно формулировать запросы, или промпты. В работах [26, 27] рассматриваются различные методики составления промптов к языковым моделям для получения наиболее достоверных и понятных результатов.

Для анализа логов системы мы разработали цикл промптов, руководствуясь методикой мысленной цепи (Chain of Thoughts, CoT) [28] и дополнив ее понятиями BDI. Вкратце, CoT в инженерии промптов относится к технике, при которой последовательность шагов рассуждения или промежуточных мыслей явно излагается в промпте, чтобы направлять языковую модель в сложных процессах решения проблем или принятия решений. Этот метод использует способность модели следовать шаблонам структурированного мышления, повышая ее способность давать последовательные и точные ответы, разбивая задачи на управляемые подзадачи.

Объединение концепций CoT и BDI позволило нам добиться объяснимости поведения агентов в рассматриваемой системе. Как предметное поле для анализа LLM использует лог игры. Разработанная методика представляется в виде следующего фреймворка (набора шагов).

- 1) Задать роль для LLM.
- 2) Описать контекст симуляции и конечную цель агентов.
- 3) Определить понятийный аппарат BDI.
 - a) Ввести структуру описания убеждений.
 - b) Ввести структуру описания желаний агентов.
 - c) Ввести структуру описания намерений агентов.
- 4) Описать фактические наблюдения и действия агентов.
- 5) Построить CoT.
 - a) Выделить интересующего агента.
 - b) Воссоздать BDI модель агента.
 - c) Обобщить поведение агента, отдельно выделить факторы, влияющие на мотивацию.

Разработанный фреймворк находится в русле работ по использованию различных аспектов мысленной цепи (Chain of Thoughts, CoT), формируемой в ходе промптинга LLM (см., например, обзор [29] и источники в нем). Однако предложенный в статье аспект, а именно отслеживание ролевых позиций участников игрового процесса через динамику изменения промптов, в доступной литературе не встречался, т.е. обладает элементами научной новизны.

Пример применения разработанного фреймворка представлен в Листинге 1. Фрагмент ответа LLM на сформированный промпт представлен в Листинге 2.

Листинг 1

Промпт к языковой модели

- 1) You are an intelligent system that assists in scheduling surgeries in a clinic. Your task is to find the optimal days for surgeries, considering multiple factors. You will work with observation data from agents, analyze this data, and make decisions based on it. Your role is to coordinate agents who analyze data from all other agents and make optimal decisions for scheduling surgeries.
- 2) The simulation context involves multiple agents, each observing various aspects of the current state of the clinic and trying to find the best day for performing surgeries. The ultimate goal of the agents is to optimally distribute surgeries across days to maximize the efficient use of the clinic's resources and meet patients' needs.
- 3) Read and remember this Belief-Desire-Intention (BDI) Conceptual Framework.
 - a) Beliefs structure. Urgency: rating from 1 to 3, where 1 is low urgency and 3 is high urgency. Completeness of Information: rating from 0 to 1, where 0 is incomplete information and 1 is complete information. Task Complexity: rating from 1 to 3, where 1 is low complexity and 3 is high complexity. Current Position in the Calendar for 14 Days: the agent's current position in the 14-day calendar. Slot Occupancy the Day Before: the occupancy of slots on the previous day. Slot Occupancy for the Current Day: the occupancy of slots on the current day. Slot Occupancy on the Next Day: the occupancy of slots on the next day.

- b) Desires Structure. Find the Best Day for Surgery. Considering all available data. Reduce Conflict and Competition for Slots. Among agents. Optimally Distribute Surgeries. For maximum clinic efficiency.
- c) Intentions Structure. 0 – enter the slot the next day, 1 – enter the slot the day before, 2 – do nothing.
- 4) Read and remember this fragment of the system logs: [system logs].
- 5) Chain of thought.
 - a) Please, identify the 1st agent
 - b) Then reconstruct the 1st Agent's BDI Model.
 - c) Finally, summarize the 1st Agent's Behavior. Characterize the agent's actions (passive, active), identify the motivating factors that make the agent act, and describe the dynamics of changes in the agent's physical condition.

