

*М.Д. Лебедев¹, Ю.В. Трофимов^{2, 3}, А.Д. Лебедев², И.А. Трусов²,
А.Н. Аверкин², Е.С. Кондрашова²*

¹Национальный исследовательский технологический университет «МИСИС», г. Москва, Россия

²Государственный университет «Дубна», г. Дубна, Россия

³Объединённый институт ядерных исследований, г. Дубна, Россия

Развитие объяснимого искусственного интеллекта усилило интерес к моделям, в которых объяснение связано с вычислительной процедурой. Для систем поддержки принятия решений важно получить результат и проверяемое основание этого результата. Методы, применяемые после обучения, полезны для анализа сложных моделей, однако часто дают приближённое описание уже сформированного «чёрного ящика» [1]. Поэтому в ХАИ 2.0 возрастает роль архитектур, где объяснение становится частью проектируемой системы [2].

Перспективной основой такого подхода являются сети Колмогорова-Арнольда (KAN). Обучаемые функции на рёбрах сети позволяют переходить от численного прогноза к анализу функциональной зависимости [3]. В ранее предложенной гибридной и иерархической ХАИ-архитектуре KAN рассматривались как центральный блок, объединяющий внутреннюю интерпретируемость, нечёткую логику и многоуровневый анализ [4]. В настоящей работе этот принцип уточняется до контура верифицируемых объяснений, где решение связывается с правилом, экспертной оценкой и текстовым пояснением.

Основная часть. В работе используется сравнительно-архитектурный подход. Сначала рассматривается классическая схема объяснимого ИИ, в которой SHAP, LIME и Grad-CAM применяются после обучения основной модели. Затем анализируется вариант, где KAN выступает интерпретируемым слоем и представляет часть зависимости через обучаемые функции. Такой переход важен, поскольку объяснение формируется не внешним комментарием к ответу, а структурой самой модели.

Предлагаемый контур строится как последовательная цепочка проверки. Входные данные проходят предварительное выделение признаков, после чего KAN-блок формирует функции на рёбрах сети. Эти функции служат основанием для нечётких правил, которые проходят экспертную проверку и преобразуются в текстовое объяснение. Общая схема контура представлена на рис. 1.



Рис. 1. Контур верифицируемого объяснения в гибридной системе на основе KAN

Представленная на рис. 1 схема показывает, как в гибридной системе на основе KAN формируется верифицируемое объяснение. После предварительного выделения признаков данные поступают в KAN-блок, где выявляются функциональные зависимости. Эти зависимости служат основой для построения нечётких правил. Затем правила проходят экспертную проверку и преобразуются в текстовое объяснение, понятное специалисту. Обратная связь от экспертной проверки к блоку нечётких правил позволяет уточнять объяснение и отсекают варианты, которые слабо согласуются с предметной областью [5].

Схема на рис. 1 также показывает, в каких условиях KAN целесообразно включать в архитектуру. При низкой или средней размерности признакового пространства KAN может выступать основным интерпретируемым блоком. При более сложных данных его разумно использовать после предварительного выделения признаков или в роли дополнительного проверочного слоя. Такой выбор помогает сохранить объяснимость и одновременно учитывать вычислительные ограничения метода [6].

Выводы. Рассмотрен подход к построению гибридной системы объяснимого искусственного интеллекта, в которой KAN служит основой для формирования верифицируемого объяснения. Предложенный контур связывает функциональную зависимость, нечёткие правила, экспертную проверку и текстовое пояснение в единую цепочку. За счёт этого объяснение становится частью архитектуры и может быть предметом содержательной проверки и дальнейшего уточнения.

Работа выполнена в рамках государственного задания Министерства науки и высшего образования Российской Федерации (тема № 124112200072-2).

Литература

1. Doshi-Velez F., Kim B. Towards a rigorous science of interpretable machine learning [Электронный ресурс]. URL: <https://arxiv.org/abs/1702.08608> (дата обращения: 20.05.2026).
2. Longo L. et al. Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions // Information Fusion. 2024. Vol. 106. Article 102301.
3. Liu Z. et al. KAN: Kolmogorov-Arnold Networks [Электронный ресурс]. URL: <https://arxiv.org/abs/2404.19756> (дата обращения: 20.05.2026).
4. Averkin A.N., Trofimov Y.V., Lebedev A.D., Ilin A.S. Hybrid and Hierarchical Explainable AI Based on Kolmogorov-Arnold Networks // Lecture Notes in Networks and Systems. 2026. Vol. 1763. P. 36-43. DOI: 10.1007/978-3-032-13612-1_4.
5. Guillaume S. Designing fuzzy inference systems from data: an interpretability-oriented review // IEEE Transactions on Fuzzy Systems. 2001. Vol. 9, No. 3. P. 426-443.
6. Tran V. D. et al. Exploring the limitations of Kolmogorov-Arnold Networks in classification [Электронный ресурс]. URL: <https://arxiv.org/abs/2407.17790> (дата обращения: 20.05.2026).