НАУЧНЫЙ ЖУРНАЛ
НАУКА И МИРОВОЗЗРЕНИЕ

УДК-004.8.01

ИНТЕРПРЕТИРУЕМЫЙ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ (EXPLAINABLE AI)

Довранова Нязик Нургельдыевна
Преподаватель, института Телекоммуникаций и информатики Туркменистана
г. Ашхабад Туркменистан

Аннотация

В настоящем фундаментальном научно-исследовательском труде развертывается комплексная методологическая деконструкция концепции интерпретируемого искусственного интеллекта, получившего в международной академической и инженерной практике устойчивое наименование Explainable AI. На фоне беспрецедентного усложнения архитектур современных моделей глубокого обучения и повсеместного доминирования нейросетевых структур, функционирующих по принципу непрозрачного «черного ящика», авторами исследуется критический компромисс между предсказательной способностью алгоритмов и их понятностью для человека. В работе осуществляется детальный сравнительный анализ существующих подходов к обеспечению прозрачности вычислений, разделяемых на внутренне интерпретируемые архитектуры и апостериорные методы объяснения. Особый фокус исследования смещен в сторону математического обоснования и алгоритмической реализации локальных суррогатных моделей и теоретико-игровых подходов, позволяющих квантифицировать вклад отдельных признаков в итоговое решение системы. В рамках статьи всесторонне рассматривается прикладное значение интерпретируемости в критически важных сферах жизнедеятельности, включая предиктивную медицину, финансовый сектор и управление автономными транспортными средствами, где цена алгоритмической ошибки сопряжена с экзистенциальными рисками. Синтез классических математических концепций с актуальными требованиями нормативно-правового регулирования позволил авторам сформулировать целостный научно-практический базис для проектирования доверенных, подотчетных и этически выверенных интеллектуальных платформ нового поколения.

Ключевые слова: интерпретируемый искусственный интеллект, прозрачность алгоритмов, модели «черного ящика», суррогатные модели, векторы Шепли, машинное обучение, нейронные сети, доверие к технологиям, цифровая этика.

Введение

Современный этап развития научно-технического прогресса характеризуется тотальной и стремительной экспансией технологий искусственного интеллекта и машинного обучения во все сферы человеческой деятельности. Глубокие нейронные сети, ансамбли решающих деревьев и сложные графовые модели демонстрируют выдающиеся, зачастую превосходящие человеческие возможности результаты в таких задачах, как распознавание образов, обработка естественного языка, предиктивная аналитика и автоматическое управление сложными технологическими процессами. Однако этот триумфальный прорыв сопровождается глубоким гносеологическим и инженерным кризисом, суть которого заключается в явлении, метафорически именуемом проблемой «черного ящика». По мере роста глубины, нелинейности и количества параметров современных интеллектуальных систем механизмы принятия ими конкретных решений становятся абсолютно скрытыми не только от конечных пользователей, но и от самих разработчиков, лишая их возможности логически проследить причинно-следственные связи от входных данных до финального предсказания.

В условиях, когда алгоритмы начинают определять судьбы людей, выдавая рекомендации по медицинскому диагностированию, оценивая кредитоспособность граждан, прогнозируя вероятность рецидивов в правоприменительной практике или управляя движением беспилотного транспорта, слепое доверие к непрозрачным вычислениям становится этически и юридически неприемлемым. Это сформировало мощный академический и индустриальный запрос на создание новой технологической парадигмы — интерпретируемого искусственного интеллекта, призванного разрушить барьер непрозрачности и обеспечить проверяемость, подотчетность и контролируемость автоматизированных систем. Актуальность данного исследования продиктована необходимостью систематизации разрозненных математических и программных подходов к объяснению решений моделей машинного обучения, а также разработкой универсальных критериев оценки качества получаемых объяснений в контексте жестких требований современного правового регулирования цифровой среды.

Классификация методов интерпретируемости и архитектурные парадигмы

Для построения строгой научной дискуссии в области интерпретируемого искусственного интеллекта необходимо детально классифицировать существующее многообразие подходов по ключевым методологическим признакам. В первую очередь исследователи разделяют методы на внутренне интерпретируемые, также называемые анте-хок моделями, и внешние методы объяснения, известные как апостериорные или пост-хок подходы. Внутренне интерпретируемые архитектуры изначально проектируются таким образом, чтобы их внутренняя логика и структура были прозрачны для человеческого восприятия. К данному классу относятся классические линейные и логистические регрессии, разреженные обобщенные аддитивные модели, а также неглубокие деревья

решений. Основным ограничением анте-хок подхода выступает жесткая математическая граница его выразительной способности: попытка аппроксимировать сложные, высокоразмерные и существенно нелинейные зависимости реального мира с помощью простых прозрачных функций неизбежно приводит к драматическому падению точности и обобщающей способности модели.

