УДК 004.89

ПОВЫШЕНИЕ ДОВЕРИЯ ЧЕРЕЗ КОНТРОЛЬ ИНТЕРПРЕТИРУЕМОСТИ НА РАЗНЫХ ЭТАПАХ ИСПОЛЬЗОВАНИЯ МОДЕЛИ

Воробьев Д.С., Оператор научной роты, III курс
Военная академия связи имени Маршала Советского Союза С. М. Буденного,
г. Санкт-Петербург

Повышение доверия к интеллектуальным системам, основанным на предобученных моделях, представляет собой одну из центральных задач современной инженерии ИИ. Среди множества направлений, формирующих доверие – включая устойчивость, безопасность, воспроизводимость и правовую совместимость – интерпретируемость занимает особое положение. Она не просто позволяет объяснить поведение модели, но обеспечивает возможность контроля, верификации и адаптивной коррекции её выводов в условиях неопределённости.

Особенно важно то, что интерпретируемость становится механизмом поддержки доверия на всех этапах жизненного цикла модели: от проектирования и предобучения до эксплуатации и постдеплойного анализа. В данной работе рассматриваются принципы и средства поэтапного контроля интерпретируемости как стратегии повышения доверия к предобученным моделям в реальных задачах [1].

На этапе проектирования и предобучения интерпретируемость может быть заложена как архитектурный и процедурный приоритет. Это означает выбор таких моделей и техник обучения, которые потенциально допускают объяснение внутренних механизмов принятия решений. Например, для языковых моделей это может быть использование модулей внимания (attention), для визуальных – архитектур с пространственно локализуемыми фильтрами, для табличных данных – обобщённые аддитивные модели (GAM) или модели с ограничением взаимодействий между признаками. В некоторых случаях разумным становится выбор в пользу интерпретируемых моделей первого уровня, даже если они уступают в точности сложным ансамблям или глубоким нейросетям. На этом же этапе могут применяться методы feature attribution analysis, направленные на выявление признаков, оказывающих системное влияние на поведение модели, что позволяет исключить некорректные или опасные зависимости ещё до запуска модели в эксплуатацию.

Во время дообучения и адаптации модели под конкретную задачу контроль интерпретируемости помогает отслеживать, как изменяется логика вывода под влиянием новых данных. Особенно это актуально для моделей, работающих в условиях слабой супервизии, активного обучения или самообучения, где нет полной уверенности в качестве меток. Интерпретируемость здесь играет роль контрольной линзы, позволяя исследовать, какие элементы входа оказывают наибольшее влияние, какие связи между признаками усиливаются, и появляются ли новые зависимости, отличающиеся от обучающих паттернов. На этом этапе могут быть также выявлены спонтанные предвзятости, возникшие из-за локальных сдвигов в тренировочных выборках, что особенно важно в задачах с социальной или медицинской чувствительностью [2].

На этапе деплоя и эксплуатации интерпретируемость служит инструментом онлайн-мониторинга поведения модели. Современные системы мониторинга, такие как Fiddler, Arize, TruEra, интегрируют методы объяснения (SHAP, LIME, Anchors, Grad-CAM, Attention Flow) непосредственно в пайплайн эксплуатации, позволяя в реальном времени оценивать, какие факторы влияют на предсказания модели. Это критически важно для задач, в которых допустимы лишь объяснимые автоматические решения (например, отказ в кредите, назначение терапии, определение приоритета вызова экстренной помощи). Интерпретируемость здесь выполняет не просто диагностическую функцию, но и защитную: в случае появления аномальных паттернов объяснений или резкого смещения распределения важности признаков система может сигнализировать об ухудшении качества модели или потенциальной атаке [3].

Особое значение приобретает контроль интерпретируемости в интерфейсе пользователя, особенно если модель предоставляет рекомендации, а не финальные решения. Объяснение вывода – будь то через визуализацию вкладов признаков, аналогичные случаи из базы, или через генеративное описание ("модель считает, что данный пациент имеет высокий риск из-за сочетания возраста, истории заболеваний и уровня сахара") – становится основой пользовательского доверия. Исследования показывают, что интерфейсы, поддерживающие интерпретируемость, повышают не только принятие ИИ, но и качество совместной работы человека и машины, поскольку позволяют пользователю корректировать поведение модели или уточнять входные данные [4].

На этапе постэксплуатационного анализа и аудита, особенно в случаях, когда необходимо объяснить спорное или ошибочное решение, интерпретируемость становится механизмом ретроспективного анализа. Если модель отказывает в медицинской процедуре или юридическом рассмотрении, необходимо доказуемо показать, какие основания лежали в основе такого вывода. В условиях жёсткого регулирования (например, AI Act в ЕС, Equal Credit Opportunity Act в США) такие объяснения должны быть не только технически корректны,

но и когнитивно понятны, юридически пригодны и воспроизводимы. Интерпретируемость в таком контексте становится гарантией процедурной прозрачности и формирует основу для ответственности – как разработчиков, так и пользователей модели.

Важно подчеркнуть, что контроль интерпретируемости не может быть «надстройкой» над моделью. Он должен быть интегрирован в инфраструктуру, поддерживаться на уровне логирования, версионности, мониторинга, визуализации и взаимодействия с внешними интерфейсами. Это требует координации между командами разработчиков, аналитиков, специалистов по этике и юридических служб. В техническом плане такая интеграция реализуется через интерпретируемые протоколы API, стандартизованные схемы отчётов (например, Model Cards, Fairness Reports), автоматическую генерацию объяснений и логов, встроенные модули объяснения в пайплайны MLOps.

В долгосрочной перспективе, интерпретируемость может стать основой для адаптивного доверия: системы, способные объяснять своё поведение, могут накапливать «репутацию» надёжности, адаптироваться к требованиям пользователя, документировать свои решения и проходить сертификацию на соответствие нормативам [5].

Таким образом, поэтапный контроль интерпретируемости превращается из факультативного аспекта разработки в фундаментальный принцип архитектуры доверенного искусственного интеллекта. Он позволяет не только повышать прозрачность и ответственность, но и формирует основу для совместной эволюции человека и ИИ в условиях усложняющейся цифровой среды.

Список литературы:

1. Bhatt, U. Explainable Machine Learning in Deployment / U. Bhatt, A. Weller, J. M. F. Moura // Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. – New York : ACM, 2020. – P. 648–657.

2. Sokol, K. Explainability Fact Sheets : A Framework for Systematic Assessment of Explainable Approaches / K. Sokol, P. A. Flach // Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. – New York : ACM, 2020. – P. 56–67. – DOI 10.1145/3351095.3372870.

3. Lakkaraju, H. How Do I Fool You? Manipulating User Trust via Misleading Black Box Explanations / H. Lakkaraju, O. Bastani // Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. – New York : ACM, 2020. – P. 79–85. – DOI 10.1145/3375627.3375833.

4. Poursabzi-Sangdeh, F. Manipulating and Measuring Model Interpretability / F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman [et al.] // Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. – New York : ACM, 2021. – Article 237. – DOI 10.1145/3411764.3445315.

5. Ehsan, U. Human-Centered Explainable AI : Towards a Reflective Sociotechnical Approach / U. Ehsan, M. O. Riedl // HCI International 2020. – Cham : Springer, 2020. – P. 449–466. – DOI 10.1007/978-3-030-60117-1_33.