Листинг 2

Ответ LLM на сформированный промпт

1. Beliefs: Urgency: 1 (low urgency), Completeness of Information: 1 (complete information), Task Complexity: 1 (low complexity), Current Position in the Calendar: [11, 11, 11, 11, 12], Slot Occupancy (Day Before): [3, 3, 2, 2, 3], Slot Occupancy (Current Day): [3, 4, 5, 5, 3], Slot Occupancy (Next Day): [1, 1, 1, 3, 6, 1, 2]. Desires: Find the best day for surgery, reduce slot conflicts among agents, optimize surgery distribution for clinic efficiency. Intentions (Actions): [2, 2, 2, 0, 2]

2. In summary, Agent 1 demonstrates a mostly passive behavior, driven by its beliefs of low urgency, complete information, and low task complexity. The agent only takes action in Step 4, possibly motivated by the favorable slot occupancy conditions on the next day. The dynamics of changes in Agent 1's physical condition are minimal and do not appear to have a substantial influence on its behavior.

Для оценки эффективности построенного фреймворка был составлен проблемно-ориентированный набор критериев, который в совокупности отвечает требованиям (1), (2), (3), (5), (6) [14], т.е. может рассматриваться как репрезентативный. В связи с отсутствием общепринятого бенчмарка для сравнительной оценки объяснимости с использованием LLM, была выполнена оценка самосогласованности разработанной методики по четырем различным LLM. Для этого с использованием разработанного набора критериев мы сравнили ответы четырех распространенных LLM по семи запускам игрового процесса с различными исходными данными.

Усредненные результаты проверки представлены в Таблице 1. Знак «+» означает, что требования критерия удовлетворены, «-» означает несоответствие и «+/-» – частичное соответствие.

Результаты свидетельствуют о том, что Claude Sonnet справляется с поставленной за

дачей лучше конкурирующих архитектур, демонстрируя лучшее выделение логической структуры игры в целом и поведения отдельных агентов, а также лучшую компоновку текстового материала.

Заключение. В работе рассмотрена проблема повышения объяснимости поведения агентов в СПКР на основе многоагентного обучения с подкреплением с учетом поведенческих особенностей отдельных агентов.

Проведен проблемно-ориентированный анализ способов реализации объяснимости поведения агентов. Показано, что выбранный для построения СПКР подход относится к наиболее сложным для интерпретации системам ИИ, и конкретные решения по формированию систем ХАИ для них слабо представлены в литературе.

В большинстве работ агенты рассматриваются как однотипные, поведенческие особенности агентов, в том числе BDI-свойства, как самостоятельный фактор не выделяются.

В то же время в ряде работ отмечается необходимость обоснованного подхода к отбору и формированию метрик для измерения уровня ХАИ.

В статье предложен фреймворк для реализации ХАИ в СПКР, ориентированный на представление поведенческих особенностей отдельных агентов в ходе игры, что очень важно для менеджера медицинского учреждения. Для экспериментальной оценки качества фреймворка разработан набор критериев оценки эффективности ХАИ, показана

его репрезентативность. Оценка проводилась путем сравнения объяснений, сформированных четырьмя распространенными LLM по семи запускам игрового процесса с различными исходными данными. Оценка показала хорошую воспроизводимость результатов, а также позволила выявить наиболее предпочтительную модель LLM. Предложенный подход к обеспечению объяснимости в многоагентных симуляциях открывает перспективы для дальнейшего развития и внедрения подобных систем в клиническую практику.

Таблица 1

Сравнение качества ответов разных LLM

Критерий	Тип LLM			
	HuggingChat	Gemini	Claude	Sonnet
Отсутствие орфографических ошибок	+	+	+	+
Отсутствие лексических ошибок	+	+	+	+
Отсутствие логических ошибок	–	+	+	+
Согласование времен	+	+	+	+
Использование профессиональной лексики	–	–	+	+
Разнообразие лексических конструкций	–	–	+	–
Отсутствие повторяющихся предложений	–	–	–	–
Соответствие заданному BDI шаблону	+/-	+/-	+	+

