НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
май–июнь 2025 Том 25 № 3 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
May–June 2025 Vol. 25 No 3 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

ОБЗОРНАЯ СТАТЬЯ

REVIEW PAPER

doi: 10.17586/2226-1494-2025-25-3-373-386
УДК 004.8

# Объяснимость и интерпретируемость — важные аспекты безопасности решений, принимаемых интеллектуальными системами (обзорная статья)

**Денис Николаевич Бирюков[1], Андрей Сергеевич Дудкин[2]✉**

[1,2] Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург, 197198, Российская Федерация

[1] Biryukov.D.N@yandex.ru, https://orcid.org/0000-0003-1300-2470
[2] andry-ll@mail.ru✉, https://orcid.org/0000-0003-0283-9048

**Аннотация**

Вопросы доверия к решениям, принимаемыми (формируемыми) интеллектуальными системами, становятся все более актуальными. Представлен систематический обзор методов и инструментов объяснимого искусственного интеллекта (Explainable Artificial Intelligence, XAI), направленных на преодоление разрыва между сложностью нейронных сетей и потребностью в интерпретируемости результатов для конечных пользователей. Проведен теоретический анализ различий между объяснимостью и интерпретируемостью в контексте искусственного интеллекта, а также их роли в обеспечении безопасности решений, принимаемых интеллектуальными системами. Показано, что объяснимость подразумевает способность системы генерировать понятные человеку обоснования, тогда как интерпретируемость сосредоточена на пассивной понятности внутренних механизмов. Предложена классификация методов XAI на основе их подхода (предварительный/последующий анализ: ante hoc/post hoc) и масштаба объяснений (локальный/глобальный). Рассмотрены популярные инструменты, такие как Local Interpretable Model Agnostic Explanations, Shapley Values и интегрированные градиенты, с оценкой их сильных сторон и ограничений применимости. Даны практические рекомендации по выбору методов для различных областей и сценариев. Обсуждается архитектура интеллектуальной системы, построенной на основе модели В.К. Финна, и адаптированной к современным требованиям к обеспечению «прозрачности» решений, где ключевыми компонентами являются информационная среда, решатель задач и интеллектуальный интерфейс. Рассмотрена проблема компромисса между точностью моделей и их объяснимостью: прозрачные модели («стеклянные ящики», например, деревья решений) уступают в производительности глубоким нейронным сетям, но обеспечивают большую бесспорность принятия решений. Приведены примеры методов и программных пакетов для объяснения и интерпретации данных и моделей машинного обучения. Показано, что развитие XAI связано с интеграцией нейро-символических подходов, объединяющих возможности глубокого обучения с логической интерпретируемостью.

**Ключевые слова**

искусственный интеллект, нейронные сети, глубокое обучение, «черный ящик», объяснимость, интерпретируемость, XAI

**Ссылка для цитирования:** Бирюков Д.Н., Дудкин А.С. Объяснимость и интерпретируемость – важные аспекты безопасности решений, принимаемых интеллектуальными системами (обзорная статья) // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 3. С. 373–386. doi: 10.17586/2226-1494-2025-25-3-373-386

# Explainability and interpretability are important aspects in ensuring the security of decisions made by intelligent systems (review article)

**Denis N. Biryukov[1], Andrey S. Dudkin[2]✉**

[1,2] Mozhaisky Military Aerospace Academy, Saint Petersburg, 197198, Russian Federation

[1] Biryukov.D.N@yandex.ru, https://orcid.org/0000-0003-1300-2470
[2] andry-ll@mail.ru✉, https://orcid.org/0000-0003-0283-9048

**Abstract**
The issues of trust in decisions made (formed) by intelligent systems are becoming more and more relevant. A systematic review of Explicable Artificial Intelligence (XAI) methods and tools aimed at bridging the gap between the complexity of neural networks and the need for interpretability of results for end users is presented. A theoretical analysis of the differences between explainability and interpretability in the context of artificial intelligence as well as their role in ensuring the security of decisions made by intelligent systems is carried out. It is shown that explainability implies the ability of a system to generate justifications understandable to humans, whereas interpretability focuses on the passive clarity of internal mechanisms. A classification of XAI methods is proposed based on their approach (preliminary/subsequent analysis: ante hoc/post hoc) and the scale of explanations (local/global). Popular tools, such as Local Interpretable Model Agnostic Explanations, Shapley Values, and integrated gradients, are considered, with an assessment of their strengths and limitations of applicability. Practical recommendations are given on the choice of methods for various fields and scenarios. The architecture of an intelligent system based on the V.K. Finn model and adapted to modern requirements for ensuring "transparency" of solutions, where the key components are the information environment, the problem solver and the intelligent interface, are discussed. The problem of a compromise between the accuracy of models and their explainability is considered: transparent models ("glass boxes", for example, decision trees) are inferior in performance to deep neural networks, but provide greater certainty of decision-making. Examples of methods and software packages for explaining and interpreting machine learning data and models are provided. It is shown that the development of XAI is associated with the integration of neuro-symbolic approaches combining deep learning capabilities with logical interpretability.

**Keywords**
artificial intelligence, neural networks, deep learning, black box, explainability, interpretability, XAI

**For citation:** Biryukov D.N., Dudkin A.S. Explainability and interpretability are important aspects in ensuring the security of decisions made by intelligent systems (review article). *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 3, pp. 373–386 (in Russian). doi: 10.17586/2226-1494-2025-25-3-373-386

## Введение

На современном этапе развития наибольшую популярность получили результаты в области моделирования глубокого обучения (Deep Learning, DL), дающие прирост в скорости и точности формируемых решений. Однако у этих результатов есть и обратная сторона — сложность моделей, что зачастую приводит к невозможности понимания того, как и за счет чего были получены те или иные решения, какие входные данные являются наиболее важными и критичными, и какие параметры модели наиболее уязвимы.

Еще предстоит много сделать, чтобы разработать подходы и инструменты, позволяющие укрепить доверие к системам и моделям DL. Возможно, в ряде случаев в обмен на понимание сути придется поплатиться скоростью и/или точностью формируемых решений. Работа в данном направлении уже ведется, но пока что наблюдается отсутствие единого тезауруса и разнообразие во взглядах на подходы к решению задач в обозначенной области исследований.

