КАРПОВ О.Э.,
академик РАН, д.м.н., профессор, ФГБУ «Национальный медико-хирургический Центр имени Н.И. Пирогова», Москва, Россия, e-mail: nmhc@mail.ru

АНДРИКОВ Д.А.,
к.т.н., ООО «Иммерсмед», Москва, Россия, e-mail: andrikovda@immersmed.ru

МАКСИМЕНКО В.А.,
д.ф.м.н., Университет Иннополис, г. Казань, Россия, e-mail: maximenkovl@gmail.com

ХРАМОВ А.Е.,
д.ф.м.н., профессор, Балтийский федеральный университет им. И. Канта, г. Калининград, Россия, e-mail: hramovae@gmail.com

ПРОЗРАЧНЫЙ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ ДЛЯ МЕДИЦИНЫ

DOI: 10.25881/18110193_2022_2_4

Аннотация.
Успех и массовое применение современных технологий искусственного интеллекта (ИИ) и, в частности, методов глубокого обучения нейронных сетей привели нас к четкому пониманию двух основных проблем: проблемы ошибок (проблема надежности) и проблемы явного объяснения решений, принимаемых ИИ (проблема прозрачности). Эти проблемы тесно связаны между собой: необъяснимые ошибки ИИ могут повторяться снова и снова. Это совершенно неприемлемо с точки зрения применения ИИ в здравоохранении, потому что является критичным для жизни и здоровья пациентов. Если оставить проблемы ошибок и объяснимости нерешенными, то непрозрачность решений ИИ может привести к отказу или существенному ограничению от использования систем ИИ в задачах медицины. В данном комментарии мы обсуждаем проблемы прозрачного объяснимого интеллекта для медицины и рассматриваем различные подходы к их решению.

Ключевые слова: *искусственный интеллект, цифровые медицинские технологии, биомаркер, нейроинтерфейс*

Для цитирования: *Карпов О.Э., Андриков Д.А., Максименко В.А., Храмов А.Е. Прозрачный искусственный интеллект для медицины. Врач и информационные технологии. 2022; 2: 4-11. doi: 10.25881/18110193_2022_2_4.*

KARPOV O.E.,
Academician of the RAS, Dr. Sci. (Medicine), Professor, Pirogov National Medical and Surgical Center, Moscow, Russia, e-mail: nmhc@mail.ru

ANDRIKOV D.A.,
PhD, Immersmed LLC, Moscow, Russia, e-mail: andrikovda@immersmed.ru

MAKSIMENKO V.A,
Dr. Sci., Innopolis University, Kazan, Russia, e-mail: maximenkovl@gmail.com

HRAMOV A.E.,
Dr. Sci., Professor, Immanuel Kant Baltic Federal University, Kaliningrad, Russia, e-mail: hramovae@gmail.com

EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR MEDICINE

DOI: 10.25881/18110193_2022_2_4

Abstract.
The success and wide-range applications of artificial intelligence (AI) technologies and, in particular, deep learning neural networks methods have led us to a clear understanding of two main problems: the problem of errors (the reliability problem) and the problem of explicitly explaining the decisions made by AI (the explanability problem). These problems are closely related: unexplained AI errors can happen again and again. This is completely unacceptable from the perspective of AI applications in health care because it is critical to the lives and health of patients. If left unresolved, the problems of error and explainability can lead to the rejection or significant restriction of AI systems in medical applications. In this paper, we discuss the problems of explainable artificial intelligence (XAI) for medicine and consider different approaches to solving them.

Keywords: *artificial intelligence, digital healthcare, biomarker, neurointerface*

For citation: *Karpov O.E., Andrikov D.A., Maksimenko V.A, Hramov A.E. Explainable artificial intelligence for medicine. Medical doctor and information technology. 2022; 2: 4-11. doi: 10.25881/18110193_2022_2_4.*

Выражение «медицинские технологии» широко используется для обозначения целого ряда инструментов, которые могут позволить специалистам здравоохранения обеспечить пациентам и обществу лучшее качество жизни путем проведения ранней диагностики, снижения осложнений и оптимизации лечения. Если раньше медицинские технологии были известны в основном как классические медицинские устройства (например, имплантаты, протезы, стенты, системы функциональной диагностики), то развитие информационных технологий привело к революции цифровых медицинских продуктов и сервисов, среди которых наибольшие надежды возлагаются на технологии искусственного интеллекта (ИИ). Методы ИИ, такие как нечеткие экспертные системы, байесовские сети, искусственные нейронные сети и гибридные интеллектуальные системы, все активнее используются в здравоохранении. Следуя многочисленным публикациям, например, в профильном журнале «Врач и информационные технологии» и докладу CB Insights [1] в последние годы наибольшие затраты на исследования в области ИИ были связаны с медицинскими применениями. В 2021 году объем инвестиций в ИИ в здравоохранении характеризуется рекордными $11.2 млрд., при этом годовой рост составил 64% по сравнению с 2020 годом [2].

