УДК 004.89

Е.А. КУЗИНА

*Национальный исследовательский ядерный университет «МИФИ», Москва*

# МЕТОДЫ ПОВЫШЕНИЯ ДОВЕРИЯ К ТЕХНОЛОГИЯМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В ЗАДАЧАХ АТОМНОЙ ЭНЕРГЕТИКИ

Естественным требованием к разрабатываемому ПО мониторинга и контроля состояния АЭС, осложняющим внедрение систем искусственного интеллекта, является безопасность, и, в частности, интерпретируемость, используемых технологий и моделей. Целью данной работы является анализ существующих подходов, способствующих повышению интерпретируемости моделей глубокого обучения. Рассмотрены три направления исследования данной проблемы, проанализированы их современное состояние и перспективы применения в компьютеризации оперативных процедур на АЭС.

## Введение

Опыт внедрения систем искусственного интеллекта (ИИ) в атомной энергетике демонстрирует их применение, в основном, в качестве систем поддержки оператора (СПО) по оперативной диагностике оборудования [1–2], что фактически реализует вспомогательную упреждающую функцию. Таким образом, в ответственности оператора частично остается оценка ситуации и полностью – планирование ответных действий. В то же время существуют работы, доказывающие эффективность применения методов ИИ в решении более широкого класса задач [3–4], направленных в том числе на повышение автономности работы АЭС. Препятствием к внедрению таких систем, в частности основанных на глубоком обучении, является недостаточный уровень доверия к ним, одним из факторов которого является низкая интерпретируемость таких моделей.

В данной работе представлен анализ подходов, направленных на повышение интерпретируемости моделей глубокого обучения и применимых в решении задач атомной энергетики.

## Подходы к повышению интерпретируемости моделей ИИ

Среди подходов, способствующих повышению интерпретируемости моделей глубокого обучения, можно выделить следующие: применение объяснимого искусственного интеллекта (англ. eXplainable Artificial Intelligence, XAI) [5]; использование гибридных моделей; переход на принципиально новые модели нейронных сетей.

Исследования в области объяснимого ИИ направлены на создание техник верификации «осмысленности» заключений модели, что фактически не совершенствует саму модель в вопросе интерпретируемости.

Гибридизация моделей позволяет вносить ограничения в закономерности, которым обучается модель ИИ. С одной стороны, такой подход позволяет сдерживать модель в рамках некоторой физически обоснованной теории, с другой – может привести к ограничению способности модели по выявлению скрытых признаков для объяснения изучаемых закономерностей, снизить качество модели.

Смена архитектуры или парадигмы обучения нейронных сетей также может стать ключом к обеспечению интерпретируемости моделей. Так, свойства сети Колмогорова-Арнольда (KAN) [6] позволяют решать вопрос интерпретируемости модели на уровне ее архитектуры. Однако обучение KAN на существующих вычислителях затруднено.

## Заключение

Исследованы проблемы применения технологии ИИ в задачах атомной энергетики. Проанализированы особенности направлений ИИ, применение которых способно повысить интерпретируемость моделей, а следовательно, способствовать достижению требуемого уровня доверия к технологии для ее распространения в решении задач атомной энергетики.

*Список литературы*

1. Поваров В.П. Принципы разработки систем принятия решений в задачах управления ядерными блоками. Вестник Воронежского государственного технического университета. 2023, т. 14, № 2, с. 87–91.

2. Сборник «2020-2021 годы: краткие результаты научно-технической деятельности АО «ВНИИАЭС». URL: https://vniiaes.ru/upload/Сборник_ОР_ВНИИАЭС_2020_2021.pdf (дата обращения 14.09.2024).

3. Николаева А.В., Увакин М.А., Пантюшин С.И., Сотсков Е.В., Антипов М.В., Николаев А.Л., Литышев А.В., Безруков Ю.А., Кавун О.Ю., Быков М.А. (АО ОКБ «ГИДРОПРЕСС») Искусственный интеллект в области использования атомной энергии - существующие возможности и перспективы. Вопросы атомной науки и техники. Серия: Физика ядерных реакторов. 2023, № 3, с. 4–16.

4. Huang Q., Peng Sh., Deng J., Zeng H., Zhang Z., Liu Y., Yuan P. A review of the application of artificial intelligence to nuclear reactors: where we are and what's next. Heliyon. 2023, v. 9, p. e13883.

5. Ali S., Abuhmed T., El-Sappagh S., Muhammad Kh., Alonso-Moral J.M., Confalonieri R., Guidotti R., Ser J.D., Díaz-Rodríguez N., Herrera F. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Information Fusion. 2023, v. 99, p. 101805.

6. Liu Z., Wang Y., Vaidya S., Ruehle F., Halverson J., Soljacc M., Hou T. Y., Tegmark M. KAN: Kolmogorov-Arnold Networks. ArXiv preprint, 2024. arXiv:2404.19756.