ORES

Научно-практический электронный журнал
Оригинальные исследования (ОРИС), *ISSN 2222-0402*
www.ores.su
original-research.ru

ОБЪЯСНИМЫЙ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ В ИНФОРМАЦИОННЫХ СИСТЕМАХ: ОБЗОР ТЕКУЩЕГО СОСТОЯНИЯ И БУДУЩИХ НАПРАВЛЕНИЙ ИССЛЕДОВАНИЙ

Хвостова Алёна Леонидовна,
соискатель, МИРЭА - Российский технологический университет
Россия, г. Москва
kartes909@yandex.ru

Аннотация

Искусственный интеллект проник во многие сферы личной и профессиональной жизни. Однако из-за растущей сложности базовых моделей и алгоритмов ИИ представляется «черным ящиком», поскольку внутренние процессы обучения, а также результирующие модели не полностью понятны. Этот компромисс между производительностью и объяснимостью может оказать существенное влияние на отдельных людей, бизнес и общество в целом. Объяснимый искусственный интеллект способен ответить на этот вызов, расширив понимание внутренней логики прогнозов, сделанных сложными моделями машинного обучения.

Ключевые слова: объяснимый искусственный интеллект, XAI, информационные системы, интерпретируемость, машинное обучение, прозрачность алгоритмов.

EXPLAINABLE ARTIFICIAL INTELLIGENCE IN INFORMATION SYSTEMS: A REVIEW OF THE CURRENT STATE OF THE ART AND FUTURE RESEARCH DIRECTIONS

Alyona L. Khvostova,
applicant, MIREA - Russian Technological University
Russia, Moscow
kartes909@yandex.ru

ABSTRACT

Artificial intelligence has permeated many areas of personal and professional life. However, due to the growing complexity of underlying AI models and algorithms, it appears to be a ‘black box’, as the internal learning processes and the resulting models are not fully understood. This trade-off between performance and explainability can have a significant impact on individuals, businesses and society as a whole. Explainable artificial intelligence is capable of addressing this challenge by enhancing our understanding of the internal logic behind predictions made by complex machine learning models.

Keywords: explainable artificial intelligence, XAI, information systems, interpretability, machine learning, algorithm transparency.

Искусственный интеллект (ИИ) уже повсеместно присутствует как в профессиональной деятельности, так и в повседневной жизни общества: в виде разнообразных технологий, таких как обработка естественного языка или распознавание изображений, а также в различных сферах применения, включая электронную торговлю, финансы, здравоохранение, управление персоналом, государственный аппарат и транспорт. Присутствие ИИ будет расширяться, поскольку около 70% компаний по всему миру намерены внедрить его к 2030 году [1]. Однако, несмотря на все перспективы и широкие возможности, использование методов ИИ сопряжено с рядом проблем.

Одна из таких проблем заключается в необъяснимых характеристиках моделей ИИ, поскольку многие его алгоритмы представляют собой так называемые «черные ящики», внутреннее устройство которых остается неизвестным пользователю. Следствием подобного положения дел нередко становится системный кризис доверия к технологическим решениям на базе ИИ. Подобный скепсис, выступая в роли критического барьера, существенно блокирует процессы масштабирования и последующей интеграции ИИ-архитектур в реальный сектор. В данном контексте исследовательское поле информационных систем (ИС) обладает наибольшим потенциалом для концептуализации принципов объяснимости ИИ. Обусловлено это тем, что в рамках ИС технологические аспекты могут быть проанализированы полиаспектно — на стыке индивидуальных паттернов восприятия, организационных структур и общеинституциональных трансформаций. Комплекс деструктивных инженерно-математических, а также специфических архитектурных эффектов, детерминированных внедрением непрозрачных («черных ящиков») моделей в структуру современных информационных комплексов, систематизирован в таблице 1.

Таблица 1 Специфика технических вызовов и системных ограничений при интеграции непрозрачных моделей ИИ в контур ИС [2,3]

Технический фактор	Количественные параметры и численные тренды	Влияние на архитектуру и функционирование ИС
Критический рост структурной сложности алгоритмов	Размерность современных нейросетевых архитектур превышает 1011 параметров, создавая стохастические графы вычислений высокой плотности	Исключает возможность пошаговой трассировки логики выполнения программы, превращая ИС в функциональный «черный ящик»
Вычислительная избыточность апостериорного анализа	Математическое ожидание времени вычисления локальных объяснений (на основе векторов Шепли) увеличивается на 200–500% относительно времени прямого прохода	Создает недопустимые задержки в контурах обработки данных реального времени (высоконагруженные системы, обнаружение аномалий)
Неустойчивость объясняющих моделей к возмущениям	Малые состязательные искажения входного вектора (менее 1–2% шума) способны изменить структуру объяснения на 80% при	Снижает надежность систем, делает компоненты контроля уязвимыми для целенаправленных деструктивных воздействий