Мы рассматриваем ИИ как часть информационных технологий, способную решать сложные задачи в областях, где накоплены большие наборы размеченных (заранее подготовленных экспертом с соответствующими отметками для обучения ИИ) данных, но для которых нет четких однозначных правил принятия решений. Технология ИИ эффективна там, где нельзя задать четкие правила, формулы и алгоритмы для решения задачи, например, есть ли на рентгенограмме легких патология? Технологии ИИ предполагают, что вместо реализации некоторой заранее сформулированной логической формулы на базе четких инструкций типа «если..., то...», алгоритм обучают с помощью большого количества заранее подготовленных данных и различных методов, которые дают компьютерной программе возможность выявить эту формулу на основе эмпирических данных и тем самым научиться выполнять задачу в будущем, даже в несколько иных условиях. Действительно, основой доказательной медицины является установление клинических взаимосвязей и представлений путем выявления корреляций, ассоциаций и закономерностей на основе собираемых и/или существующих баз данных. Традиционно в биомедицине для установления этих закономерностей и корреляций использовались статистические методы. Методы ИИ в случае медицинских приложений предполагают обучение интеллектуальной системы с помощью повторяющихся алгоритмов распознаванию того, как выглядят определенные группы симптомов или определенные клинические/радиологические изображения или временные ряды, то есть фактически классифицировать биомаркеры тех или иных заболеваний [3–5].

Биомаркер указывает на медицинский (биологический) признак, который можно измерить объективно, точно и воспроизводимо и который можно использовать в качестве индикатора состояния всего организма [5]. Например, высокий уровень свинца в крови может указывать на необходимость проверки нервной системы и когнитивных расстройств, особенно у детей. Высокий уровень холестерина является распространенным биомаркером риска сердечных заболеваний. Всемирная организация здравоохранения определила биомаркер как «практически любое измерение, отражающее взаимодействие между биологической системой и потенциальной опасностью, которая может быть химической, физической или биологической. Измеренный ответ может быть функциональным и физиологическим, биохимическим на клеточном уровне или молекулярным взаимодействием» [6].

Медицинские системы искусственного интеллекта (СИИ) существуют во многих формах, от чисто виртуальных (например, системы управления медицинской информацией на основе глубокого обучения и помощи врачами при принятии решений о лечении) до киберфизических (например, роботы, используемые для помощи лечащему хирургу, и нанороботы для адресной доставки лекарств). Возможности технологий ИИ по распознаванию сложных моделей и скрытых структур позволили многим системам обнаружения и диагностики на основе изображений в здравоохранении работать не

хуже, а в некоторых случаях и лучше врачей [7]. Системы поддержки принятия врачебных решений с использованием ИИ могут уменьшить количество диагностических ошибок, расширить интеллектуальные возможности для более эффективной диагностики и лечения, а также повысить эффективность ведения электронных медицинских карт и документирования. Появляющиеся вычислительные усовершенствования в области обработки естественного языка (англ. natural language processing, NLP), идентификации биомаркеров, эффективного поиска, прогнозирования и беспристрастного рассуждения приведут к дальнейшему развитию возможностей ИИ для решения неразрешимых в настоящее время проблем.

В связи с появлением новых медицинских устройств, использующих ИИ, разгорелась дискуссия о том, должна ли логика, лежащая в основе ИИ, быть понятной врачу и пациенту. Иными словами, переходя к широко используемой в кибернетике аналогии «черного ящика», технология медицинского ИИ должна быть «прозрачной», то есть прозрачность (или объяснимость) можно понимать как характеристику системы, управляемой ИИ, позволяющую человеку восстановить, почему ИИ пришел к тем или иным выводам. В интересной заметке, опубликованной в ведущем научном журнале по медицине Nature Medicine [8], рассказывается о следующем случае. Чтобы проанализировать настроения в обществе, на одной из конференций участникам был задан следующий вопрос: «*Предположим, у вас рак и вам нужна операция по удалению опухоли. Какого из двух хирургов вы бы выбрали, если бы вам пришлось выбирать между хирургом-человеком, вероятность смерти которого составляет 15%, и хирургом-роботом, вероятность смерти которого составляет 2%, с оговоркой, что никто не знает, как работает робот, и ему нельзя задавать вопросы?*» Все присутствующие, кроме одного, предпочли человека.

