Объяснительный ИИ 2.0: концептуальные сдвиги и новые требования

А. Н. Аверкин
ФИЦ ИУ РАН
averkin2003@inbox.ru

Е. Н. Волков
ФИЦ ИУ РАН
envolkoff@gmail.com

Аннотация. **В условиях стремительной интеграции искусственного интеллекта в критически важные сферы – от медицины и промышленности до законодательства и образования – возрастают требования к прозрачности и объяснимости интеллектуальных систем. Настоящий доклад посвящён анализу концептуального перехода от первой волны объяснимого ИИ (ОИИ 1.0), сосредоточенной на постфактум атрибутивных методах, к парадигме ОИИ 2.0, ориентированной на когнитивно осмысленные, человеко-центричные и контекстуально релевантные объяснения. В ходе исследования были поставлены четыре исследовательских вопроса, охватывающих различия между ОИИ 1.0 и ОИИ 2.0, требования прикладных сценариев, современные методы и вызовы развития. Методология основана на библиографическом анализе публикаций 2020–2024 годов и критическом осмыслении междисциплинарных источников. Работа выявляет, что успешное развитие ОИИ 2.0 требует выхода за пределы технической интерпретации и требует междисциплинарного подхода, включающего когнитивные науки, юриспруденцию и социотехнический анализ. Полученные результаты позволяют сформировать более целостное представление о требованиях к объяснениям и направлениях дальнейших исследований.**

Ключевые слова: объяснимый искусственный интеллект, интерпретируемость, человеко-центричный ИИ, концептуальные объяснения, XAI 2.0

I. Введение

Объяснительный искусственный интеллект (ОИИ, англ. Explainable Artificial Intelligence) в последние годы стал неотъемлемым направлением исследований в области машинного обучения, особенно в контексте использования сложных, «чёрных ящиков» — таких как глубокие нейронные сети — в высокорисковых и социально значимых приложениях. От медицины [1] до автономного транспорта [2], от промышленного производства [3] до судебных решений [4] — повсеместное внедрение ИИ-систем требует прозрачности и доверия к их предсказаниям и поведению. Именно эти цели и ставит перед собой ОИИ.

Традиционно под ОИИ понимался набор методов, обеспечивающих постфактум интерпретацию решений моделей, чаще всего посредством визуализации вкладов признаков (feature attribution). Такие подходы, как LIME [5], SHAP [6], Grad-CAM [7] и другие, стали стандартными инструментами в арсенале исследователей и практиков. Однако в условиях растущей сложности моделей, их масштабируемости и применения в задачах, требующих не только интерпретации, но и понимания, существующие ОИИ-инструменты демонстрируют ряд ограничений [8].

Становится очевидным, что XAI 1.0, основанный преимущественно на визуальной и статистической атрибуции, не может удовлетворительно отвечать на более глубокие вопросы: «почему модель приняла то или иное решение?», «какие концепты она распознала?», «можно ли доверять её выводу в конкретном контексте?» Это особенно критично в контексте Индустрии 5.0, где человек и ИИ должны взаимодействовать на уровне семантики, а не только данных [9].

На этом фоне формируется новая парадигма — ОИИ 2.0, ориентированная не только на объяснение модели, но и на интерактивную передачу знаний, адаптацию к пользователю, согласование с предметной областью и действие. Новые требования к объяснимости связаны с необходимостью объяснять не только что модель делает, но и как и почему, а также — что это означает в контексте предметной области. Это требует перехода от инженерной визуализации к когнитивно и семантически насыщенным объяснениям [10].

Современные методы, такие как Concept Relevance Propagation (CRP) и Relevance Maximization (RelMax) [11], представляют собой попытки объединить локальные и глобальные объяснения, позволяя ответить одновременно на вопросы, где и что использует модель при принятии решений. Такие подходы формируют основу ОИИ, направленного на интеграцию структурных, концептуальных и интерпретируемых объяснений на всех уровнях архитектуры модели.

Таким образом, цель данной — проанализировать основные концептуальные сдвиги, сопровождающие переход от ОИИ 1.0 к ОИИ 2.0, выявить новые требования, формируемые реальными сценариями применения, рассмотреть передовые методы объяснения и обозначить ключевые вызовы, стоящие перед исследовательским сообществом. В работе будут последовательно рассмотрены четыре исследовательских вопроса, направленных на раскрытие фундаментальных оснований и перспектив ОИИ нового поколения.