Не заканчиваются споры о том, какую систему можно считать интеллектуальной и что же такое искусственный интеллект (ИИ), но мало кто спорит с тем, что важным качеством интеллектуальной системы (ИС) является ее способность объяснить потребителю на понятном ему языке, почему и как она пришла к тому или иному решению. Только в этом случае пользователь сможет оценить результат, полученный ИС, и обоснованно принять решение. Ввиду этого видится важным рассмотреть различные взгляды на «объяснимость» и «интерпретируемость» применительно к моделям машинного обучения и к ИС в целом.

## Общая модель ИС

Вопрос, связанный с тем, какую систему можно считать интеллектуальной, остается открытым до сих пор. Ввиду этого, предлагается под ИС в общем случае понимать такую систему, которая соответствует положениям, сформулированным В.К. Финном [1, 2].

ИС обладают специфической архитектурой, допускающей определенные вариации. Схематически эта архитектура может быть представлена следующим образом [2, 3] (рисунок):

ИС = (1) решатель задач + (2) информационная среда + (3) интеллектуальный интерфейс.

(1) Решатель задач = (1.1) рассуждатель + (1.2) вычислитель + (1.3) синтезатор.

(1.1) Рассуждатель реализует синтез и взаимодействие познавательных процедур, образующих автоматизированное рассуждение, областью применимости которого является класс задач, решаемых посредством формализованной эвристики. Логическим средством формализации этой эвристики и являются рассуждения. Важно понимать, что правдоподобные рассуждения, формализующие эвристику решения задач, адекватные цели применения ИС, являются основным инструментом ее решателя задач, реализуемым в рассуждателе [1].

Существуют два типа рассуждателей.

Рассуждатели первого типа применимы к неизменяющемуся множеству исходных высказываний, характеризующих «замкнутый мир», а их логическим средством являются дедуктивные выводы (возможны варианты, когда используются неклассические логики).

Рассуждатели второго типа реализуют формализованные эвристики для решения классов задач, исходными данными которых являются изменяемые и пополняемые множества высказываний (под изменением высказываний понимается пересмотр его истинностного значения, соответствующие базы фактов называют эпистемическими). Формализованные эвристики этого типа осуществляют синтез познавательных процедур, включающих эмпирическую индукцию, основанную на установлении сходства фактов, и абдукцию, по-

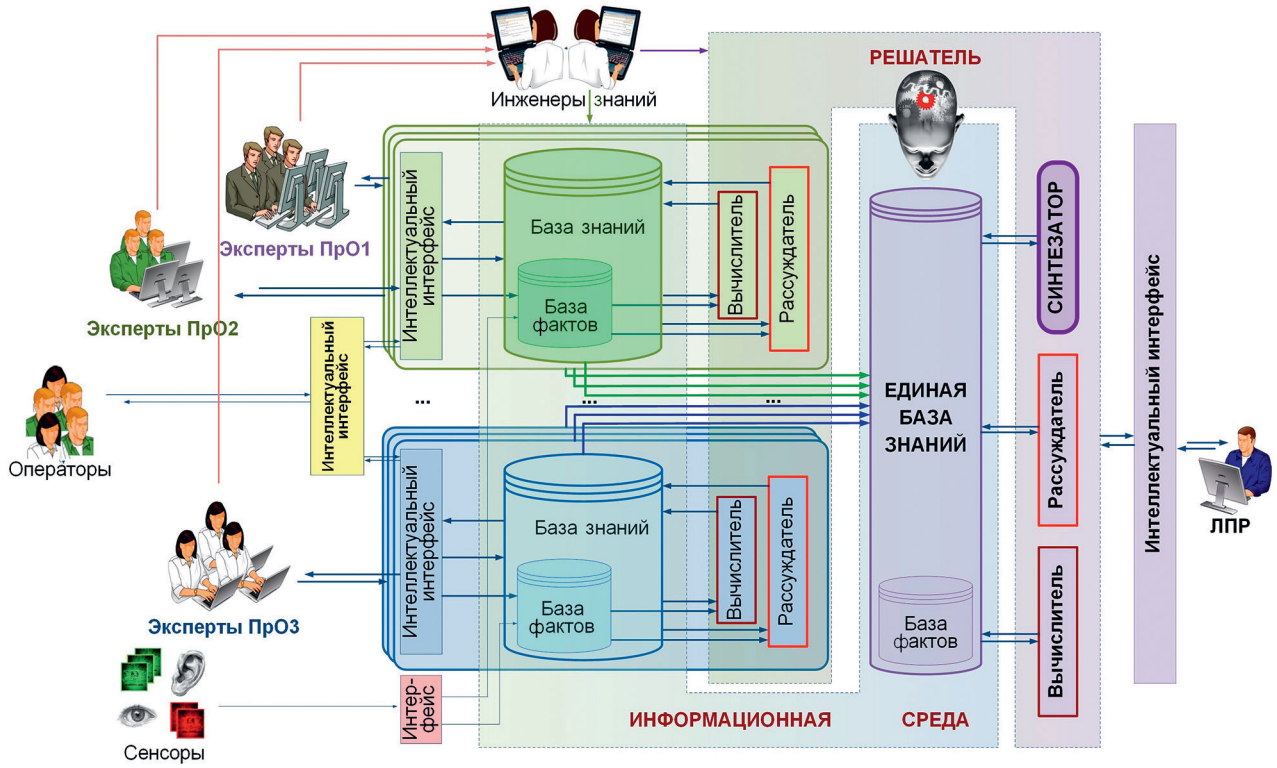


*Рисунок*. Структурно-функциональная схема интеллектуальной системы.
ПрО1 — предметная область 1; ПрО2 — предметная область 2; ПрО3 — предметная область 3; ЛПР — лицо, принимающее решения

*Figure*. Structural and functional diagram of the intelligent system.
Subject area 1 — ПрО1; subject area 2 — ПрО2; subject area 3 — ПрО3; the decision maker — ЛПР

средством которой объясняется начальное состояние базы фактов и, в случае необходимости реализуется ее пополнение (рассуждатели второго типа применимы к «открытым мирам»). Рассуждения рассуждателей второго типа называют когнитивными (правдоподобными) рассуждениями [2].

(1.2) Вычислитель применяется к числовым данным, используя численные методы, релевантные целям ИС (например, различные статистические методы анализа данных).

(1.3) Синтезатор выбирает стратегии, адекватные не только цели ИС, но и состоянию базы факторов (БФ), и результатам предыдущих применений решателя задач.

