# ЭЛЕКТРОТЕХНИКА

**УДК 621.311:004.89**

## МЕТОДЫ ОБЪЯСНИМОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ ЭНЕРГЕТИЧЕСКИХ И ЭЛЕКТРОЭНЕРГЕТИЧЕСКИХ СИСТЕМ: ОБЗОР, ПРОБЛЕМЫ И ВОЗМОЖНОСТИ

***М.А. Куликова, О.Н. Торгованова, К.Б. Корнеев***
*Тверской государственный технический университет (г. Тверь)*



**Аннотация.** В статье рассмотрены перспективы использования методов объяснимого искусственного интеллекта (XAI) в сфере электроэнергетики, связанные с повышением объяснимости моделей машинного обучения и улучшением понимания их результатов. Представлены общие проблемы и ограничения, вызванные внедрением методов XAI в сфере энергетики и электроэнергетических систем.

**Ключевые слова:** электроэнергетика, энергосистема, нейронные сети, глубокое обучение, объяснимый искусственный интеллект, XAI.

**DOI: 10.46573/2658-7459-2024-4-50-57**

Развитие методов глубокого обучения (от англ. deep learning – DL) привело к созданию более совершенных классификаторов и алгоритмов машинного обучения (от англ. machine learning – ML), которые нашли применение в энергосистемах [1]. В некоторых случаях эти методы демонстрируют значительные преимущества перед традиционными алгоритмами оптимизации с точки зрения эффективности, устойчивости к помехам и точности.

Тем не менее, несмотря на очевидный успех названных алгоритмов, существует важная проблема. Модели ML зачастую бывают очень сложными, и оператору может быть непонятно, как и почему они принимают те или иные решения, а также каким образом они обрабатывают входящие данные. Планированием и эксплуатацией энергосистем занимаются исключительно специалисты в данной области. Они опираются на имеющиеся знания об этих системах, специализированные программы и накопленный опыт. Экспертам зачастую тяжело доверять решениям и рекомендациям, предложенным алгоритмами ML, и это ограничивает их практическое применение. Данная трудность особенно заметна в случаях, требующих высокого уровня надежности, что характерно для энергетической отрасли [2]. Верно и обратное: специалисты в области программирования алгоритмов не знают специфики работы энергосистем, поэтому не могут в полной мере организовать обучение моделей искусственного интеллекта (ИИ).

В последние годы активно разрабатываются новые подходы и концепция, направленная на повышение объяснимости и понятности результатов моделей ML. Эта

концепция получила название «объяснимый искусственный интеллект» (от англ. Explainable Artificial Intelligence – XAI). Цель XAI заключается в том, чтобы помочь исследователям, разработчикам, специалистам и пользователям лучше понимать, как работают модели ML, сохраняя при этом их эффективность и точность [3]. На рис. 1 представлены ключевые этапы, связанные как с ИИ, так и с энергетическими системами. Видно, что использование XAI в области энергетики находится пока что на начальном этапе.

Одним из минусов многих алгоритмов ML является их непонятность. Это значит, что такие алгоритмы трудно объяснить, даже специалисты в данной области не каждый раз могут это сделать. Если пользователи считают модель «черным ящиком», они не всегда доверяют ее прогнозам и поэтому ее не используют [4]. Кроме того, многие нейронные сети – это очень сложные модели, которые трудно понять даже экспертам в области ИИ. Их архитектура разрабатывается методом проб и ошибок и может содержать сотни уровней и миллиарды входных параметров.