II. Методы и методология

Методологическая основа настоящего исследования сочетает элементы библиографического анализа и концептуального обобщения на основе сформулированных исследовательских вопросов. Целью являлось систематическое выявление отличий между парадигмами ОИИ 1.0 и ОИИ 2.0, а также формулировка новых требований к объяснимому искусственному интеллекту в условиях его внедрения в социально и промышленно значимые области.

В качестве первичного этапа был проведён поиск, по ключевым словам, в международных научных библиографических базах, включая Scopus, Google Scholar, arXiv. Ключевые словосочетания включали: «Explainable AI», «XAI 2.0», «interpretable models», «concept-level explanation», «CRP» (Concept Relevance Propagation), «causal explanations» и др. Особое внимание уделялось публикациям последних пяти лет, а также материалам, рекомендованным конференциями NeurIPS, ICLR, AAAI, IJCAI. Однако, в связи с тем, что первые работы по обозначенной теме появились лишь в 2024 году, полученная выборка оказалась небольшой.

На основе критического анализа отобранных источников были сформулированы четыре исследовательских вопроса, направленных на всестороннее рассмотрение темы: выявление отличий ОИИ 2.0, изучение требований со стороны практики, обзор современных методов и постановка открытых исследовательских проблем. Ответы на эти вопросы составляют основную структуру данного исследования.

III. Результаты и обсуждение

A. RQ1: В чём заключаются ключевые отличия между XAI 1.0 и XAI 2.0 как парадигмами объяснимого ИИ?

Переход от ОИИ 1.0 к ОИИ 2.0 представляет собой фундаментальное смещение научной и инженерной перспективы в области объяснимого искусственного интеллекта. Если ОИИ 1.0 можно охарактеризовать как набор методов постфактум интерпретации, направленных на объяснение решений моделей — в частности, глубоких нейронных сетей — то ОИИ 2.0 представляет собой трансдисциплинарную исследовательскую парадигму, ориентированную на формирование обоснованных, человеко-ориентированных и многоаспектных объяснений.

Методы ОИИ 1.0, включая такие как LIME, SHAP и Grad-CAM, фокусировались на атрибутивных объяснениях, выделяя важные признаки или области входных данных, повлиявшие на вывод модели. Эти подходы доминировали в задачах классификации и регрессии, особенно в изображениях и табличных данных. Однако, как подчёркивается в манифесте, данные объяснения зачастую страдают от неустойчивости, неоднозначности и когнитивной непрозрачности, особенно в высокоразмерных и нестандартных пространствах.

ОИИ 2.0 переопределяет объяснимость как процесс взаимодействия между ИИ и человеком, опирающийся на когнитивные, этические и социальные принципы. В этой парадигме объяснение не ограничивается локальным вкладом признаков, а включает концептуальные, причинные и адаптивные уровни анализа, ориентированные на интерпретацию, действия и доверие. Например, ОИИ 2.0 предполагает, что объяснение должно быть осмысленным, проверяемым и пригодным для применения в конкретных сценариях принятия решений.

Одним из важнейших отличий ОИИ 2.0 является ориентация на человеко-центричность и контекстуальность. Это означает, что объяснение должно учитывать индивидуальные особенности пользователя, задачи, уровень знаний и цели взаимодействия. Кроме того, ОИИ 2.0 требует поддержки многоаспектных объяснений, включающих правовые, этические и операционные характеристики модели — таких как надёжность, справедливость, устойчивость к искажениям и прозрачность.

Также XAI 2.0 предполагает расширение круга моделей, подлежащих объяснению: от классификаторов и регрессоров к генеративным моделям и крупным языковым моделям (LLMs), для которых классические методы ОИИ оказываются неэффективными. В этой связи особое внимание уделяется методам механистической интерпретации, концептуального анализа и онтологической сопоставимости, а также интердисциплинарной кооперации как обязательному элементу в построении объясняемых ИИ-систем.