	неизменном итоговом прогнозе	
Отсутствие формальной верификации в программной инженерии	Более 70% применяемых методов оценивания прозрачности используют эмпирические, а не строгие математические критерии	Препятствует автоматизированному тестированию и отладке кода ИС, не позволяя гарантировать воспроизводимость результатов

Объяснимый искусственный интеллект (XAI) сегодня часто рассматривают как один из способов справиться с этими трудностями. По сути, XAI — это направление в ИИ, которое старается сделать модели машинного обучения более понятными для людей. Оно включает методы и приемы, которые помогают пользователям разбираться, откуда берутся результаты, и больше им доверять, когда их выдают алгоритмы машинного обучения [4].

Существуют также непосредственные операционные преимущества с технологической, организационной и экономической точек зрения. В частности, интерпретируемость модели представляет собой условие, позволяющее:

(i) оптимизировать и отлаживать модели;

(ii) выявлять неточные дискриминационные закономерности;

(iii) отслеживать непрерывные процессы обучения;

(iv) внедрять технологии, ориентированные на пользователя;

(v) обеспечивать подотчетность и ответственность;

(vi) применение моделей в качестве обучающих агентов для повышения уровня знаний и развития навыков.

Анализ существующей научно-практической базы позволяет выделить три ключевых направления, характеризующих текущее состояние применения XAI в ИС.

Во-первых, в практике проектирования современных систем реализуются различные типы интерпретируемости, направленные на обеспечение достаточного уровня понимания логики работы алгоритмов со стороны пользователей.

Во-вторых, осуществляется активный перенос методов объяснимости между разнородными прикладными областями. Например, подходы к обеспечению прозрачности, изначально разработанные для диалоговых систем, в настоящее время успешно адаптируются для применения в рамках электронных торговых площадок [5].

В-третьих, вместо тестирования в искусственно упрощенных симуляционных сценариях, текущая практика смещается в сторону проведения валидации в реальных условиях эксплуатации, что выражается в выполнении натурных и долгосрочных эмпирических исследований непосредственно в целевой операционной среде.

Кроме того, с практической точки зрения, область XAI охватывает два широких направления: интерпретируемость «ante hoc», когда модели создаются таким образом, чтобы быть понятными по своей сути, и анализ «post hoc», при котором объяснения генерируются после обучения сложной модели. Методы варьируются от моделей, прозрачных по своей сути (например, деревья решений и линейные модели), до подходов, не зависящих от конкретной модели, которые генерируют локальные или глобальные объяснения (например, карты атрибуции признаков, суррогатные модели и контрфактические примеры) [6].

В этой области также рассматриваются метрики качества объяснений, оценка с учетом интересов человека и взаимодействие с такими понятиями, как надежность, точность, устойчивость и причинно-следственная связь.

Возможности XAI для верификации нейросетей и генетических алгоритмов в ИС очевидны. Однако архитектурное усложнение моделей создает тупик для методов

объяснения. Речь идет о генеративных системах, а также алгоритмах распределенного и совместного обучения. Их производительность растет за счет масштаба. Миллиарды или даже триллионы параметров — норма для современного ИИ. Такой объем парализует стандартные инструменты XAI. Причина заключается в высокой размерности данных. Она экспоненциально увеличивает вычислительную сложность и блокирует извлечение обученных концепций.

В частности, одно из препятствий, связанных с последним моментом, заключается в полисемантической природе нейронов в генеративных моделях (то есть один нейрон может представлять несколько концепций), которая, как считается, возникает из суперпозиции множества независимых признаков. Исторически сложившийся фокус методологии XAI на решении конвенциональных задач классификации и регрессионного анализа предопределяет сегодня острую необходимость проектирования принципиально иных подходов, адаптированных к специфике генеративной парадигмы. Вектор технологического развития смещается в сторону самообучающихся и нейросетевых генеративных архитектур — в частности, вариационных автокодировщиков (VAE) и генеративно-состязательных сетей (GAN), — чья растущая популярность актуализирует новые вызовы. Сложность здесь кроется, к примеру, в процедурах верификации и анализа латентных (скрытых) пространств, формируемых данными моделями, а также в последующем синтезе интерпретируемых метаописаний для них, что до сих пор составляет самостоятельную фундаментальную проблему.

Ряд ключевых векторов и целевых ориентиров, определяющих долгосрочные контуры развития XAI, систематизирован и концептуально обобщен в таблице 2.