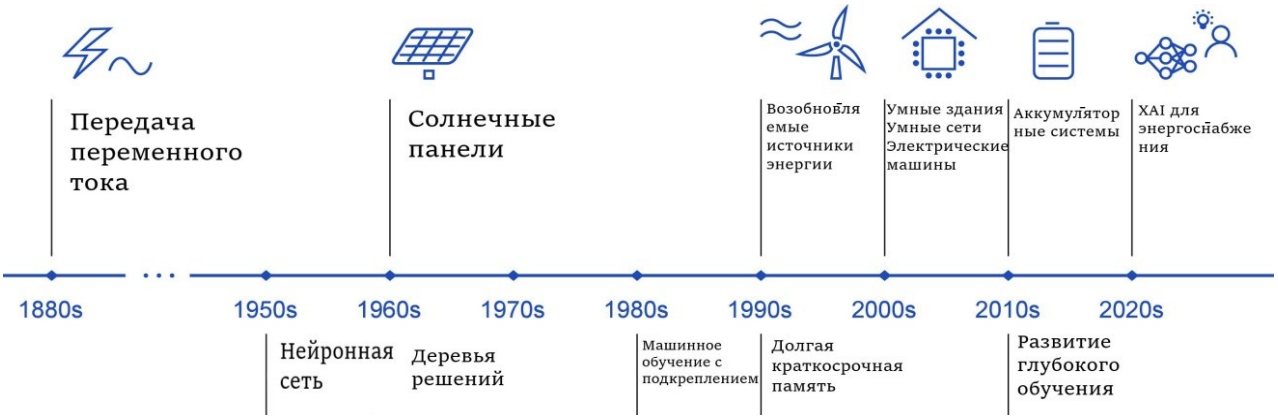



Рис. 1. Путь к XAI для энергетики и энергетических систем

С учетом указанной проблемы следует уточнить, что главная цель XAI – дать исследователям, разработчикам и пользователям возможность лучше понимать результаты моделей ML. Кроме того, предполагается, что добавление объяснимости позволит сделать обобщение моделей ML более надежным, и в дальнейшем это поможет улучшить их управляемость. Данные идеи были реализованы в сфере энергетических систем (рис. 2). На рис. 3 представлен конкретный пример оценки эффективности энергопотребления здания [5]. Здесь XAI используется для предоставления пользователю сведений о том, какие функции являются важными.

Есть две основные области объяснения, на которые ориентированы методы XAI. Первая область – это «локальное объяснение», при котором входными данными для XAI являются отдельные элементы входных данных, т.е. карта объяснения $g$ генерируется каждый раз для отдельного элемента данных. Другая область – «глобальное объяснение», когда целью является понимание функционирования всей модели. Для этого используются группы наборов данных, на их основе создается объяснение $g$. На рис. 4 представлена общая схема локальных и глобальных объяснений.

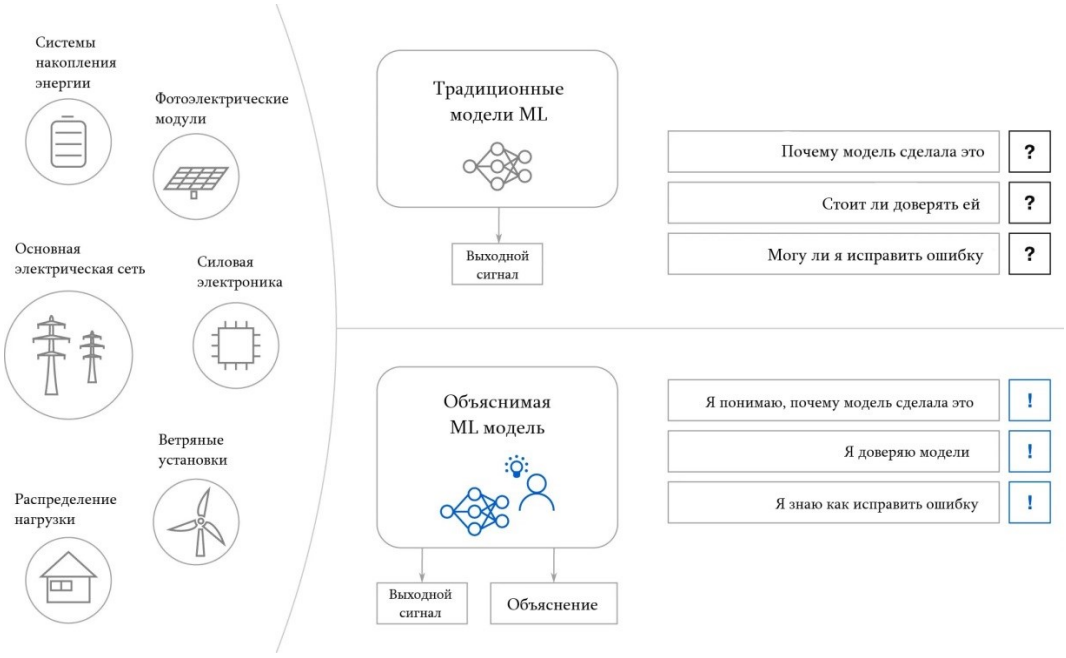
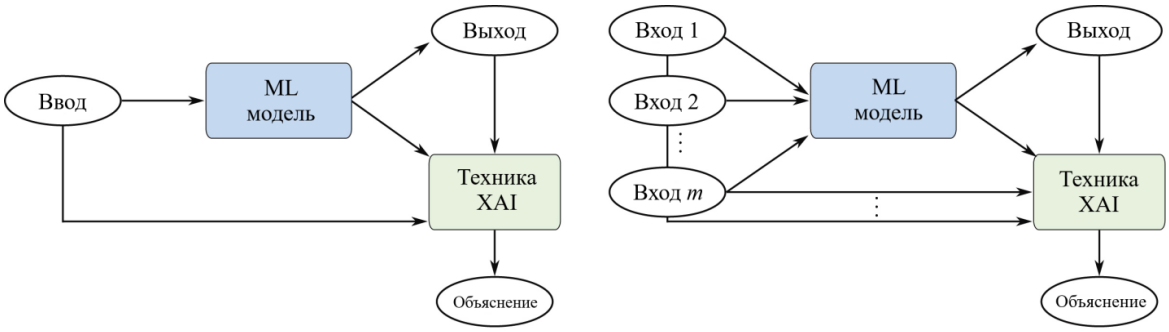