Если через Г — обозначить множество правил вывода, содержащих как правила достоверного вывода, так и правила правдоподобных выводов, посредством которых осуществляются различные правдоподобные рассуждения, а через С — множество вычислительных процедур, то их комбинирование осуществляет синтезатор.

(2) Информационная среда = (2.1) БФ + (2.2) база знаний (БЗ).

(2.1) БФ представляет рассматриваемую предметную область («замкнутый мир» или «открытый мир»; в первом случае БФ не изменяется, во втором — возможно ее пополнение в соответствии с результатами, полученными решателем задач, и желаниями пользователя ИС как человеко-машинной системы).

Наличие БФ как подсистемы ИС создает возможность осуществления машинного обучения, а следовательно, расширения БЗ.

(2.2) БЗ — подсистема представления знаний.

Обычно выделяют три типа знаний для ИС: декларативные, процедурные и концептуальные [1].

(3) Интеллектуальный интерфейс включает в себя диалог (наилучший вариант — диалог на естественном языке), демонстрацию как результатов работы ИС, так и процесса их получения, графическое представление результатов, обучение пользователя работе с ИС, поддержку интерактивного режима работы ИС.

Рассуждения и вычисления, представление знаний и интерфейс являются практическими реализациями принципов функционирования ИС. Посредством этих компонент функционирования ИС осуществляется интеллектуальная обработка данных.

Для того чтобы повысить степень доверия к ИС, как к системе, способствующей принятию решений лицом, принимающим решения (ЛПР), необходимо иметь возможность понимать то, каким образом и почему ИС пришла к тому или иному решению [4, 5]. Для этого необходимо, чтобы была возможность объяснить, каким образом решатель задач породил решение, а интеллектуальный интерфейс позволил бы интерпретировать сформированное решений. При этом следует учесть наличие разного уровня компетенций как ЛПР, так и инженеров по знаниям.

### Доверие к нейронным сетям: вызовы и стратегии обеспечения прозрачности

В последние годы в области проектирования, разработки и применения ИС наметилась тенденция к разработке «белых ящиков» — моделей, поведение которых можно объяснить и которые являются более понятными для людей. Для этого в ряде случаев предлагается разбивать ИС (решатель задач) на модули, каждый из которых может быть интерпретирован человеком, или изначально строить модели с локальной прозрачностью. Однако на данный момент «белый ящик» все еще остается достаточно нишевой идеей, а основным методом обучения всех моделей ИИ все еще является «черный ящик» (ЧЯ) (см. ИИ от Alpha Zero, ChatGPT, GPT-3, GPT-4 и др.). Быстрота и простота ЧЯ в настройке делают его более выгодным для компаний, которые ищут более быстрых результатов. Также следует отметить, что сейчас для решения большинства задач предпочтительнее использовать подходы ЧЯ, так как они показывают более высокую эффективность, чем интерпретируемые модели. Даже если интерпретируемые модели выдают хорошие результаты, все равно приходится идти на компромисс между их возможностями и пониманием. Эта проблема, по всей видимости, все еще будет актуальной достаточно продолжительное время.

Видится, что ситуацию может изменить только введение ограничений со стороны регуляторов, основанных на понимании того, что решения, принимаемые ИС, функционирующей на базе модели, относящейся к типу ЧЯ, могут содержать опасность. Предотвратить или хотя бы снизить возможную опасность может только четкое понимание того, как и почему ИС пришла к тому или иному выводу. К сожалению, подавляющее большинство популярных на данный момент инструментов и моделей не позволяют этого сделать: большинство систем ИИ работает как классические модели ЧЯ, т. е. разработчики ориентируются на вход и выход модели, не обращая внимание на процессы внутри ИС, которая занимается поиском результата. Они просто получают результаты и сравнивают их с желаемым результатом, не пытаясь понять, как ИС это достигает.

Для понимания того, как ИС (ИИ) приходит к тому или иному заключению, применяют методы и инструменты, разработанные для «объяснения» и «интерпретации» решений, однако при анализе классификаций данных методов можно заметить, что зачастую методы «объяснения» и методы «интерпретации» разделяют по-разному. Ввиду этого обозначим на начальном этапе различие между «объяснимостью» и «интерпретируемостью».

Объяснимость и интерпретируемость ИИ являются одними из важнейших качеств, которыми должен обладать современный ИИ. Объяснимость ИИ означает, что его принятие решений должно быть понятным и прозрачным для людей. Это качество особенно важно в областях, где принимаемые ИИ решения имеют серьезные последствия, например, в медицине, финансах, правоохранительных органах или в военной сфере. Если решения ИИ непонятны или необоснованны, это может привести к серьезным проблемам и негативным последствиям.

Интерпретируемость ИИ подразумевает, что люди могут понимать, как ИИ принимает свои решения и на основе каких данных. Это может быть полезно для того, чтобы проверять работу ИИ и корректировать его ошибки и недостатки. Интерпретируемость ИИ также может помочь улучшить его работу, например, путем обучения ИИ новым способам обработки данных или принятия решений. Объяснимость же означает способность объяснить, как работает ИС, используя понятный человеку язык и концепции. Это означает, что ИС должна быть способна объяснить свое решение на основе входных данных и ее параметров.

Таким образом, интерпретация и объяснимость — два разных понятия, причем объяснимость является более широким концептом, который включает в себя не только интерпретацию, но и способность объяснить работу ИС в целом.

Интерпретируемое машинное обучение, применяемое в системах ИИ, не является подмножеством Explainable Artificial Intelligence (XAI) или наоборот [6]. В то время как первая направлена на создание прогностических моделей «стеклянного ящика» («белого»/«прозрачного»), вторая стремится понять ЧЯ, используя объяснительную модель, суррогатную модель, подход атрибуции, важность релевантности или другие статистические данные. Существует опасение, что определения, подходы и методы не совпадают, что приводит к непоследовательной классификации систем и моделей DL для интерпретации и объяснения.

Несмотря на недавние успехи в XAI, до сих пор неясно, как конкретная глубокая нейронная сеть (Deep Neural Network, DNN) приходит к тому или иному решению, насколько она уверена в своем решении, можем ли мы доверять или не доверять ей, и когда ее нужно корректировать. Основная причина этого заключается в том, что DNN приводят к сложным моделям ЧЯ, т. е. когда известны только входные признаки и выходные прогнозы, что затрудняет понимание природы обучения в рамках их структуры. Вопросы, над которыми нам следует подумать: «Можем ли мы доверять решениям, принятым моделями глубокого обучения?» или «Как модель глубокого обучения, принимает решение?».

