МЕТОДЫ ОЦЕНКИ ДОВЕРИЯ И ПРОЗРАЧНОСТИ РЕШЕНИЙ ИИ-СИСТЕМ

Вакарь Е. И., Коростелёв А. Д., Романюк М. В.
Центр информатизации и инновационных разработок,
Белорусский государственный университет информатики и радиоэлектроники
Минск, Республика Беларусь
E-mail: {e.vakar, a.korostelev, romanuk}@bsuir.by

В статье рассматриваются современные подходы к оценке доверия и прозрачности решений систем искусственного интеллекта. Приведена классификация методов объяснимого ИИ (XAI), описаны метрики доверия и уверенности в предсказаниях моделей, а также анализируются факторы, влияющие на восприятие прозрачности пользователями. Отмечается актуальность разработки интерпретируемых решений для повышения безопасности и надёжности применения ИИ в критически важных областях.

Введение

Развитие искусственного интеллекта (ИИ) приводит к широкому внедрению интеллектуальных систем в различных сферах – от медицины до образования. Однако принятие решений такими системами зачастую непрозрачно для пользователя, что вызывает сомнения в достоверности и надёжности их выводов. В связи с этим возрастает необходимость разработки методов, обеспечивающих доверие к ИИ и прозрачность его решений [1].

I. Методы объяснимости и интерпретации

Современные методы объяснимого искусственного интеллекта (XAI, Explainable Artificial Intelligence [2]) делятся на локальные и глобальные. Локальные методы направлены на объяснение отдельных решений модели, то есть того, почему именно в данном случае ИИ выдал определённый результат.

К числу наиболее распространённых относятся:

- LIME (Local Interpretable Model-agnostic Explanations) – создаёт упрощённую линейную модель, приближающую поведение сложной модели в окрестности конкретного примера. Позволяет определить, какие признаки оказали наибольшее влияние на конкретный прогноз.
- SHAP (Spapley Additive Explanations) – основан на теории игр и вычисляет «вклад» каждого признака в итоговое решение модVели, что обеспечивает теоретическую интерпретируемость и универсальность подхода.
- Grad-CAM (Gradient-weighted Class Activation Mapping) – используется в свёрточных нейронных сетях для визуализации областей изображения, которые повлияли на классификацию, что особенно полезно для анализа моделей компьютерного зрения.

Глобальные методы ориентированы на интерпретацию поведения модели в целом. К ним относятся:

- Surrogate Models – построение интерпретируемой модели (например, дерева решений), которая аппроксимирует поведение сложной модели «чёрного ящика».
- Feature Importance – оценка значимости признаков, используемых моделью при обучении. Такой анализ помогает выявить наиболее критичные факторы, влияющие на решения системы.
- Partial Dependence Plots (PDP) и Accumulated Local Effects (ALE) – визуализируют зависимость прогнозов от отдельных признаков или их комбинаций, позволяя анализировать взаимное влияние входных данных.

II. Методы оценки доверия

Доверие к системам искусственного интеллекта является ключевым фактором их успешного внедрения и принятия пользователями. Если объяснимость (XAI) помогает понять почему модель выдала конкретный результат, то методы оценки доверия позволяют количественно и качественно определить, насколько можно полагаться на это решение [3].

Выделяют два основных аспекта доверия:

1. Техническое доверие – оценка корректности, устойчивости и предсказуемости поведения модели.
2. Психологическое (пользовательское) доверие – восприятие надёжности и справедливости системы со стороны человека.

Классическим подходом технической оценки доверия является анализ уверенности модели (confidence) – вероятностной оценки, выдаваемой системой вместе с результатом предсказания. Однако высокая уверенность не всегда гарантирует правильность ответа. Поэтому применяются дополнительные метрики и подходы:

- Оценка калибровки модели – проверка, насколько вероятностные оценки совпадают

с фактическими результатами. Хорошо откалиброванная модель с уверенностью 0.8 должна быть права примерно в 80% случаев. Для анализа используются такие показатели, как Expected Calibration Error (ECE) и Brier Score.