Рис. 2. Концепции XAI для общих задач энергосистем

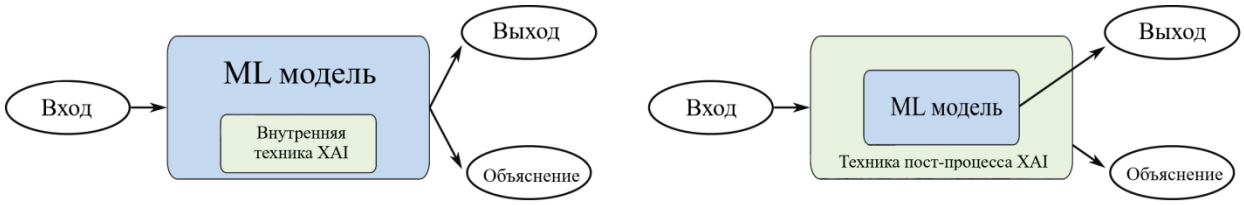

Рис. 3. Пример модели классификатора с XAI
для оценки эффективности энергопотребления в здании




а б

Рис. 4. Концептуальная иллюстрация объяснимых алгоритмов:
а – локальных; б – глобальных

Еще одна важная концепция XAI – это интеграция объяснений в саму модель. Объяснения могут быть частью конкретной модели ML или применяться к любой модели в качестве постпроцесса. Существует два подхода к интеграции методов XAI: внутренний, который зависит от модели, и постспециальный, который не зависит от нее. При внутреннем подходе объяснимость становится частью архитектуры модели ML и не может быть перенесена на другие архитектуры. При постспециальном подходе метод XAI не связан с конкретной архитектурой и может быть применен к любой обученной модели ML. На рис. 5 показана высокоуровневая схема для внутреннего и постспециального XAI.




а б

Рис. 5. Высокоуровневая иллюстрация алгоритмов объяснимой модели:
а – внутреннего; б – постспециального

Таким образом, локальные объяснения помогают понять, почему было принято то или иное решение. Глобальные объяснения позволяют узнать систему в целом и оптимизировать ее на основе знаний, полученных моделью. Встроенные методы полезны при обучении новой модели, для которой понимание имеет ключевое значение. Специальные методы дают возможность использовать уже обученные модели или проверенные методы ML.

При внедрении XAI в системы электроснабжения необходимо учитывать ряд проблем и ограничений. В таблице представлено их краткое описание.