Использование DL многообещающе, поскольку оно может обрабатывать сложные наборы данных и моделировать сильно нелинейные внутренние представления данных.

Растущий интерес исследователей к XAI привел к тому, что исследовательское сообщество DL сосредоточилось на методах интерпретируемости и объяснимости DNN. Основная проблема с объяснимостью, согласно DARPA [7], состоит в том, чтобы обеспечить достаточное обоснование выводов AI/ML, чтобы пользователи знали, почему вывод был сделан или нет, чтобы пользователь знал, когда алгоритм будет успешным или неудачным, и когда можно доверять. Чтобы укрепить доверие к системам и моделям DL [8, 9], заинтересованные стороны должны получить представление

о процессе принятия решений в системе или модели, точно узнав, почему и как система, модель и алгоритм DL пришли к определенному решению.

Можно предположить, что два ключевых слова «интерпретировать» и «объяснять» никогда не бывают полностью взаимозаменяемыми в языковом употреблении, особенно в техническом и научном языках [10]. Хотя и некоторые считают [11], что нет четких различий между двумя основными концепциями XAI — интерпретацией и объяснением.

Исследования в области моделирования DL традиционно были сосредоточены на улучшении качества, алгоритмов или скорости прогнозирования модели нейронной сети [12]. В то же время DNN обычно рассматриваются как модели ЧЯ из-за их многослойной нелинейности и глубоко вложенной структуры, которые часто критикуются как непрозрачные и непонятные для человека [13]. Поскольку они обучены, а не запрограммированы напрямую, может быть трудно понять, как именно они приходят к своим решениям.

Термин ЧЯ относится к модели, которая принимает последовательность входных данных запроса и выдает соответствующие выходные данные, скрывая при этом внутренние состояния, такие как архитектура модели [14]. Существуют методы объяснения ЧЯ, которые пытаются объяснить существующие модели DL без учета внутренней структуры модели. Этот класс методов объяснения не зависит от модели и может быть легко интегрирован в модели DL, от деревьев решений до сложных нейронных сетей.

Методы объяснения ЧЯ также называют постфактум (post hoc-методами), потому что их можно использовать для опроса моделей DL после обучения и развертывания, не зная процедур обучения. Объяснения, полученные таким образом, не гарантируют, что они будут удобными для человека, полезными или осуществимыми, и могут быть опасными в решениях с высокими ставками.

Противоположностью ЧЯ является «стеклянный ящик» [15, 16], который по своей сути прозрачен. Модули или задачи интерпретации заранее интегрируются в архитектуру и алгоритмы DL называются ante hoc. Преимущество этого подхода заключается в том, что специалисты-практики могут переводить модели в DL, обнаруживать ошибки в данных и/или маркировке и, в некоторых случаях, редактировать решения моделей, когда они не соответствуют значениям или знаниям предметной области. Этот тип подхода решает проблему компромисса между точностью и интерпретацией, которая является проблемой ЧЯ и post hoc-модели.

Растет интерес к пониманию того, как эти модели DL достигают своих успешных прогнозных задач. Работа по объяснению этих сетей ЧЯ была сосредоточена на понимании того, как фиксированная модель DL приводит к конкретному прогнозу [17]. Проблемы использования моделей DL для объяснения их решений в основном связаны со следующим: отсутствием прозрачности [16, 18]; отсутствием объяснимости [19–21]; большой сложностью и потребностью в больших вычислительных ресурсах [22]; недостаточной устойчивостью к атакам со стороны противника [23] и неспособностью объяснить решения и действия так, чтобы ЛПР могли их понять [24].

Целью XAI является создание набора методов, которые создают более объяснимые модели, сохраняя при этом высокий уровень решения задач поиска, обучения, планирования и рассуждений за счет оптимизации и повышения точности. Основная идея XAI состоит в том, чтобы вскрыть алгоритм работы ЧЯ, объяснить, как ЧЯ принимал решения, и учесть выполняемые шаги и модели. Эти функции ЧЯ могут быть чрезвычайно сложными для понимания ЛПР.

Таким образом:

— цель XAI [25] состоит в том, чтобы сделать поведение системы более понятным для людей, предоставляя объяснения;
— задача XAI состоит в том, чтобы дать объяснения, которые были бы полными и интерпретируемыми [26].

Системы XAI должны быть в состоянии предоставлять исторически масштабируемые объяснения того, что система делала, что она делает сейчас и что произойдет дальше, а также раскрывать ключевую информацию, на которую она реагирует. При этом XAI предполагает, что существует много типов конечных пользователей или групп конечных пользователей.

### XAI: теоретические основы объяснимости

Поскольку каждое решение DNN представляет собой комбинацию тысяч нейронов и весов, взаимодействующих друг с другом, объяснение зачастую оказывается очень сложной задачей. Потому существует большая потребность в возможности задать системе дополнительные вопросы, чтобы понять, почему система пришла к тому или иному прогнозу/решению.

Внутренней целью алгоритмов объяснения является облегчение человеческого понимания [27]. Если ЛПР поймет объяснения, оно будут более склонно доверять системам DL и применять их. Безопасность системы DL зависит от объяснений, честности, безопасности/конфиденциальности и отладки модели [28].

Однозначного и принятого всем научным сообществом определения «объяснимости» применительно к DL и ИИ в целом пока нет. «Объяснимость — это совокупность признаков интерпретируемой области, которые в данном примере способствовали принятию решения». Объяснения могут быть полными или частичными [25]. Полностью объяснимые модели обеспечивают полное объяснение и являются прозрачными [29]. Модели, которые можно объяснить лишь частично, раскрывают только важные части процесса рассуждений.

В научном исследовании научное объяснение должно включать как минимум две части [30]: объект, подлежащий объяснению, и содержание объяснения.

Выделяют четыре типа объяснений: аналитические (дидактические) высказывания, кейсы, визуализация, альтернативный выбор.

Некоторые исследователи считают [31], что визуальные объяснения должны удовлетворять двум критери-

ям: они должны различать классы и точно описывать конкретный экземпляр изображения.

Текстовые объяснения [32] представляют собой высказывания на естественном языке, которые словесно формулируются или описываются. Текстовые пояснения могут быть основаны либо на шаблонах [27], либо на правилах [33].

Объяснения на примерах [16, 34] выделяют частные экземпляры.

Объяснения упрощением [35] включают разделение сложного пространства признаков на более простые, объяснимые области.