Таким образом, ОИИ 1.0 и ОИИ 2.0 различаются не только по используемым методам, но и по своим целям, объектам объяснения, ожидаемым результатам и научным основаниям. ОИИ 1.0 акцентирует внимание на модели как объекте интерпретации; ОИИ 2.0 — на человеке как центральном участнике совместного когнитивного процесса. [11]

B. RQ2: Какие новые требования к объяснительности выдвигают современные прикладные сценарии?

С распространением искусственного интеллекта в критически важных и социально значимых сферах, таких как медицина, финансы, образование и промышленность, существенно возросли требования к характеру и качеству объяснений, предоставляемых ИИ-системами. Согласно манифесту XAI 2.0, в прикладных сценариях объяснения должны не только раскрывать внутреннюю логику модели, но и быть понятными, проверяемыми, адаптивными и соответствовать нормативным требованиям конкретной области применения.

В медицине, где решения ИИ напрямую влияют на здоровье и жизнь пациентов, ключевыми требованиями становятся доверие, прозрачность и поддержка принятия решений. Врачи и клиницисты используют выводы ИИ-систем в диагностике, назначении терапии и общении с пациентами. При этом необходимо, чтобы объяснения были не только технически корректными, но и этически приемлемыми, юридически обоснованными и когнитивно доступными. Например, при использовании ИИ в распознавании хронических заболеваний или COVID-19, объяснения, основанные на методах LIME, SHAP и Anchors, позволяли клиницистам верифицировать предсказания и повышать доверие к системе.

В сфере финансов объяснения подчиняются строгим регуляторным нормам. Согласно требованиям GDPR и законодательным актам США (например, Equal Credit Opportunity Act), организации обязаны предоставлять чёткие и обоснованные объяснения отказов в кредитовании или финансовых решениях. Это порождает запрос на стабильные и юридически корректные объяснения, которые можно проследить и воспроизвести. Нарушение этих требований чревато штрафами и потерей доверия клиентов.

В промышленности, в частности в контексте Индустрии 5.0, XAI должен обеспечивать операционную

применимость, позволяя инженерам и операторам интерпретировать поведение модели в реальном времени. Например, в задачах предиктивного обслуживания оборудования объяснение должно быть связано с причинно-следственными связями, выявляющими факторы риска отказов, а не только визуализировать данные. Здесь особое значение приобретают контрфактические и причинные объяснения, способные поддерживать действия пользователя.

В образовании и рекомендательных системах особое внимание уделяется персонализированным объяснениям, адаптированным под уровень подготовки учащихся. ИИ-инструменты, такие как обучающие ассистенты или генераторы заданий, должны сопровождаться объяснениями, способствующими развитию метакогнитивных навыков, включая саморефлексию и саморегуляцию.

Объединяя эти сценарии, XAI 2.0 должен удовлетворять четырём универсальным требованиям:

- контекстная релевантность,
- многоаспектность (в том числе правовая и этическая),
- доказуемость и воспроизводимость,
- адаптивность к целевой аудитории.

Таким образом, прикладные сценарии выдвигают перед XAI требования, выходящие за пределы технической интерпретации модели, превращая объяснение в социально и юридически значимую форму коммуникации, требующую междисциплинарного подхода к её проектированию и оценке.

C. RQ3: "Какие методы и подходы XAI 2.0 на сегодняшний день демонстрируют наибольший потенциал с точки зрения интерпретируемости, доверия и адаптивности?"

Современный этап развития XAI характеризуется значительным разнообразием методов, направленных на обеспечение интерпретируемости, надёжности и адаптивности ИИ-систем. Наиболее перспективные подходы XAI 2.0 выходят за рамки простых атрибутивных карт и стремятся к созданию человеко-понятных объяснений, обладающих когнитивной связностью, концептуальной выразительностью и правдоподобием.

Одной из ключевых групп методов остаются атрибутивные подходы (LIME, SHAP, Grad-CAM, LRP), позволяющие выделить значимые входные признаки. Несмотря на популярность, они демонстрируют ограниченную интерпретируемость в сложных и многомерных задачах, где необходимы не только локальные визуализации, но и высокоуровневое понимание.