Существующие ограничения применения моделей ИИ в энергетике

| Категория | Проблемы и ограничения |
|---|---|
| Компромиссы | Как объяснение, так и результаты должны быть точными |
| Стандартизация | Отсутствие согласия в определении XAI и объяснимости.<br>Различные типы пользователей: исследователи ИИ, эксперты в области энергетики, разработчики энергетической политики и потребители.<br>Должны быть даны обоснованные объяснения |
| Показатели оценки | Следует использовать объективные показатели,<br>если объяснение не может быть определено.<br>Необходимо определить и оценить показатель объяснимости |
| Безопасность | Вредоносные атаки могут быть сгенерированы<br>с использованием информации, полученной<br>с помощью методов XAI.<br>Конфиденциальность моделей ML может быть нарушена |

*Окончание таблицы*

| Категория | Проблемы и ограничения |
|---|---|
| Пользователи | Методы XAI должны оптимально подходить для различных типов пользователей. Сложность сотрудничества между экспертами в области энергетики и экспертами в других областях |
| Рекомендации | Потребители могут быть введены в заблуждение и полагаться на недостоверные результаты. Нельзя допускать, чтобы методы XAI выдавали вводящие в заблуждение объяснения |

Одной из главных проблем, относящихся непосредственно к сфере энергетики, является возможность использования модели, которая обеспечивает высокую производительность и при этом остается понятной [6]. Обычно точные модели бывают сложными для восприятия и понимания. Компромисс имеет особое значение в области электроэнергетики, поскольку пользователю (персоналу энергосистемы) требуется как высокая производительность, так и точное объяснение, чтобы обеспечить высокий уровень надежности.

Еще одним существенным недостатком XAI можно выделить отсутствие стандартизации и четких определений. В настоящее время нет единого мнения о том, как именно следует понимать XAI и объяснимость его работы. Некоторые исследователи сосредоточены на методах визуализации, в то время как остальные используют концепцию важности объектов или их релевантности. Одна из причин отсутствия единого стандарта заключается в том, что существуют разные типы пользователей, применяющих модели ML и XAI: исследователи в области ИИ, эксперты в энергетике, разработчики энергетической политики и потребители [7]. Каждый из них использует методы ML для своих нужд на основе различных факторов и уровней абстракции. Еще один важный вопрос состоит в том, как определить оптимальное объяснение. Во многих приложениях не существует четко установленного метода для предоставления понятного и оптимального объяснения. Без возможности решить эту задачу использование XAI может стать неясным, а результаты работы алгоритма XAI – трудно интерпретируемыми. Кроме того, хотя данные, как правило, содержат классификационные метки для процесса обучения, обычно не приводятся обоснованные объяснения. Без таких объяснений результаты XAI не будут иметь ориентира для сравнения. Таким образом, при создании баз входных данных для любого энергетического приложения важно, по возможности, включать в них обоснованные объяснения [8].

Существенным недостатком методов XAI является отсутствие объективных критериев оценки качества объяснений. Даже если можно дать четкое определение объяснимости, желательно иметь показатель, который оценивал бы, насколько модель «объяснима». Эти показатели должны измерять объяснимость для каждого метода XAI и любого классификатора. Значение оценки должно основываться на том, насколько предполагаемое объяснение соответствует обоснованию. В отличие от многих других распространенных задач ML, в области энергетики существует множество приложений, где правильное объяснение может быть четко определено (например, в задаче обнаружения и классификации аномальных событий в электросети). В этом случае объяснение может включать время возникновения события или его причину [9]. Таким

образом, объяснение можно представить в виде двоичного вектора, который показывает наличие или отсутствие события в каждый момент времени выборки. Определение оценочного показателя основывается на корректном объяснении и сравнивается с результатами, полученными с помощью XAI. Если объяснение не может быть точно определено, то установить объективные показатели для каждого пользователя сложно.

Еще одним критически важным аспектом, который необходимо учитывать, является информационная безопасность. В сфере энергетики и систем энергоснабжения данный вопрос стоит особенно остро. Когда применяются методы XAI для объяснения модели ML, конфиденциальность последней может быть нарушена из-за необходимости обработки большого объема входных данных, получаемых не только по исследуемому, но и по сторонним объектам электроэнергетики, с существующей фактической привязкой последних к смежным объектам энергосистемы. В силу наличия у системы доступа к такому объему входных данных существует некоторая вероятность обратного извлечения исходных данных, а также возможность управления системой их обработки. Любая информация, полученная с помощью методов ИИ, может быть использована для создания эффективных вредоносных атак, направленных на дезориентацию модели. Эти атаки включают, к примеру, манипулирование моделью путем предоставления системе определенных ложных данных, что приводит к искаженным результатам. Основываясь на данных, полученных с помощью XAI, такие атаки могут проявить себя как эффективные, поскольку они позволяют найти минимальные изменения во входных данных, необходимые для изменения решения модели ML. Последствия данных атак для электрических сетей и систем управления энергопотреблением оказываются серьезными и даже катастрофическими. Кроме того, надежность объяснений должна поддерживаться за счет постоянного получения актуальных данных о состоянии электроэнергетической системы. Открытый для систем ИИ доступ к сбору и обработке этих данных может привести к разрушению внутренней системы безопасности объекта электроэнергетики или энергосистемы в целом [10].