Робот-хирург в этом примере моделировал опасность «черного ящика» — отсутствие прозрачности в логике работы современных медицинских систем, использующих ИИ. Даже когда традиционный «непрозрачный» ИИ может выявить закономерность, указывающую на неизбежность заболевания, мы обычно не можем объяснить логику, лежащую в основе этого решения. Существуют технологические причины, по которым создание объяснимых СИИ является сложной задачей; «логика черного ящика» остается камнем преткновения.

Во-многом это определяется тем, что современные системы ИИ имеют сложную архитектуру и включают миллионы элементарных вычислительных элементов. Алгоритмы обучения также развиваются, например трансферное обучение позволяет использовать ИИ, обученный решать определенную задачу, для решения другой задачи. Все это приводит к тому, что ИИ находит закономерности и решения, которые нельзя найти другими средствами. Однако, алгоритмы нахождения этих закономерностей становятся все менее понятными и интерпретируемыми.

Итак, можно ли доверять таким решениям и быть уверенным в том, что ИИ не ошибется в критической ситуации, когда на кону стоит здоровье и даже жизнь человека? Возможно, одним из определяющих моментов является тот прирост эффективности, который дает ИИ в сравнении с традиционными подходами. Если он очень велик, возможно ли, что это перевесит вероятность ошибки, какую может допустить ИИ в силу своих скрытых особенностей, которые мы не можем интерпретировать?

Мы считаем, что в решении проблемы прозрачности медицинского ИИ видится два пути:

1. Пытаться сделать алгоритмы ИИ полностью интерпретируемыми.
2. Использовать системы ИИ в качестве ассистентов и систем поддержки принятия решений врачом.

Что касается первого направления, то есть мнение, что сделать ИИ полностью интерпретируемым не получится. Повышение интерпретируемости в общем случае будет вызывать упрощение моделей и стоить их эффективности, так что «игра не будет стоить свеч».

Второе направление оказалось выходом из ситуации, которая возникла на фоне первоначальных достижений в вычислительных возможностях ИИ, заключающейся в опасности, что технологии ИИ в конечном итоге заменят врачей. На смену этой парадигме пришло понятие «*дополненного интеллекта*», предложенного У.Р. Эшби еще в 1950-х годах [9], которое

Рисунок 1 — Иллюстрация фундаментальной теоремы биоинформатики Фридмана, описывающая влияние дополненного интеллекта. Применительно к теме нашего обсуждения она будет звучать как «Система здравоохранения с искусственным интеллектом будет лучше, чем система здравоохранения без него».

наиболее точно определяет взаимодействия между данными, вычислениями и медицинскими работниками и, возможно, является лучшим определением для «ассистирующего ИИ» в медицине. Версия дополненного интеллекта, описанная в фундаментальной теореме Фридмана о биомедицинской информатике [10], имеет непосредственное отношение к роли ИИ в здравоохранении (Рис. 1). В соответствии с описанием дополненного интеллекта Фридмана, ее можно сформулировать так «*Врачи, использующие искусственный интеллект, заменят врачей, которые его не используют*». Хотим обратить внимание, что мы не утверждаем факт замены врачей искусственным интеллектом, а подчеркиваем именно ассистирующую роль ИИ.

В 2019 году стартовал инициативный проект по созданию ассистирующей системы для врача-эпилептолога (система поддержки принятия врачебных решений), который выполняется на клинической базе ФГБУ «НМХЦ им. Н.И. Пирогова» Минздрава России). Совместно с резидентом Сколково ООО «Иммерсмед» было создано программное решение по экспресс-диагностике эпилепсии. Это настоящий междисциплинарный проект, медицинская составляющая которого включает проведение анализа большого числа клинически исследований, а фундаментальная часть проекта предусматривает создание математических основ описания таких сложных нестационарных сигналов как ЭЭГ с эпилептическими разрядами и разработкой системы ИИ для их выделения [11].