Атрибуция релевантности признаков [36] присваивает оценку важности каждому признаку для конкретного ввода.

Можно различать системы объяснения с самоанализом, которые объясняют, как модель определяет свой окончательный вывод, и системы объяснения с обоснованием, которые производят предложения, подробно описывающие, как визуальные свидетельства согласуются с выводом системы.

Визуальные объяснения выделяют области DNN, которые описывают интересующие классы, или, в более общем смысле, визуализируют поведение модели [37].

В общем случае хорошее объяснение должно быть, по крайней мере, верным и интерпретируемым [38].

Достоверное объяснение — точная характеристика поведения модели, в то время как интерпретируемое объяснение легко понять специалисту-человеку [39].

Существует два типа объяснений: что научилась делать ИС и как она это делает?

Вопрос «что» относится к внешним свойствам преобразований, например, является ли они инвариантными по отношению к входным данным.

Вопрос «как» относится к внутреннему функционированию, т. е. к тому, как скрытые элементы обрабатывают информацию для достижения полученного результата.

Имеют место пять общих требований к полезному объяснению системы DL [40]: точность, понятность, достаточность, низкая ресурсоемкость и эффективность.

Не без оснований считается [11], что объяснимость модели ML обычно обратно пропорциональна ее производительности. Часто модели DL являются самыми мощными, но наименее объяснимыми. Деревья решений являются наиболее объяснимыми, но с наименьшей точностью [25]. Также считается, что для моделей ЧЯ, чем выше точность прогноза, тем ниже объяснительная способность модели [26, 41].

Объяснения можно классифицировать также, как и интерпретацию по нескольким основаниям. Так, например, объяснения могут быть глобальными и локальными.

Глобальные объяснения пытаются сосредоточиться на модели в целом. В этом случае вся модель может быть объяснена, и можно проследить рассуждения от входных данных до каждого возможного результата. Этот подход позволяет получить лучшее представление обо всей модели. Это может быть, например, визуализация распределения веса в DNN или визуализация глубоких слоев сети, распространяющихся по сети.

Локальное объяснение пытается описать отдельные результаты, например, объяснить каждое предсказание. Цель состоит в том, чтобы объяснить, почему ЧЯ делает конкретное предсказание на основе локальных признаков, например пикселов. Эти объяснительные методы можно использовать для небольшой части сети, например, при рассмотрении одного фильтра в глубокой сети. Локальная объяснимость имеет дело с ситуацией, в результате которой можно понять только причины конкретного решения.

Также объяснимые методы подразделяют на post hoc- и ante hoc-методы.

Ante hoc (также известный как внутренний) объяснение (до этого события) решений модели ЧЯ заранее включается в архитектуру модели или в концептуальные ограничения. Ante hoc-системы предоставляют объяснения, которые идут от начала модели или ввода к выводу [40].

Post hoc (после события) объяснения решения модели ЧЯ могут быть даны постфактум. Post hoc-объяснения применяются для моделей, которые сложно объяснить. Потому post hoc-методы предполагают создание второй модели (эксплейнера), которая предоставляет объяснения. Это могут быть визуальные объяснения, текстовые пояснения, локальные пояснения, пояснения на примере, пояснения по упрощению и атрибуция релевантности признаков.

Основная проблема с post hoc-объяснениями заключается в том, что они могут не точно определять, как работает модель DL. Тем не менее, они предоставляют пользователям полезную информацию. Post hoc-модели позволяют объяснить процесс с точки зрения его результата, например, путем определения того, какая часть входных данных отвечает за конечный результат.

Основное различие между post hoc- и ante hoc-методами заключается в компромиссе между точностью модели и точностью объяснения.

Также методы объяснимости могут зависеть от модели DL, которую необходимо объяснить. Методы объяснимости, специфичные для модели, ограничены определенным классом моделей. Они пытаются понять модель DL, анализируя внутренние компоненты сети и то, как они взаимодействуют, исследуя функции активации или активацию обратного прохода на входе. Объяснимость, специфичная для модели, может быть связана с конкретным типом модели ЧЯ или входными данными.

Объяснители ЧЯ, не зависящие от модели, обычно требуют доступа только к функции предсказания модели, в то время как объяснители «стеклянного ящика» обычно требуют доступа к внутренним компонентам модели.

Методы, не зависящие от модели, можно разделить на три категории: упрощение модели, оценка атрибуции (релевантность признаков) и методы визуализации.

Независимые от модели объяснения (post hoc-объяснители) предназначены для применения к любой модели DL посредством получения некоторой информации из ее процедуры прогнозирования.

Чтобы повысить объяснимость моделей DL, необходимо принять во внимание, по крайней мере, несколько наиболее важных требований к объяснимости DL:

— причинность: способность метода прояснять взаимосвязь между входом и выходом в заданном контексте использования [16, 42, 43];
— корректируемость: способность метода вносить необходимые коррективы обратно в модель обучения [44];
— эффективность: способность метода поддерживать правильное принятие решений [45];
— производительность: способность метода поддерживать более быстрый наилучший вариант для принятия решений конечным пользователем [41];
— явность: способность метода немедленно и понятно давать объяснения [46, 47];
— добросовестность: способность метода давать объяснения, указывающие на истинные релевантные признаки [46, 47];
— достоверность: способность метода согласовываться с отображением ввода-вывода глубинной модели [48];
— информативность: способность метода предоставлять полезную информацию конечному пользователю посредством его вывода [16, 49];
— стабильность: согласованность метода предоставления аналогичных объяснений для аналогичных или соседних входных данных [46, 47];
— переносимость: способность метода обобщать и переносить новые знания на незнакомые ситуации [16, 48, 49];
— надежность: постоянство метода, позволяющего выдерживать небольшие возмущения входных данных, которые не меняют прогноз модели [46, 47, 50];
— убедительность: способность метода убеждать пользователей выполнять определенные действия [46, 47];
— исследуемость: способность метода проверять процесс обучения, который не сходится или не обеспечивает приемлемой производительности [46, 47].

### Интерпретируемость моделей: методы и критерии оценки

По сравнению с системами ИИ, ориентированными на задачи, системы XAI предназначены для выполнения конкретных задач, которые приводят к объяснениям или созданию объяснительных моделей для решения проблем ЧЯ или систем Interface-mock-live, где интерпретации или понятные модели должны быть «стеклянными ящиками». Это может способствовать созданию доверительного ИИ и реализации удобных для человека методов ИИ, ориентированных на прозрачность, справедливость и подотчетность модели [51].

