Объяснимый искусственный интеллект (Explainable artificial intelligence)

> Синонимы: Интерпретируемый искусственный интеллект, Объяснимое машинное обучение, Explainable AI, XAI, Interpretable AI, Explainable machine learning, XML

Объяснимый искусственный интеллект (XAI) — направление в ИИ, связанное с разработкой и применением технологий, методов и процессов, позволяющих сделать сложные модели машинного обучения и результаты их работы понятными и прозрачными для человека. Модель не только формирует выводы и решения, но и объясняет, почему они были получены.

В рамках XAI стремятся создавать системы и модели ИИ, способные объяснять свои действия и принимать решения понятным для людей образом, чтобы повысить доверие к ним. Объяснимый ИИ используется для описания алгоритмов, а также их ожидаемых последствий и возможных отклонений. Для этого используются методы визуализации, более простые алгоритмы, а также интерактивные интерфейсы с подсказками.

Иными словами, когда человек использует ИИ для принятия решения и получает результат, у него, естественно, возникает вопрос: почему именно такой ответ? При этом сама модель, ее структура и параметры обычно не позволяют объяснить это решение, поэтому для внешнего наблюдателя она выглядит как «черный ящик». Пока ответ на этот вопрос не получен, доверие к результатам модели и принятым на их основе решениям снижается.

Методы XAI позволяют сформировать представление о поведении модели, оценить значимость признаков, обнаружить ее предвзятость и потенциальные проблемные точки. Главной целью при этом является преодоление проблемы «черного ящика», которая заключается в том, что зачастую поведение модели и выдаваемые ею результаты не могут понять и объяснить даже сами ее создатели.

Основными результатами применения технологий XAI являются повышение доверия со стороны пользователей к результатам работы моделей машинного обучения, возможность оценки их корректности и осознанного принятия решений, формирование пользовательского опыта. Кроме этого, объяснимость позволяет подтвердить имеющиеся знания, оспорить их или сформировать новые гипотезы.

Использование XAI также позволяет эффективнее выявлять и устранять ошибки ИИ, особенно в тех областях, где они могут привести к серьезным негативным последствиям, что способствует повышению безопасности и соблюдению этических требований.

Необходимость объяснения, почему с помощью ИИ было принято определенное решение, закреплена на законодательном уровне в ряде стран.

Чем выше объяснимость ИИ и доверие к нему, тем активнее он внедряется. Поэтому компании, разрабатывающие ИИ-решения, заинтересованы в их объяснимости, так как это помогает привлекать больше клиентов и увеличивать доходы. Именно этим объясняется рост инвестиций в исследования и разработки в области XAI.

В основе XAI лежат три базовых концепции:

- **прозрачность** — возможность понимать и описывать, как из данных извлекаются признаки и как формируются целевые значения при обучении модели;
- **интерпретируемость** — возможность объяснить, как модель принимает решения, в понятной для человека форме;
- **объяснимость** — возможность понять, какие признаки сильнее всего повлияли на результат работы модели.

В контексте XAI все модели могут быть разделены на две группы:

- **прозрачные (интерпретируемые)** — модели, внутренняя структура и логика работы которых по своей природе проста и понятна человеку. К ним относятся линейная и логистическая регрессия, деревья решений и другие модели на основе правил, кластеризация методом k-средних, классификаторы на основе метода k-ближайших соседей. Например, последний легко объясняет, что конкретный объект относится к данному классу, потому что к нему же относится большинство ближайших объектов.
- **непрозрачные** — модели, структура и логика работы которых скрыты внутри или слишком сложны для понимания. Типичным примером являются нейронные сети, их веса практически не поддаются интерпретации, а также вероятностные модели кластеризации.

Некоторые подразумевают под XAI генеративный ИИ, который способен формировать контент в максимально понятном и доступном виде. Однако это не совсем так. Задача генеративного ИИ — создавать новый контент, который может и не содержать логики, объясняющей, почему он именно такой. В то же время XAI фокусируется на обеспечении прозрачности внутренних процессов модели с целью дать понимание, почему был получен данный результат.

Методы, используемые XAI, можно разделить на два направления:

- использование простых и понятных (самообъясняемых) моделей;
- использование сложных моделей с последующим объяснением их работы с помощью дополнительных методов.

Очевидно, что первое направление является более перспективным, поскольку генерация объясняющей информации требует дополнительных затрат.

В настоящее время разработано достаточно много подходов для реализации XAI. Наиболее популярными являются:

- **локальные интерпретируемые модельно-независимые объяснения (LIME — Local Interpretable Model-agnostic Explanations)** — метод, который объясняет работу

сложной модели, приближая ее поведение вблизи конкретного примера более простой и понятной моделью.
- **аддитивные объяснения Шэпли (SHAP — Shapley Additive exPlanations)** — метод, который показывает, какой вклад каждый признак вносит в конкретное предсказание модели.
- **аддитивная глобальная важность Шэпли (SAGE — Shapley Additive Global Importance)** — метод, который оценивает, насколько в целом модель зависит от каждого признака.

Несмотря на достаточно интенсивное развитие технологий XAI, у них есть ряд ограничений. Одной из проблем, с которыми сталкиваются исследования, является отсутствие консенсуса в отношении определений ряда ключевых понятий. В частности, точные определения объяснимого ИИ различаются в разных работах и контекстах.

Некоторые исследователи используют термины «объяснимость» и «интерпретируемость» как взаимозаменяемые для обозначения концепции, позволяющей сделать модели и их результаты более понятными человеку. Другие проводят различия между ними. Например, одни считают, что «объяснимость» относится к априорным объяснениям, а «интерпретируемость» — к апостериорным.

Другой проблемой является отсутствие практических рекомендаций по выбору, внедрению и тестированию объяснений для конкретных случаев. Хотя показано, что объяснения улучшают понимание систем машинного обучения для многих аудиторий, их способность укреплять доверие среди неспециалистов в области ИИ вызывает споры. Наибольшую перспективность показали интерактивные объяснения в форме вопросов и ответов.

Еще одним предметом дискуссий является поиск компромисса между точностью и объяснимостью моделей. Действительно, повышение объяснимости в большинстве случаев связано с упрощением модели, что, как правило, ведет к уменьшению точности. Поэтому модели, к которым применены методы XAI для повышения прозрачности, должны подвергаться усиленной проверке перед развертыванием.

Кроме этого, проблемами являются высокая сложность разработки систем XAI, а также темпы развития ИИ, за которыми разработка новых методов повышения объяснимости просто не успевает.