Список источников

1. Abdalkareem Z. A., Amir A., Al-Betar M. A., Ekhan, P., Hammouri A. I. Healthcare scheduling in optimization context: a review // *Health and Technology*. 2021. Т. 11. С. 445–469. (In Eng.).
2. Cardoen B., Demeulemeester E., Beliën J. Operating room planning and scheduling: A literature review // *European journal of operational research*. 2010. Т. 201. № 3. С. 921–932. (In Eng.).
3. Hancerliogullari G., Koksalmis E., Hancerliogullari K. O. Heuristics and simulated annealing algorithm for the surgical scheduling problem // *Heuristics, metaheuristics and approximate methods in planning and scheduling*. 2016. С. 225–241. (In Eng.).
4. Castro P. M., Marques I. Operating room scheduling with generalized disjunctive programming // *Computers & Operations Research*. 2015. Т. 64. С. 262–273. (In Eng.).
5. Hassanzadeh H., et al. Daily surgery caseload prediction: towards improving operating theatre

References

1. Abdalkareem Z. A., Amir A., Al-Betar M. A., Ekhan, P., Hammouri A. I. Healthcare Scheduling in Optimization Context: a Review. *Health and Technology*. 2021. Vol. 11. pp. 445–469.
2. Cardoen B., Demeulemeester E., Beliën J. Operating Room Planning and Scheduling: A Literature Review. *European Journal of Operational Research*. 2010. Vol. 201. No. 3. pp. 921–932.
3. Hancerliogullari G., Koksalmis E., Hancerliogullari K. O. Heuristics and Simulated Annealing Algorithm for the Surgical Scheduling Problem. *Heuristics, Metaheuristics and Approximate Methods in Planning and Scheduling*. 2016. pp. 225–241.
4. Castro P. M., Marques I. Operating Room Scheduling with Generalized Disjunctive Programming. *Computers & Operations Research*. 2015. Vol. 64. pp. 262–273.
5. Hassanzadeh H., et al. Daily Surgery Caseload Prediction: Towards Improving Operating Theatre