Таблица 2 Векторы девелопмента и релевантные целевые параметры эволюции систем XAI в структуре ИС [5,6]

Перспективное научно-техническое направление	Математический аппарат и архитектурные решения	Целевые технические критерии и метрики эффективности
Нейносимвольная интеграция	Сочетание непрерывных нейросетевых моделей с дискретными графами знаний, онтологиями и логическим выводом первого порядка	Снижение частоты генерации недостоверных логических связей более чем на 85%. Точность доказательства истинности вывода > 95%.
Причинно-следственный (каузальный) анализ	Замена корреляционных методов суррогатного моделирования структурными уравнениями причинности и контрфактуальным анализом	Снижение метрики погрешности объяснения до значений < 0,02. Повышение устойчивости к сдвигу распределения данных в 3,5 раза.
Оптимизация для распределенных и маломощных систем	Применение процедур прореживания весовых коэффициентов, квантования вычислительных графов до 8-битных целых чисел и дистилляции знаний	Время формирования объяснения на один запрос < 10 мс. Сокращение объема задействованной оперативной и видеопамяти на 75–80%

Формализованная стандартизация и автоматическое тестирование	Внедрение автоматизированных сред оценки на основе математических критериев монотонности признаков и инвариантности	Достижение согласованности объяснений > 98% при контролируемых модификациях входных сигналов. Полная автоматизация металгоритмического контроля.
Внутренне интерпретируемые архитектуры	Проектирование сетей с явным выделением разреженных признаков и встроенными механизмами взвешивания внимания	Снижение точности аппроксимации целевой функции по сравнению с базовой моделью не более чем на 1–1,5%. Отсутствие дополнительных затрат времени на генерацию объяснений (0% вычисли-тельных издержек)

Подведем итоги.

Функциональная основа систем объяснимого ИИ (XAI) устроена так, чтобы при определённых условиях давать проверяемое объяснение того, как разные модели принимают решения, и в основном сводится к тому, чтобы оценивать, какой причинный «вес» имеют ключевые признаки, влияющие на классификацию. Однако эволюция архитектур машинного обучения носит нелинейный характер. Нейросети постоянно усложняются, и из-за этого инструменты, которые помогают объяснять их работу, тоже приходится регулярно пересматривать и подстраивать; если считать, что методы можно просто оставить как есть, это может дорого обойтись. В итоге скрытые уязвимости и проблемы в конструкции — например, риски из-заложных корреляций — легко остаются незамеченными, если полагаться только на классический мониторинг.

Список литературы:

1. Аверкин А.Н. Объяснимый искусственный интеллект как часть искусственного интеллекта третьего поколения // Речевые технологии. 2023. № 1. С. 4-10.
2. Акулин А.А. О важности развития объяснимого искусственного интеллекта // Вестник Военного инновационного технополиса "Эра". 2025. Т. 6. № 1. С. 100-104.
3. Хаджиева Л.К. Методы объяснимого искусственного интеллекта (EXPLAINABLE AI): подходы, проблемы и перспективы // Экономика и управление: проблемы, решения. 2025. Т. 8. № 11 (164). С. 170-176.
4. Кожевникова Я.И. Искусственный интеллект и машинное обучение: объяснимый ИИ (XAI) и прозрачность алгоритмов // Аллея науки. 2025. Т. 1. № 4 (103). С. 824-827.
5. Косов П.И. Повышение достоверности объяснимого искусственного интеллекта посредством нечеткой логики и онтологии // Моделирование, оптимизация и информационные технологии. 2025. Т. 13. № 2 (49).
6. Сарин К.С. Методика построения интерпретируемых нечетких классификаторов для систем объяснимого искусственного интеллекта // Доклады Томского государственного университета систем управления и радиоэлектроники. 2025. Т. 28. № 2. С. 73-87.

References:

1. Averkin, A.N. Explainable artificial intelligence as part of third-generation artificial intelligence // Speech Technologies. 2023. No. 1. pp. 4–10.
2. Akulin, A.A. On the importance of developing explainable artificial intelligence // Bulletin of the 'Era' Military Innovation Technopolis. 2025. Vol. 6. No. 1. pp. 100–104.
3. Khadzhieva L.K. Methods of explainable artificial intelligence (EXPLAINABLE AI): approaches, problems and prospects // Economics and Management: Problems, Solutions. 2025. Vol. 8. No. 11 (164). pp. 170–176.
4. Kozhevnikova Y.I. Artificial Intelligence and Machine Learning: Explainable AI (XAI) and Algorithm Transparency // Alley of Science. 2025. Vol. 1. No. 4 (103). pp. 824–827.
5. Kosov P.I. Improving the reliability of explainable artificial intelligence through fuzzy logic and ontology // Modelling, Optimisation and Information Technologies. 2025. Vol. 13. No. 2 (49).
6. Sarin K.S. Methodology for constructing interpretable fuzzy classifiers for explainable artificial intelligence systems // Proceedings of Tomsk State University of Control Systems and Radio Electronics. 2025. No. 2. pp. 73–87.