Некоторые авторы [17] определяют интерпретацию как «отображение абстрактного понятия, например, предсказанного класса, в область, которую человек может понять». Другие [25] определяют цель интерпретируемости как описание внутренностей системы способом, понятным человеку («интерпретируемость определяется как способность объяснить или передать значение в понятных человеку терминах»). Не существует математического определения интерпретируемости. Однако есть взгляды [52] на то, как можно измерить интерпретируемость. Нематематическое определение [43] гласит: «интерпретируемость — степень, в которой человек может понять причину решения».

Интерпретируемость относится к пассивному свойству модели и уровню, на котором конкретная модель имеет смысл для человека. Объяснимость же, напротив, можно рассматривать как активную сторону модели, которая относится к раскрытию внутренних функций модели. Другими словами, интерпретируемость — результат модели DL, а объяснимость — инструмент, который должен «открыть» этот результат. Чем выше интерпретируемость модели DL, тем легче кому-то понять, почему были сделаны те или иные решения или прогнозы. Модель более интерпретируема, чем другая модель, если ее решения легче понять человеку, чем решения последней. Отсюда можно сделать вывод, что понятие интерпретируемости имеет более широкую перспективу по сравнению с понятием объяснимости.

Выделяют [52] глобальную интерпретируемость и локальную интерпретируемость. Также важным аспектом является то, сколько времени пользователь должен или может потратить на понимание объяснения. Интерпретируемость модели, конечно, зависит и от опыта пользователя. Наряду с интерпретируемостью человеком, которая помогает людям понимать машины, существует также интерпретируемость машин. Это относится к тому, как машины «понимают» решения друг друга в рамках многоагентных ИС. Таким образом, глядя на цель DL с точки зрения моделирования, интерпретируемость тесно связана с ключевыми аудиториями конечных пользователей с точки зрения системы, а соответственно зависит от возможностей, заложенных в интеллектуальный интерфейс, через который ИС взаимодействует с ЛПР и инженерами по знаниям.

Глобальная интерпретируемость связана с пониманием того, как общая модель DL принимает решения, которые являются результатом прогноза. Это можно сделать, исследуя сложную структуру и параметры всей модели, какие входные шаблоны фиксируются и как они преобразуются для получения выходных данных [53]. Локальная интерпретируемость исследует локально причины поведения модели DL с учетом конкретного прогноза. Это достигается путем идентификации сопоставления каждой функции в конкретном входе для прогноза, сделанного DNN [39].

Считается [16], что интерпретируемость не является монолитным понятием, а отражает несколько различных идей и носит квазинаучный характер.

Рассмотрим основные концепции, описывающие интерпретируемость.

Понятность — свойство модели позволять человеку понимать ее работу без выяснения ее внутренней структуры или внутренних операций, с помощью которых модель обрабатывает данные [54]. Это свойство относится к ответу на вопрос: как работает модель DL? Естественно, что аудитория является краеуголь-

ным камнем XAI, когда дело доходит до понимания модели [51].

Когда дело доходит до понимания DNN, имеют дело с двумя взглядами на «понятность»:

— механистическое понимание (какой механизм сеть использует для решения проблемы или реализации функции);
— функциональное понимание (как сеть связывает входные переменные с выходными переменными).

Понятность является наиболее важным понятием в интерпретируемом DL.

Прозрачность обращается к свойству системы, чтобы объяснить, как она функционирует, даже если она ведет себя неожиданно. Очень важная проблема, которую приходится решать, заключается в том, что основная истина интерпретируемости зависит от полной прозрачности того, как модель DL приходит к своему решению [55].

Прозрачность можно дифференцировать [56] на прозрачность модели, прозрачность дизайна (построения), и алгоритмическую прозрачность.

Некоторые авторы [16] разделяют прозрачность модели по свойству понятности модели на три уровня: уровень всей модели, уровень отдельных компонентов и уровень алгоритмов.

В некоторых исследованиях интерпретируемость понимается как необходимое условие доверия [16]. Это указывает на то, что основная идея интерпретируемости состоит в том, чтобы помочь людям понять задачи прогнозирования моделей DL и доверять им.

Также интерпретируемость можно рассматривать как состоящую из трех категорий: интерпретируемость данных (какие измерения данных наиболее важны для задачи); интерпретируемость модели (как модель, принадлежащая к определенной категории, обычно выглядит в соответствии с моделью); интерпретируемость предсказания (объясняет, почему определенный шаблон $x$ был классифицирован определенным образом $f(x)$).

Исходя из вышеизложенного, можно сосредоточиться на конкретных понятиях, описывающих интерпретируемость: понимание, прозрачность (доверие) и решение, а определение интерпретируемости можно дать следующим образом: «Интерпретируемость означает способность человека понимать и доверять результатам, полученным моделью глубокого обучения».

### Инструментарий, применяемый в области объяснения решений, принимаемых ИИ

Приведем перечень программных реализаций частных методов, применяемых в области объяснения решений, принимаемых ИИ, а также ряд программных пакетов (таблица), которые могут помочь исследователям при решении задач объяснения и интерпретации наборов данных и моделей машинного обучения.

*Local Interpretable Model Agnostic Explanations (LIME)*

Исходный код: https://github.com/marcotcr/lime.

Лицензия: BSD 2-Clause «Simplified» License.

Объяснения, не зависящие от локальной интерпретируемой модели. Рассматривая модели машинного обучения как функции ЧЯ, методы объяснения, не зависящие от модели, обычно имеют доступ только к выходным данным модели.

*Anchors*

Исходный код: https://github.com/marcotcr/anchor.

Лицензия: BSD 2-Clause «Simplified» License.

Основная идея заключается в том, что отдельные предсказания любой модели классификации ЧЯ объясняются путем нахождения решающего правила, которое в достаточной степени «закрепляет» предсказание — отсюда и название «якоря». Результирующие объяснения представляют собой правила принятия решений в форме операторов IF-THEN, которые определяют области в пространстве объектов

*GraphLIME*

Исходный код: https://github.com/WilliamCCHuang/GraphLIME.

Лицензия: MIT license.

Метод, который использует основную идею LIME, но не является линейным. Он применяется к особому типу архитектуры нейронных сетей, а именно к графовым нейронным сетям. Эти модели могут обрабатывать неевклидовы данные, поскольку они организованы в виде графовой структуры.