- efficiency // *BMC Medical Informatics and Decision Making*. 2022. Т. 22. № 1. С. 151. (In Eng.).
6. Bruni M. E., Beraldi P., Conforti D. A stochastic programming approach for operating theatre scheduling under uncertainty // *IMA Journal of Management Mathematics*. 2015. Т. 451 26. № 1. С. 99–119. (In Eng.).
7. ГОСТ Р 59276–2020 Системы искусственного интеллекта. Способы обеспечения доверия. Общие положения. Дата введения 23 декабря 2020 г. – М.: Стандартинформ, 2021. – 16 с.
8. Schwartz R., et al. A Proposal for identifying and managing bias in artificial intelligence. Draft // National Institute of Standards and Technology. Special Publication. 2021. (In Eng.).
9. Li B., et al. Trustworthy AI: From Principles to Practices // *ACM Computing Surveys*. 2023. Т. 55. № 9. С. 1–46. DOI: 10.1145/3555803. (In Eng.).
10. Linardatos P., Papastefanopoulos V., Kotsiantis S. Explainable AI: A Review of machine learning interpretability methods // *Entropy*. 2021. № 23 (1). С. 18. (In Eng.). DOI: 10.3390/e23010018.
11. Schulman J., et al. Proximal Policy Optimization Algorithms [Электронный ресурс]. – Режим доступа: <https://www.semanticscholar.org/reader/dce6f9d4017b1785979e7520fd0834ef8cf02f4b> (In Eng.).
12. Саттон Р. С., Барто Э. Г. Обучение с подкреплением = Reinforcement Learning. 2-е издание. – М.: ДМК пресс, 2020. – 552 с.
13. Georgeff M., et al. The belief-desire-intention model of agency // *Intelligent Agents V: Agents Theories, Architectures, and Languages: 5th International Workshop, ATAL'98 Paris, 469 France, July 4–7, 1998 Proceedings 5*. – Springer Berlin Heidelberg, 1999. – С. 1–10. (In Eng.).
14. De Gasperis G., Facchini S.D. MAS, DAO and DLT: a 3 legs architecture for intelligent services // *6th Distributed Ledger Technologies Workshop (DLT2024)*, May 14-15, 2024, Turin, Italy. (In Eng.).
15. Alzetta F, Giorgini P, Najjar A, Schumacher MI, Calvaresi D. In-time explainability in multi-agent systems: challenges, opportunities, and roadmap // *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*. 2020. Т. 12175. С. 39–53. DOI: 10.1007/978-3-030-51924-7_3. (In Eng.).
16. Ciatto G., et al. Towards XMAS: eXplainability through multi-agent systems // *Artificial Intelligence and Internet of Things*. 2019. Т. 2502. Ст. 3. (In Eng.).
17. Heuillet A., Couthouis F., Díaz-Rodríguez N. Explainability in deep reinforcement learning // *Knowledge-Based Systems*. 2021. Т. 214. С. 106685. DOI: 10.1016/j.knosys.2020.106685. (In Eng.).
18. Milani S., et al. Explainable reinforcement learning: a survey and comparative review // *ACM Computing Surveys*. 2023. Т. 56. № 7. С. 1–36. DOI: 10.1145/36168. (In Eng.).
- Efficiency. *BMC Medical Informatics and Decision Making*. 2022. Vol. 22. No. 1. P. 151.
6. Bruni M. E., Beraldi P., Conforti D. A Stochastic Programming Approach for Operating Theatre Scheduling under Uncertainty. *IMA Journal of Management Mathematics*. 2015. Vol. 451 26. No. 1. pp. 99–119.
7. GOST R 59276–2020 Artificial Intelligence Systems. Methods of Ensuring Trust. General Provisions. Date of introduction December 23, 2020. *Moscow, Standartinform*. 2021. 16 p. (In Russ.).
8. Schwartz R., et al. A Proposal for Identifying and Managing Bias in Artificial Intelligence. Draft. *National Institute of Standards and Technology. Special Publication*. 2021.
9. Li B., et al. Trustworthy AI: From Principles to Practices. *ACM Computing Surveys*. 2023. Vol. 55. No. 9. pp. 1–46. DOI: 10.1145/3555803.
10. Linardatos P., Papastefanopoulos V., Kotsiantis S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*. 2021. No. 23 (1). pp. 18. DOI: 10.3390/e23010018.
11. Schulman J., et al. Proximal Policy Optimization Algorithms. Available at: <https://www.semanticscholar.org/reader/dce6f9d4017b1785979e7520fd0834ef8cf02f4b>
12. Sutton R. S., Barto E. G. Reinforcement Learning. 2nd edition. *Moscow, DMK press*. 2020. 552 p. (In Russ.).
13. Georgeff M., et al. The belief-desire-intention model of agency. *Intelligent Agents V: Agents Theories, Architectures, and Languages: 5th International Workshop, ATAL'98 Paris, 469 France, July 4–7, 1998 Proceedings 5*. *Springer Berlin Heidelberg*. 1999. pp. 1–10.
14. De Gasperis G., Facchini S.D. MAS, DAO and DLT: a 3 Legs Architecture for Intelligent Services. *6th Distributed Ledger Technologies Workshop (DLT2024)*, May 14-15, 2024, Turin, Italy.
15. Alzetta F, Giorgini P, Najjar A, Schumacher MI, Calvaresi D. In-Time Explainability in Multi-Agent Systems: Challenges, Opportunities, and Roadmap. *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*. 2020. Т. 12175. С. 39–53. DOI: 10.1007/978-3-030-51924-7_3.
16. Ciatto G., et al. Towards XMAS: eXplainability Through Multi-Agent Systems. *Artificial Intelligence and Internet of Things*. 2019. Vol. 2502. Paper 3.
17. Heuillet A., Couthouis F., Díaz-Rodríguez N. Explainability in Deep Reinforcement Learning. *Knowledge-Based Systems*. 2021. Vol. 214. pp. 106685. DOI: 10.1016/j.knosys.2020.106685.
18. Milani S., et al. Explainable Reinforcement Learning: A Survey and Comparative Review. *ACM Computing Surveys*. Vol. 56. No. 7. pp. 1–36. DOI: 10.1145/36168.