Нашей задачей стало создание системы ИИ, которая позволяла бы разметить многодневные записи ЭЭГ для диагностики эпилептических приступов в условиях записей низкого качества, различного используемого оборудования для регистрации, различного числа каналов регистрации и т.д. Обычно такие данные анализирует непосредственно врач-эпилептолог, и их рутинная расшифровка может занимать от нескольких часов до целого рабочего дня. Мы предположили, что редкость эпилептических приступов может стать важной особенностью нашей системы и пошли нестандартным путем. Мы не стали использовать классическую систему обучения ИИ с учителем, когда по ранее размеченным данным, на которых выделялись эпилептические разряды, нейронную сеть обучали на выявление подобных событий (эпилептических разрядов). Наоборот, мы учли особенности несбалансированности данных, когда количество интересующих нас событий (разрядов) на порядки меньше, чем фоновая активность (в течении 2–3 суток типично наблюдается от 1 до 4 разрядов длительностью несколько минут). Был применен подход классификации без учителя и выделены физиологические особенности приступа, которые выделялись одноклассовым классификатором [12] на базе машины опорных векторов. В результате мы получили систему, которая не зависела ни от оборудования, ни от условий регистрации сигналов, потому что классификатор «работал» отдельно с каждой индивидуальной записью пациента. Не все выделенные эпизоды были истинными разрядами, но это было заложено в основу работы системы, когда сужали число интересующих событий до примерно 20 для четырехсуточной записи ЭЭГ. Далее врач-эпилептолог рассматривал только эти события и

выделял те, которые истинно являлись приступами. Время рутинной работы врача сократилось до 5–10 минут.

Мы получили полностью прозрачный ИИ за счет объединения возможностей ИИ сузить количество «подозрительных» участков записи, которые вероятно могли содержать эпилептические приступы, и последующей интерпретации этих данных естественным интеллектом врача. «Симбиоз» ИИ и врача оказался весьма успешным, и система в настоящее время внедряется в клиническую практику [13].

Итак, ассистирующие врачу технологии ИИ позволяют решить ряд проблем прозрачного ИИ, но не для всех задач. Например, если речь идет об автоматизированных системах, поддерживающих жизнь и здоровье человека в реальном времени:

- нейроинтерфейсы для контроля эпилепсии [14];
- системы автоматического введения инсулина [15];
- кардиостимуляторы и имплантируемые кардиовертеры-дефибрилляторы [16] и др.;
- нейроинтерфейсы для реабилитации и управления внешними устройствами [17],

то в таких системах решение принимается не один раз, а каждую секунду. Чем больше решений — тем больше вероятность ошибки из-за скрытых особенностей ИИ, которые мы не можем интерпретировать и объяснить. А скорость и автономность работы таких систем не позволяет каждый их шаг «согласовывать» с врачом.

Рассмотрим эту ситуацию на примере конкретного случая. В 2013 году был начат проект по предсказанию в реальном времени эпилептических приступов при абсанс-эпилепсии совместно Университетом Радбауд (Неймеген, Нидерланды) и Вестфальским Университетом (Мюнстер, Германия). Задачей проекта являлась разработка нейроинтерфейса, который предсказывает приближение приступа и предотвращает его путем электрической стимуляции мозга. Система тестировалась на часто используемой животной модели абсанс-эпилепсии — крысах линии WAG/Rij [18].

- Первая версия системы, которая основывалась чисто на интерпретациях известных принципов формирования эпилептического разряда, практически в 100% случаев указывала на приближение приступа, но также могла принять за приступ другое состояние. Во время каждого предсказания в мозг посылается электрический импульс. Поэтому нельзя было допустить большого числа ложных стимуляций [19].
- За несколько лет работы мы пытались решить проблему различными методами, основанными на интерпретируемых подходах физики и нейробиологии. В результате вторая версия системы предсказывала около 60% приступов, но число ложных предсказаний было на 80% меньше [20].
- Наконец, применение ИИ позволило еще на 71% сократить число стимуляций и поднять точность предсказания до 80%. По своим показателям третья версия системы безопасна для испытания на человеке, однако она теперь содержит элемент ИИ, и интерпретировать все решения такой системы невозможно [21].

Хотим отметить, что процедура формирования признаков, по которым ИИ отличает приступ от нормальной активности, в разработанной системе прозрачна и основана на фундаментальных знаниях о работе мозга. Однако в силу сложности и вариабельности регистрируемых сигналов для каждого отдельного события эти признаки не всегда иллюстративны. ИИ позволил «выучить» эти вариабельности на предварительно размеченных наборах данных и обработать их в режиме реального времени.

Известно, что «хороший врач лечит болезнь, а великий врач лечит пациента, у которого есть болезнь». Отношения между врачом и пациентом основаны на общении и доверии. Без объяснимого с медицинской точки зрения ИИ врачу будет практически нечего сообщить пациенту, что приведет к потере доверия и удовлетворенности пациента. Прозрачность логики, которая может быть реализована через дополненный ассистирующий искусственный интеллект, может расширить возможности врачей, не лишая их самостоятельности, что может открыть дверь для более широкого использования ИИ в здравоохранении.

Работа поддержана Президентским грантом (проект НШ-589.2022.1.2).