Другая проблема заключается в том, что современные методы, используемые для объяснения результатов работы XAI, разработаны с учетом потребностей экспертов в области ИИ, а не специалистов по энергетическим системам [11]. В настоящее время самые передовые алгоритмы XAI созданы учеными-компьютерщиками и исследователями ИИ. Таким образом, наиболее распространенным способом объяснения результатов является использование так называемых «тепловых карт», которые обеспечивают наглядность работы, но не всегда содержат достаточное количество информации для пользователя. Во многих приложениях более сложное представление выходных данных XAI может привести к более эффективному объяснению алгоритмов работы и принятия решений. В связи с этим было бы целесообразно наладить сотрудничество между экспертами в области энергетики и другими специалистами, чтобы создать эффективные и специализированные методы XAI, идеально подходящие для приложений в области энергетики. Например, в методы, где объясняются и даются рекомендации по эксплуатации электрических систем, можно было бы включить информацию, основанную на представительных знаниях системных операторов.

Ключевая задача систем ИИ в энергетике заключается в том, чтобы предотвратить предоставление пользователями неверных объяснений в рамках XAI и обеспечить надежные рекомендации. Хотя объяснения могут усилить доверие пользователей к системе, в долгосрочной перспективе результаты работы моделей не всегда оказываются

точными. Это может привести к тому, что пользователи начнут доверять ошибочным выводам, создавая ложную убежденность. Кроме того, доверие к неверным объяснениям или прогнозам заставляет их использовать неэффективные или небезопасные модели, что в будущем может привести к серьезным проблемам. Чтобы обеспечить доверие к локально объяснимому алгоритму, необходимо проверить его на различных входных данных. При этом подход, основанный на глобально объяснимом алгоритме, позволяет получить объяснения, используя весь набор данных [12]. Сравнение этих внутренних и временных, локальных и глобальных подходов к энергетическим приложениям требует дальнейшего изучения.

## СПИСОК ЛИТЕРАТУРЫ

1. Ozcanli A.K., Yaprakdal F., Baysal M. Deep Learning Methods and Applications for Electrical Power Systems: A Comprehensive Review // *International Journal of Energy Research*. 2020. No. 44 (9), pp. 7136–7157.
2. Массель Л.В. Современный этап развития искусственного интеллекта (ИИ) и применение методов и систем ИИ в энергетике // *Информационные и математические технологии в науке и управлении*. 2021. № 4 (24). С. 5–18.
3. Machlev R., Perl M., Belikov J., Levy K., Levron Y. Measuring Explainability and Trustworthiness of Power Quality Disturbances Classifiers Using XAI – Explainable Artificial Intelligence // *IEEE Transactions on Industrial Informatics*. 2021, pp. 5127–5137.
4. Adadi A., Berrada M. Peeking Inside the Black-box: A Survey on Explainable Artificial Intelligence (XAI) // *IEEE Access*. 2018. No. 6, pp. 52138–52160.
5. Корнеев К.Б., Павлова Ю.М., Осеи-Овусу Р. Алгоритмические модели управления электрической нагрузкой в системах электроснабжения // *Вестник Тверского государственного технического университета. Серия «Строительство. Электротехника и химические технологии».* 2022. № 3 (15). С. 40–50.
6. Kruse J., Schäfer B., Witthaut D. Revealing Drivers and Risks for Power Grid Frequency Stability with Explainable AI // *Patterns*. 2021. No. 2 (11), pp. 1–26.
7. Любарский Ю.Я. Оперативный диспетчерский анализ нештатных ситуаций в электрических сетях промышленных предприятий – компьютерная поддержка на основе технологии экспертных систем // *Промышленная энергетика*. 2017. № 9. С. 2–6.
8. Donti P.L., Kolter J.Z. Machine Learning for Sustainable Energy Systems // *Annual Review of Environment and Resources*. 2021. No. 46 (1), pp. 719–747.
9. Ковалев С.П. Проектирование информационного обеспечения цифровых двойников энергетических систем // *Системы и средства информатики*. № 1. 2020. Т. 30. С. 66–81.
10. Корнеев К.Б., Окунева В.В., Павлова Ю.М. Открытость и защищенность протоколов передачи критической информации на объектах энергетики // *Вестник Тверского государственного технического университета. Серия «Строительство. Электротехника и химические технологии».* 2019. № 2 (2). С. 50–57.
11. Рассел С., Норвиг П. Искусственный интеллект: современный подход. 2-е изд. М.: И.Д. Вильямс, 2016. 1408 с.
12. Бушуев В.В. Интеллектуальное (когнитивное) прогнозирование и управление в энергетике // *Системные исследования в энергетике: методология и результаты* / под ред. А.А. Макарова, Н.И. Воропая. М.: МЭИ, 2018. С. 102–112.