- Методы оценки неопределённости (uncertainty estimation) – позволяют количественно описывать степень неуверенности модели в своём предсказании. Применяются байесовские нейронные сети, Dropout-аппроксимация, ensemble-модели и bootstrapping.
- Оценка устойчивости (robustness) – анализ того, как изменяются результаты при незначительных изменениях входных данных. Высокая чувствительность к шуму свидетельствует о низком уровне доверия.
- Методы детектирования аномалий и выбросов – позволяют выявить ситуации, в которых модель работает вне обучающего распределения, что повышает осознанность системы относительно собственных ограничений.

Немаловажным аспектом является восприятие доверия со стороны человека, взаимодействующего с ИИ-системой. Даже технически надёжная модель может вызвать недоверие при недостатке прозрачности интерфейса или отсутствии объяснений.

Для анализа этого аспекта применяются:

- Опросы и UX-исследования – оценка субъективного восприятия доверия, удобства, справедливости и понятности решений системы.
- Методы интерактивного XAI – предоставление пользователю возможности запросить объяснение, проверить альтернативные сценарии и наблюдать реакцию модели на изменения данных.
- Визуальные интерфейсы доверия (Trust Dashboards) – отображают уверенность модели, степень неопределённости и происхождение данных, что повышает прозрачность системы в реальном времени.

Современные исследования направлены на встраивание доверительных механизмов непосредственно в архитектуру ИИ-систем. Применяются подходы, при которых модель одновременно обучается на задачу предсказания и задачу оценки собственной уверенности (self-assessment). Кроме того, развивается направление Trustworthy AI, объединяющее методы XAI, оценку неопределённости, устойчивость к атакам и этические аспекты использования ИИ.

Таким образом, оценка доверия представляет собой комплексную задачу, включающую технические, поведенческие и когнитивные аспекты. Эффективная система доверия должна не только корректно предсказывать результаты, но и уметь аргументировать их, демонстрировать устойчивость и давать пользователю возможность понять границы своей надёжности.

III. Современные тенденции

С ростом популярности больших языковых моделей (LLM), таких как GPT, всё актуальнее становятся вопросы прозрачности их ответов и интерпретируемости рассуждений. В отличие от традиционных алгоритмов, LLM используют масштабные текстовые корпуса и вероятностные зависимости, что усложняет понимание логики их работы. Активно исследуются подходы к объяснению вывода таких моделей, включая трассировку источников знаний, анализ цепочек рассуждений и визуализацию внутренних представлений. Эти методы повышают уровень доверия и делают процесс генерации более прозрачным [4].

Также активно развивается нормативное регулирование доверия к ИИ. Европейский AI Act определяет требования к прозрачности и контролю решений моделей в зависимости от степени риска их применения [5].

Заключение

Прозрачность и доверие являются ключевыми характеристиками безопасного и ответственного использования ИИ. Перспективными направлениями остаются разработка универсальных метрик доверия, интеграция XAI-инструментов в промышленные системы и создание стандартов оценки объяснимости решений.

Дальнейшее развитие в этой области требует тесного взаимодействия исследователей, разработчиков и регуляторов. Необходимы единые подходы к оценке интерпретируемости и надёжности моделей, а также внедрение принципов этичного ИИ на этапах проектирования и эксплуатации. Совместное развитие технологий, методологий и нормативных требований позволит сформировать экосистему искусственного интеллекта, основанную на прозрачности, ответственности и уверенности пользователей в результатах его работы.

IV. Список литературы

1. Long, B. Explainable AI the Latest Advancements and New Trends / B. Long, E. Liu, R. Qiu, Y. Duan // arXiv preprint arXiv:2505.07005, 2025.
2. Saarela, M. Recent Applications of Explainable AI (XAI): A Systematic Literature Review / M. Saarela, V. Podgorelec. //Appl. Sci. 2024
3. Zahra Atf. Is Trust Correlated With Explainability in AI? A Meta-Analysis. / Zahra Atf, P. R. Lewis // arXiv:2504.12529, 2025
4. Palikhe, A. Towards Transparent AI: A Survey on Explainable Large Language Models. / A. Palikhe, Z. Yu, Z. Wang, W. Zhang // arXiv preprint arXiv:2506.21812, 2025
5. European Commission / Artificial Intelligence Act – Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024