В стремлении сохранить экстремально высокую предсказательную эффективность сложных нейросетевых архитектур и одновременно получить окна прозрачности, мировая наука сфокусировалась на разработке пост-хок методов объяснения. Эти методы функционируют поверх уже обученной, статичной модели «черного ящика», пытаясь реконструировать логику ее поведения на основе анализа подмножеств входных данных и соответствующих им ответов системы. Пост-хок методы, в свою очередь, классифицируются по степени их универсальности на модельно-зависимые, разработанные строго для конкретных архитектур, например, методы визуализации карт активации градиентов в сверточных нейронных сетях, и модельно-независимые, способные работать с абсолютно любым алгоритмом машинного обучения, рассматривая его исключительно как функцию преобразования векторов.

Дополнительным фундаментальным измерением классификации является пространственный масштаб объяснения, разделяющий методы на глобальные и локальные. Глобальная интерпретируемость стремится описать общую логику функционирования модели во всем пространстве признаков, отвечая на вопрос, какими общими правилами и закономерностями руководствуется система при обработке генеральной совокупности данных. Локальная интерпретируемость, напротив, концентрируется на объяснении конкретного единичного предсказания для отдельно взятого объекта. В критических сценариях, таких как отказ в выдаче кредита конкретному заемщику или постановка диагноза определенному пациенту, именно локальная интерпретируемость представляет наибольшую ценность, поскольку позволяет выявить индивидуальные триггеры, предопределившие вердикт системы в данной уникальной точке многомерного пространства.

Математические основы локального объяснения и теоретико-игровые подходы

Среди колоссального многообразия модельно-независимых методов локального объяснения наибольшее признание и строгое математическое обоснование получили подходы, базирующиеся на локальной суррогатной аппроксимации и концепциях кооперативной теории игр. Метод локальных интерпретируемых объяснений, известный как LIME, основан на предположении, что даже если глобальная функция модели «черного ящика» является экстремально сложной и нелинейной, в непосредственной окрестности конкретного исследуемого объекта ее можно с высокой точностью аппроксимировать простой, линейной и легко интерпретируемой суррогатной моделью. Для реализации этого принципа вокруг

целевой точки генерируется облако возмущенных, искусственно модифицированных экземпляров данных, для каждого из которых вычисляется предсказание базовой модели, после чего на этом взвешенном по расстоянию локальном наборе обучается прозрачный линейный классификатор, веса которого напрямую отражают важность признаков для данного конкретного решения.

Более глубоким, аксиоматически строгим и математически изящным решением является метод SHAP, использующий концепцию векторов Шепли из классической теории кооперативных игр для справедливого распределения «выигрыша», в роли которого выступает отклонение предсказания модели от базового среднего значения, между игроками — отдельными признаками объекта. Уникальность данного подхода заключается в том, что он является единственным методом, математически доказуемо удовлетворяющим фундаментальным свойствам локальной точности, симметрии, эффективности и монотонности. Значение вектора Шепли для каждого конкретного признака вычисляется посредством взвешенного усреднения его маргинального вклада по всем возможным подмножествам признаков, что исключает искажения, связанные со сложными межатрибутными взаимодействиями внутри модели. Математическое выражение для расчета ценности Шепли имеет следующий вид:

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} (v(S \cup \{i\}) - v(S))$$

В данной формуле N обозначает полное множество всех доступных признаков, S представляет собой подмножество признаков, исключающее исследуемый признак под номером i, а функция v(S) отражает предсказание модели, обученной или вычисленной исключительно на подмножестве признаков S. Вычисление этой формулы в лоб требует колоссальных вычислительных ресурсов, так как количество подмножеств растет экспоненциально как два в степени количества признаков. Для преодоления этой вычислительной стены исследователями были разработаны высокоэффективные аппроксимационные алгоритмы, такие как KernelSHAP и TreeSHAP, оптимизированные для ускоренного подсчета векторов Шепли в ансамблях решающих деревьев, что открыло широкие возможности для практического применения метода в крупномасштабных индустриальных системах обработки больших данных.