ЛИТЕРАТУРА/REFERENCES

1. CB Insights Research. Healthcare remains the hottest AI category for deals. April 12, 2017. [Last accessed on 2022 Feb 11]. Available at: https://www.cbinsights.com/research/artificial-intelligence-healthcare-startups-investors.
2. На основе данных State of AI Q3'21. [Last accessed on 2022 March 22]. Available at: https://www.cbinsights.com/research/report/ai-trends-q3-2021.
3. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nature medicine. 2019; 25(1): 30-36.
4. Johnson Kevin B, et al. Precision medicine, AI, and the future of personalized health care. Clinical and translational science. 2021;14(1): 86-93.
5. Карпов О.Э., Храмов А.Е. Прогностическая медицина // Врач и информационные технологии. — 2021. — №3. — С.20-37. [Karpov OE, Hramov AE. Prognosticheskaya medicina. Vrach i informacionnye tekhnologii. 2021; 3: 20-37. (In Russ).] doi: 10.25881/18110193_2021_3_20.
6. Strabo K, Tavel JA. What are biomarkers? Current Opinion in HIV and AIDS. 2010; 5(6): 463.
7. Toprol E.J. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. Basic Books, New York, NY, 2019.
8. Kundu S. AI in medicine must be explainable. Nature Medicine. 2021; 27(8): 1328.
9. Ashby WR. An Introduction to Cybernetics. Chapman & Hall Ltd., London, UK, 1957.
10. Friedman CP. A «fundamental theorem» of biomedical informatics. J. Am. Med. Inform. Assoc. 2009; 16: 169-170.
11. Karpov OE, Grubov VV, Maksimenko VA, Utaschev N, Semerikov VE, Andrikov DA, Hramov AE. Noise amplification precedes extreme epileptic events on human EEG. Physical Review E. 2021; 103: 022310.
12. Perera P, Patel VM. Learning deep features for one-class classification. IEEE Transactions on Image Processing. 2019; 28(11): 5450-5463.
13. Кучин А.С., Грубов В.В., Максименко В.А., Утяшев Н.П. Автоматизированное рабочее место врача эпилептолога с возможностью автоматического поиска приступов эпилепсии // Врач и информационные технологии. — 2021. — №3. — С.62-73. [Kuchin AS, Grubov VV, Maksimenko VA, Utyashev NP. Avtomatizirovannoe rabochee mesto vracha epileptologa s vozmozhnost'yu avtomaticheskogo poiska pristupov epilepsii. Vrach i informacionnye tekhnologii. 2021; 3: 62-73. (In Russ).] doi: 1025881/18110193_2021_3_62.
14. Chaudhary U, Birbaumer N, Ramos-Murguialday A. Brain–computer interfaces for communication and rehabilitation. Nature Reviews Neurology. 2016; 12(9): 513-525.
15. Shah RB, Patel M, Maahs DM, Shah V N. Insulin delivery methods: Past, present and future. International journal of pharmaceutical investigation. 2016; 6(1): 1.
16. Ricci RP, Morichelli L, Santini M. Home monitoring remote control of pacemaker and implantable cardioverter defibrillator patients in clinical practice: impact on medical management and health-care resource utilization. Europace. 2008; 10(2): 164-170.
17. Hramov AE, Maksimenko VA, Pisarchik AN. Physical principles of brain–computer interfaces and their applications for rehabilitation, robotics and control of human brain states. Physics Reports. 2021; 918: 1-133.

18. Sarkisova K, van Luijtelaar G. The WAG/Rij strain: a genetic animal model of absence epilepsy with comorbidity of depressiony. Progress in Neuro-Psychopharmacology and Biological Psychiatry. 2011; 35(4): 854-876.

19. van Luijtelaar G, Lüttjohann A, Makarov VV, Maksimenko VA, Koronovskii AA, Hramov AE. Methods of automated absence seizure detection, interference by stimulation, and possibilities for prediction in genetic absence models. Journal of Neuroscience Methods. 2016; 260: 144-158.

20. Maksimenko VA, Heukelum S, Makarov VV, Kelderhuis J, Lüttjohann A, Koronovskii AA, Hramov AE, Luijtelaar G. Absence Seizure Control by a Brain Computer Interface. Scientific Reports. 2017; 7: 2487.

21. Budde B, Maksimenko V, Sarink K, Seidenbecher T, van Luijtelaar G, Hahn T, Pape HC, Lüttjohann A. Seizure prediction in genetic rat models of absence epilepsy: improved performance through multiple-site cortico-thalamic recordings combined with machine learning. Eneuro. 2022; 9(1).