На смену им приходят концептуальные и нейро-семантические методы, среди которых особое место занимают Concept Bottleneck Models, ProtoPNet, Concept Activation Vectors и Concept Relevance Propagation (CRP). Эти подходы позволяют отображать предсказания в терминах человеко-понятных концептов и создавать «глобально-локальные» (“glocal”) объяснения, что существенно облегчает когнитивную интерпретацию модели. Кроме того, они повышают доверие за счёт соответствия объяснений логике предметной области.

Другим важным направлением являются аргументативные и символические методы объяснения, в частности – вычислительная аргументация и правила вывода. Такие подходы обеспечивают прозрачное представление логики принятия решений, позволяют верифицировать цепочки рассуждений, и особенно актуальны в правовых и этически нагруженных сценариях.

В контексте генеративных моделей и больших языковых моделей (LLMs) наиболее перспективными являются методы механистической интерпретации. Они направлены на декомпозицию поведения модели на составные алгоритмы и выявление скрытых представлений, отвечающих за принятие решений. Механистический анализ масштабируем, и может быть дополнен информационно-геометрическими и причинными моделями для оценки стабильности и достоверности объяснений.

Наконец, особое внимание в XAI 2.0 уделяется адаптивности объяснений. Методы, учитывающие контекст пользователя, уровень подготовки, цели и задачи, становятся приоритетными. Это достигается за счёт персонализированных объяснений, онтологического сопоставления и возможности интерактивной настройки глубины и формы интерпретации. Использование пользовательских моделей, а также включение элементов обратной связи через Human-in-the-Loop XAI расширяют возможности применения ИИ в критически важных областях.

Таким образом, методы XAI 2.0 представляют собой синтез статистических, концептуальных, логических и когнитивных механизмов, ориентированных не только на понимание ИИ, но и на формирование надежных, контекстуальных и практически полезных объяснений.

D. RQ4: Каковы основные вызовы и открытые проблемы в развитии XAI 2.0, требующие междисциплинарного подхода?

Несмотря на значительный прогресс в разработке методов объяснимого ИИ, XAI 2.0 сталкивается с рядом фундаментальных вызовов, решение которых невозможно без привлечения знаний и методов из различных дисциплин — философии, психологии, юриспруденции, лингвистики, когнитивных наук и социологии. Эти вызовы касаются не только технической стороны объяснений, но и вопросов понимания, доверия, нормативного регулирования и адаптивности.

Во-первых, XAI 2.0 нуждается в унификации понятийного аппарата. В настоящее время в литературе отсутствует единое понимание таких терминов, как explainability, interpretability, transparency, trustworthiness. Это порождает концептуальную неоднозначность и затрудняет междисциплинарную коммуникацию. Решение данной проблемы требует создания глоссариев, согласования терминов и анализа их использования в различных дисциплинах.

Во-вторых, существует проблема оценки качества объяснений. На сегодняшний день отсутствует

общепринятый набор метрик или стандартов, позволяющих сравнивать XAI-методы между собой. Особое значение здесь имеет вовлечение пользователей — текущие подходы редко включают человеко-ориентированные исследования, что снижает достоверность и применимость выводов. Для этого требуется интеграция методов юзабилити-тестирования, эмпирических исследований из HCI и психологии восприятия.

Третья ключевая проблема — создание объяснений для новых типов ИИ, включая генеративные модели (GAN, VAE), крупные языковые модели (LLM) и концепто-ориентированные архитектуры. Для этих систем традиционные XAI-подходы оказываются неэффективными. Возникает необходимость в разработке новых методов, таких как механистическая интерпретация, основанная на принципах причинности и декомпозиции поведения модели.

Следующий вызов — персонализация и адаптивность объяснений. Пользователи различных категорий (регуляторы, инженеры, клиницисты, граждане) предъявляют различные требования к форме, глубине и языку объяснений. XAI 2.0 требует проектирования систем, способных адаптировать объяснения под когнитивные особенности и цели пользователя.

Наконец, XAI 2.0 сталкивается с вызовом многоаспектности объяснимости, где объяснение должно учитывать технические, правовые, этические и когнитивные компоненты одновременно. Это требует создания мультифасеточных объяснений, которые объединяют данные о надёжности, справедливости, устойчивости и соответствию нормативным требованиям.

Таким образом, ключевые вызовы XAI 2.0 — это не просто проблемы алгоритмического характера, а комплексные междисциплинарные задачи. Их решение требует не только синтеза технических и гуманитарных подходов, но и институциональной поддержки междисциплинарных исследований, развития стандартов и активного взаимодействия между академией, индустрией и обществом.