*Layer-wise Relevance Propagation (LRP)*

Исходный код: https://github.com/chr5tphr/zennit.

Лицензия: GPL-3.0, LGPL-3.0.

Исходный код: https://github.com/albermax/innvestigate.

Лицензия: The copyright in this software is being made available under the BSD License, included below. This software is subject to other contributor rights, including patent rights, and no such rights are granted under this license. All rights reserved.

Послойное распространение релевантности (LRP) — метод объяснения, основанный на распространении, т. е. он требует доступа к внутренним элементам модели (топологии, весам, активациям и т. д.).

*Deep Taylor Decomposition (DTD)*

Исходный код: https://github.com/chr5tphr/zennit.

Лицензия: GPL-3.0, LGPL-3.0.

Метод объяснения на основе распространения, который объясняет решения нейронной сети путем декомпозиции. Он перераспределяет значение функции (т. е. выходные данные нейронной сети) на входные переменные послойно, используя при этом математический инструмент (первого порядка) разложения Тейлора для определения пропорции или релевантности, присвоенной элементам нижнего уровня в процессе перераспределения (т. е. их соответствующие вклады). Этот подход тесно связан с методом LRP.

*Prediction Difference Analysis (PDA)*

Исходный код: https://github.com/lmzintgraf/DeepVis-PredDiff.

Лицензия: MIT license.

Метод основан на идее, когда для данного прогноза каждой входной функции присваивается значение релевантности по отношению к классу. Идея PDA заключается в том, что релевантность функции может быть оценена измерением того, как изменяется прогнозирование, когда признак неизвестен.

*Testing with Concept Activation Vectors (TCAV)*

Исходный код: https://github.com/tensorflow/tcav.

Лицензия: Apache-2.0 license.

Целью подхода является количественная оценка того, насколько сильно концепция, такая как цвет, влияет на классификацию. Чтобы рассчитать такой TCAV, сначала необходимо собрать и объединить два набора данных: набор данных, содержащий изображения, представляющие концепцию, и набор данных, состоящий из изображений, в которых эта концепция отсутствует.

*Explainable Graph Neural Networks (XGNN)*

Исходный код: https://github.com/divelab/DIG/tree/dig-stable/benchmarks/xgraph.

Лицензия: GPL-3.0 license.

Является методом post hoc, который работает на уровне модели, что означает, что он не стремится предоставлять объяснения на уровне отдельных примеров (метод объяснения изобретен специально для задачи классификации графов).

*Shapley Values (SHAP)*

Исходный код: https://github.com/slundberg/shap.

Лицензия: MIT license.

Методы в этом семействе связаны с объяснениями модели функции в какой-то отдельной точке. Значения SHAP симметричны. Это означает, что, если две переменные оказывают одинаковое влияние на поведение модели, например, потому, что они принимают одинаковые значения, они получат одинаковые атрибуты.

*Asymmetric Shapley Values (ASV)*

Исходный код: https://github.com/nredell/shapFlex.

Лицензия: MIT license.

ASV позволяют использовать дополнительные знания о причинно-следственных связях между переменными в процессе объяснения модели. Причинно-следственная связь, описанная в форме причинного графика, позволяет перераспределять атрибуцию переменных таким образом, чтобы исходные переменные имели большую атрибуцию, оказывая влияние как на другие зависимые переменные, так и на предсказания модели.

*Break-Down*

Исходный код: https://github.com/ModelOriented/DALEX.

Лицензия: GPL-3.0 license.

Если в модели есть взаимодействия, разный порядок переменных может привести к разным вкладам. Значения SHAP усредняются по всем возможным порядкам, что приводит к аддитивным вкладам и игнорированию взаимодействий. Альтернативой является анализ различных порядков, чтобы обнаружить, когда одна переменная имеет разные вклады в зависимости от того, какие другие переменные предшествуют ей. Метод Break-Down анализирует различные порядки для идентификации и визуализации взаимодействий в модели.

*Shapley Flow*

Исходный код: https://github.com/nathanwang000/Shapley-Flow.

Лицензия: не указано.

Shapley Flow также позволяет использовать структуру зависимостей между переменными в процессе объяснения. Как и в ASV, взаимосвязь описывается причинно-следственным графом. Однако, в отличие от ASV и других методов, атрибуция присваивается не узлам (переменным), а ребрам (отношениям между переменными). Ребро на графике имеет значение, если его удаление изменит прогнозы модели.

*Textual Explanations of Visual Models*

Исходный код: https://github.com/LisaAnne/ECCV2016.

Лицензия: All Rights Reserved. Permission to use, copy, modify, and distribute this software and its documentation for educational, research, and not-for-profit purposes, without fee and without a signed licensing agreement, is hereby granted, provided that the above copyright notice, this paragraph and the following two paragraphs appear in all copies, modifications, and distributions.

Генерация текстовых описаний изображений решается несколькими моделями машинного обучения, которые содержат как часть, обрабатывающую входные изображения — обычно сверточную нейронную сеть, так и ту, которая изучает адекватную текстовую последовательность, обычно рекуррентную нейронную сеть. Эти две части взаимодействуют для создания описательных предложений с изображениями, что предполагает успешное выполнение задачи классификации. Важно отметить, что простое описание содержимого изображения не эквивалентно объяснению процесса принятия решений в модели нейронной сети.

*Integrated Gradients*

Исходный код: https://github.com/ankurtaly/Integrated-Gradients.

Лицензия: не указано.

Метод Integrated Gradients основан на двух фундаментальных аксиомах: чувствительности и инвариантности реализации. Чувствительность означает, что ненулевые атрибуты присваиваются каждому входному сигналу и базовой линии, которые отличаются по одному признаку, но имеют разные прогнозы. Инвариантность реализации означает, что если две модели ведут себя идентично/функционально эквивалентны, то атрибуции должны быть идентичными. Хотя эти две аксиомы звучат очень естественно, оказывается, что многие методы атрибуции не обладают этими свойствами. В частности, когда модель имеет сглаженные прогнозы для конкретной интересующей точки, градиент в интересующей точке обнуляется и не несет информации, полезной для объяснения.

*Causal Models*

Исходный код: нет.

Лицензия: нет.

Модель Causal Models можно считать расширением байесовских моделей среды RL с использованием контрфактуалов. Она учитывает события, которые могут произойти, или состояния среды, которые могут быть достигнуты при различных действиях, предпринятых агентом RL. В конечном счете цель любого агента RL — максимизировать долгосрочное вознаграждение; объяснение обеспечивает причинно-следственные связи до тех пор, пока не будет достигнуто состояние получения вознаграждения.