19. Doshi-Velez F., Kim B. Towards a rigorous science of interpretable machine learning. 2017 [Электронный ресурс]. – Режим доступа: <https://arxiv.org/pdf/1702.08608.pdf> (In Eng.).
20. Wells L., Bednarz T. Explainable AI and reinforcement learning – a systematic review of current approaches and trends. 2021 [Электронный ресурс]. – Режим доступа: <https://doi.org/10.3389/frai.2021.550030> (In Eng.).
21. Zhao H., et al. Explainability for Large Language Models: a survey [Электронный ресурс]. – Режим доступа: <https://arxiv.org/abs/2309.01029> (In Eng.).
22. Luo H., Specia L. From understanding to utilization: a survey on explainability for Large Language Models [Электронный ресурс]. – Режим доступа: <https://arxiv.org/abs/2401.12874> (In Eng.).
23. Han S., Zhang, Yao Q., Y., Jin W., Xu Z., He Ch. LLM Multi-Agent systems: challenges and open problems [Электронный ресурс]. – Режим доступа: <https://arxiv.org/abs/2402.03578> (In Eng.).
24. Agiollo A., Siebert L. C., Murukannaiah P. K., et al. From Large Language Models to small logic programs: building global explanations from disagreeing local post-hoc explainers // SpringerNature Link [Электронный ресурс]. – Режим доступа: <https://doi.org/10.1007/s10458-024-09663-8> (In Eng.).
25. Zhang X., et al. Explaining agent behavior with Large Language Models [Электронный ресурс]. – Режим доступа: <https://arxiv.org/abs/2309.10346> (In Eng.).
26. Sahoo P., et al. A Systematic survey of prompt engineering in Large Language Models: techniques and applications [Электронный ресурс]. – Режим доступа: <https://arxiv.org/abs/2402.07927> (In Eng.).
27. Chen B., et al. Unleashing the potential of prompt engineering in Large Language Models: 476 a comprehensive review [Электронный ресурс]. – Режим доступа: <https://arxiv.org/abs/2310.14735> (In Eng.).
28. Wei J., et al. Chain-of-thought prompting elicits reasoning in Large Language Models // 478 Advances in neural information processing systems. 2022. Т. 35. С. 24824–24837. (In Eng.).
29. Sahoo P., et al. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications [Электронный ресурс]. – Режим доступа: <https://arxiv.org/abs/2402.07927> (In Eng.).
19. Doshi-Velez F., Kim B. Towards a Rigorous Science of Interpretable Machine Learning. 2017. Available at: Режим доступа: <https://arxiv.org/pdf/1702.08608.pdf> (In Eng.).
20. Wells L., Bednarz T. Explainable AI and Reinforcement Learning – A Systematic Review of Current Approaches and Trends. Available at: <https://doi.org/10.3389/frai.2021.550030>
21. Zhao H., et al. Explainability for Large Language Models: A Survey. Available at: <https://arxiv.org/abs/2309.01029>
22. Luo H., Specia L. From Understanding to Utilization: A Survey on Explainability for Large Language Models. Available at: <https://arxiv.org/abs/2401.12874>
23. Han S., Zhang, Yao Q., Y., Jin W., Xu Z., He Ch. LLM Multi-Agent Systems: Challenges and Open Problems. Available at: <https://arxiv.org/abs/2402.03578>
24. Agiollo A., Siebert L. C., Murukannaiah P. K., et al. From Large Language Models to Small Logic Programs: Building Global Explanations from Disagreeing Local Post-hoc Explainers. *SpringerNature Link* Available at: <https://doi.org/10.1007/s10458-024-09663-8>
25. Zhang X., et al. Explaining agent behavior with Large Language Models. Available at: <https://arxiv.org/abs/2309.10346>
26. Sahoo P., et al. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. Available at: <https://arxiv.org/abs/2402.07927>
27. Chen B., et al. Unleashing the Potential of Prompt Engineering in Large Language Models: 476 a Comprehensive Review. Available at: <https://arxiv.org/abs/2310.14735>
28. Wei J., et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *478 Advances in Neural Information Processing Systems*. 2022. Vol. 35. pp. 24824–24837.
29. Sahoo P., et al. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. Available at: <https://arxiv.org/abs/2402.07927>