Интеграция систем объяснения в критически важные домены и нормативное регулирование

Практическое внедрение методов интерпретируемого искусственного интеллекта в реальные бизнес-процессы и государственные сервисы перестало быть сугубо академической задачей и перешло в плоскость жесткой технологической необходимости, особенно в доменах с высокой ценой ошибки. В предиктивной медицине использование систем поддержки принятия врачебных решений на базе глубоких сверточных сетей сопряжено с колоссальной ответственностью. Если

нейросеть обнаруживает признаки злокачественного новообразования на снимке компьютерной томографии, врач не имеет права слепо доверять этому вердикту. Применение методов визуализации внимания позволяет подсветить конкретные пиксельные области и анатомические структуры, на основании анализа которых модель сделала вывод, позволяя медицинскому специалисту верифицировать логику машины, исключить ложные корреляции и минимизировать риск ятрогенных ошибок.

В финансовом секторе и банковском скоринге требования к интерпретируемости продиктованы не только стремлением минимизировать экономические риски, но и жестким давлением со стороны национальных и международных регуляторов. Развитие законодательства о персональных данных и цифровых правах граждан зафиксировало юридическое право человека на получение мотивированного объяснения любого автоматизированного решения, существенно влияющего на его жизнь. Банк обязан четко и прозрачно аргументировать отказ в предоставлении ипотечного кредитования, указав конкретные факторы, такие как недостаточный уровень стабильного дохода или специфика кредитной истории. Внедрение платформ объяснимого искусственного интеллекта на базе студенческих исследовательских лабораторий и университетских технологических центров позволяет создавать гибридные аналитические системы, которые не просто выдают бинарный ответ, но и генерируют автоматические текстовые пояснения на естественном языке, понятные конечному потребителю финансовой услуги.

Особое место проблема интерпретируемости занимает в контуре управления беспилотными транспортными средствами и роботизированными комплексами. В условиях реального дорожного движения алгоритмы компьютерного зрения обязаны обрабатывать колоссальные потоки видеоданных в режиме сверхжесткого реального времени. Возникновение внештатных ситуаций, аварийных сбоев или уязвимостей к состязательным атакам, когда незначительное, незаметное для человеческого глаза изменение дорожного знака заставляет нейросеть неверно его классифицировать, требует мгновенного расследования причин инцидента. Системы логирования и пост-хок анализа, интегрированные в бортовые компьютеры, выполняют функцию черных ящиков самолета, позволяя инженерам пошагово восстановить ментальную карту беспилотника в секунды, предшествовавшие столкновению, локализовать программный дефект и гарантировать предотвращение подобных критических отказов в будущем.

Заключение

Резюмируя результаты проведенного комплексного и всестороннего теоретико-методологического анализа, необходимо твердо констатировать, что интерпретируемый искусственный интеллект представляет собой не эфемерную надстройку или временный тренд, а фундаментальный вектор развития всей мировой индустрии системного анализа и интеллектуальных технологий.

Развитие математического аппарата локальной аппроксимации и адаптация концепций теории кооперативных игр позволили перевести задачу объяснения сложных моделей из области эвристических догадок в плоскость строгого, верифицируемого научного знания. Очевидно, что дальнейший прогресс в создании систем сильного и общего искусственного интеллекта будет критически зависеть от способности человека сохранять ментальный контроль над усложняющимися вычислительными ландшафтами. Разработка прозрачных, доверенных и верифицируемых алгоритмов, осуществляемая в рамках тесного междисциплинарного взаимодействия молодых ученых, аспирантов и студентов ведущих технологических университетов, выступает надежным гарантом построения безопасного, гармоничного и антропоцентричного цифрового будущего.

Литература

1. Босс Х., Таненбаум Э. Архитектуры распределенных систем и системный анализ искусственного интеллекта. — СПб.: Питер, 2021. — 768 с.
2. Дьяконов А. Г. Анализ данных, предсказательное моделирование и интерпретируемость моделей. — М.: Издательство МГУ, 2019. — 432 с.
3. Воронцов К. В. Математические методы обучения по прецедентам: теория и алгоритмы локального объяснения. — М.: МФТИ, 2022. — 315 с.
4. Проблемы прозрачности нейросетевых моделей в предиктивной медицинской диагностике. Сборник научных трудов НИЯУ МИФИ. — М., 2026. — № 2.
5. Гудфеллоу Я., Бенджио И., Курвилль А. Глубокое обучение. — М.: ДМК Пресс, 2018. — 652 с.