## СВЕДЕНИЯ ОБ АВТОРАХ

*КУЛИКОВА Мария Александровна* – магистрант, ФГБОУ ВО «Тверской государственный технический университет», 170026, Россия, г. Тверь, наб. А. Никитина, д. 22. E-mail: mashakilikova@mail.ru
*ТОРГОВАНОВА Ольга Николаевна* – старший преподаватель кафедры иностранных языков, ФГБОУ ВО «Тверской государственный технический университет», 170026, Россия, г. Тверь, наб. А. Никитина, д. 22. E-mail: maerz@mail.ru
*КОРНЕЕВ Константин Борисович* – кандидат технических наук, доцент кафедры электроснабжения и электротехники, ФГБОУ ВО «Тверской государственный технический университет», 170026, Россия, г. Тверь, наб. А. Никитина, д. 22. E-mail: energy-tver@mail.ru

## БИБЛИОГРАФИЧЕСКАЯ ССЫЛКА

Куликова М.А., Торгованова О.Н., Корнеев К.Б. Методы объяснимого искусственного интеллекта для энергетических и электроэнергетических систем: обзор, проблемы и возможности // Вестник Тверского государственного технического университета. Серия «Строительство. Электротехника и химические технологии». 2024. № 4 (24). С. 50–57.

---

# EXPLAINABLE ARTIFICIAL INTELLIGENCE METHODS FOR POWER AND ELECTRICITY SYSTEMS: REVIEW, CHALLENGES AND POSSIBILITIES

***M.A. Kulikova, O.N. Torgovanova, K.B. Korneev***
*Tver State Technical University (Tver)*

**Abstract.** The paper considers the prospects of using explainable artificial intelligence (XAI) methods in the field of electric power industry related to increasing the explainability of machine learning models and improving the understanding of their results. Common problems and limitations caused by the introduction of XAI methods in the field of power engineering and electric power systems are presented.

**Keywords:** electric power industry, power system, neural networks, deep learning, explainable artificial intelligence, XAI.

## INFORMATION ABOUT THE AUTHORS

*KULIKOVA Mariya Aleksandrovna* – Master's Student, Tver State Technical University, 22, embankment of A. Nikitin, Tver, 170026, Russia. E-mail: mashakilikova@mail.ru
*TORGOVANOVA* Olga Nikolaevna – Senior Lecturer of the Department of Foreign Languages, Tver State Technical University, 22, embankment of A. Nikitin, Tver, 170026, Russia. E-mail: maerz@mail
*KORNEEV Konstantin Borisovich* – Candidate of Technical Sciences, Associate Professor of the Department of Power Supply and Electrical Engineering, Tver State Technical University, 22, embankment of A. Nikitin, Tver, 170026, Russia. E-mail: energy-tver@mail.ru

## CITATION FOR AN ARTICLE

Kulikova M.A., Torgovanova O.N., Korneev K.B. Explainable artificial intelligence methods for power and electricity systems: review, challenges and possibilities // Vestnik of Tver State Technical University. Series «Building. Electrical engineering and chemical technology». 2024. No. 4 (24), pp. 50–57.