IV. Заключение

Исследование, представленное в данном докладе, позволяет сделать ряд содержательных выводов о направлении эволюции объяснимого искусственного интеллекта в контексте перехода к парадигме XAI 2.0. Анализ показал, что ОИИ 1.0, основанный на постфактум визуализации признаков, недостаточен для удовлетворения потребностей реальных прикладных сценариев, требующих доверия, адаптивности и правовой обоснованности решений. Напротив, XAI 2.0 ориентирован на человека как ключевого участника взаимодействия с ИИ и предполагает не только объяснение решения модели, но и её когнитивную и контекстуальную верификацию.

Прикладные сценарии, охватывающие медицину, промышленность, финансовую сферу и образование, формируют специфические требования к объяснительности — от юридической воспроизводимости до поддержания действий в реальном времени. Это обуславливает необходимость разработки объяснений, сочетающих интерпретируемость, доказуемость и персонализацию. Современные методы XAI 2.0, включая концепто-ориентированные, механистические и аргументативные подходы, демонстрируют высокий потенциал в обеспечении доверительных взаимодействий между ИИ и пользователем.

Одновременно с этим, в работе выявлен ряд открытых проблем, решение которых невозможно без междисциплинарной кооперации. К ним относятся стандартизация терминологии, разработка метрик оценки качества объяснений, обеспечение персонализированного взаимодействия, а также адаптация объяснимости к новым типам моделей, включая генеративные и языковые архитектуры.

В заключение следует отметить, что развитие XAI 2.0 требует не только технологического прогресса, но и синтеза знаний из гуманитарных и социальных наук. Только в этом случае объяснимый ИИ сможет стать основой доверительных, ответственных и эффективных интеллектуальных систем будущего.

Список литературы

[1] Sadeghi Z., Alizadehsani R., Cifci M.A. et al. A review of Explainable Artificial Intelligence in healthcare // Computers and Electrical Engineering. 2024. Vol. 118. P. 109370. DOI: 10.1016/j.compeleceng.2024.109370.

[2] Kuznietsov A., Gyevnar B., Wang C. et al. Explainable AI for safe and trustworthy autonomous driving: a systematic review //IEEE Transactions on Intelligent Transportation Systems. 2024. DOI: 10.1109/TITS.2024.3474469.

[3] Alexander Z., Chau D.H., Saldaña C. An interrogative survey of explainable AI in manufacturing //IEEE Transactions on Industrial Informatics. 2024. DOI: 10.1109/TII.2024.3361489.

[4] Richmond K. M. G. et al. Explainable AI and law: An evidential survey //Digital Society. 2024. Vol. 3. No. 1. P. 1. DOI: 10.1007/s44206-023-00081-z.

[5] Ribeiro M.T., Singh S., Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. P. 1135–1144. DOI: 10.1145/2939672.2939778.

[6] Lundberg S.M., Lee S.-I. A Unified Approach to Interpreting Model Predictions // Advances in Neural Information Processing Systems (NeurIPS). 2017. Vol. 30.

[7] Selvaraju R.R., Cogswell M., Das A. et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization // Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2017. P. 618–626. DOI: 10.1109/ICCV.2017.74.

[8] Rudin C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead // Nature Machine Intelligence. 2019. Vol. 1, No. 5. P. 206–215. DOI: 10.1038/s42256-019-0048-x.

[9] Bobek S., Nowaczyk S., Gama J. et al. Why Industry 5.0 Needs XAI 2.0? // Proceedings of the 1st World Conference on eXplainable Artificial Intelligence. 2023. CEUR Workshop Proceedings, Vol. 3432.

[10] Achtibat R., Dreyer M., Eisenbraun I. et al. From Attribution Maps to Human-Understandable Explanations through Concept Relevance Propagation // Nature Machine Intelligence. 2023. Vol. 5, No. 12. P. 1006–1019. DOI: 10.1038/s42256-023-00711-8.

[11] Longo L., Brcic M., Cabitza F. et al. Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions //Information Fusion. 2024. Vol. 106. P. 102301. DOI: 10.1016/j.inffus.2024.102301.