*Таблица.* Программные пакеты, которые могут применяться при решении задач объяснения/интерпретации наборов данных и моделей машинного обучения

*Table.* Software packages that can be used to solve problems of explaining/interpreting data sets and machine learning models

| № | Инструмент | Исходный код | Лицензия |
|---|---|---|---|
| 1 | AI Fairness 360 (AIF360) | https://github.com/Trusted-AI/AIF360 | Apache-2.0 license |
| 2 | AI Explainability 360 (AIX360) | https://github.com/Trusted-AI/AIX360 | Apache-2.0 license |
| 3 | Alibi Explain | https://github.com/SeldonIO/alibi | Apache-2.0 license |
| 4 | Analysis by synthesis (ABS) | https://github.com/bethgelab/AnalysisBySynthesis | Apache-2.0 license |
| 5 | Captum | https://github.com/pytorch/captum | BSD-3-Clause license |
| 6 | DALEX | https://github.com/ModelOriented/DALEX | GPL-3.0 license |
| 7 | DeepExplain | https://github.com/marcoancona/DeepExplain | MIT license |
| 8 | Deep visualization tool | https://github.com/yosinski/deep-visualization-toolbox | MIT license |
| 9 | ELI5 | https://github.com/TeamHG-Memex/eli5 | MIT license |
| 10 | explainX | https://github.com/explainX/explainx | MIT license |
| 11 | FAT Forensics | https://github.com/fat-forensics/fat-forensics | BSD-3-Clause license |
| 12 | InterpretML | https://github.com/interpretml/interpret | MIT license |
| 13 | iNNvestigate | https://github.com/albermax/innvestigate | в соответствии с BSD с указанными дополнениями |
| 14 | H2O.ai | https://github.com/h2oai/mli-resources | Не указано |
| 15 | L2X | https://github.com/Jianbo-Lab/L2X | Не указано |
| 16 | Rectified gradient | https://github.com/1202kbs/Rectified-Gradient | Не указано |
| 17 | Saliency relevance propagation | https://github.com/Hey1Li/Salient-Relevance-Propagation | MIT license |
| 18 | Sensitivity analysis library (SALib) | https://github.com/SALib/SALib | MIT license |
| 19 | Skater | https://github.com/oracle/Skater | UPL-1.0 license |
| 20 | tfexplain | https://github.com/sicara/tf-explain | MIT license |
| 21 | treeinterpreter | https://pypi.org/project/treeinterpreter/ | BSD License (BSD) |
| 22 | XAI | https://github.com/EthicalML/xai | MIT license |

*Meaningful Perturbations*

Исходный код: https://github.com/ruthcfong/perturb_explanations.

Лицензия: не указано.

Метод может рассматриваться как независимый от модели метод объяснения, основанный на возмущениях. Таким образом, объяснение вычисляется исключительно на основе реакции модели на возмущенную (или закрытую) входную выборку.

*EXplainable Neural-Symbolic Learning (X-NeSyL)*

Исходный код: https://github.com/JulesSanchez/X-NeSyL.

Лицензия: не указано.

Исходный код: https://github.com/JulesSanchez/MonuMAIAutomaticStyleClassification.

Лицензия: GPL-3.0 license.

Нейро-символические методы включают в себя предварительные человеческие знания для различных задач, таких как изучение концепций, и в то же время они дают более понятный результат, такой как математические уравнения или языки, специфичные для предметной области.

**Заключение**

При проектировании и создании интеллектуальных систем важно тщательно продумать, как представлять факты и знания в системе, какие механизмы обработки реализовывать в решателе, каким должен быть интеллектуальный интерфейс взаимодействия с лицами, принимающими решение, какова их квалификация, какой язык взаимодействия инженеров по знаниям с интеллектуальной системой выбрать и т. п. Ответ на эти вопросы во многом обусловит потенциальную возможность реализации подсистемы, способной объяснить конечному пользователю ход принятия решения и донести эту информацию на доступном ему языке.

Следует определять допустимую точность формируемых решений и время, отводимое на их порождение, так как в ряде случаев, учитывая данные факторы, можно отдать предпочтение менее точным и оперативным системам, но более «прозрачным», т. е. основанным на применении моделей из класса «белых ящиков», что позволит повысить степень объяснимости их работы. В случае невозможности отказаться от применения моделей типа «черного ящика», нужно понимать, какие методы и инструменты их «объяснимости» и «интер-

претируемости» существуют и какие применимы для той или иной модели.

Отсутствие возможности понять, как та или иная интеллектуальная система приходит к тому или иному решению, таит опасности, связанные как с безопасностью самой модели, так и с безопасностью реализации решений, ею порождаемых. Ввиду этого объяснимость и интерпретируемость искусственного интеллекта являются одними из важнейших качеств, которыми должен обладать современный искусственный интеллект. Чтобы объяснить, как искусственный интеллект принимает определенные решения, используются методы и инструменты, предназначенные для их интерпретации и объяснения.

Несмотря на достижения в области разработки объяснимого искусственного интеллекта, по-прежнему не понятно, как глубокие нейронные сети принимают решения, насколько они уверены в своих выводах, и когда следует корректировать их решения, чтобы было можно доверять им. Если же пользователь поймет объяснения, он будет более склонен доверять системам глубокого обучения и применять их.

Интерпретация же относится к процессу понимания того, как интеллектуальная система (и нейронная сеть — в частности) обрабатывает информацию и принимает решения на основе предоставленных ей данных. Целью интерпретации является выявление взаимосвязей между входными данными и выходом, а также выявление слабых мест и возможных ошибок.

С другой стороны, объяснимость означает способность объяснить, как работает нейронная сеть и интеллектуальная система в целом, используя понятный человеку язык и концепции. Это означает, что интеллектуальная система должна быть способна объяснить свое решение на основе входных данных и параметров, используемых для обучения.

Однозначного и принятого всем научным сообществом определения «объяснимости» применительно к глубокому обучению и искусственному интеллекту в целом пока нет, но не вызывает сомнения, что объяснение должно быть точным, понятным, достаточным и не требовать большого количества ресурсов для его осуществления.

Таким образом, интерпретация и объяснимость — два разных понятия, причем объяснимость является более широким концептом, который включает в себя не только интерпретацию, но и способность объяснить работу нейронной сети и интеллектуальной системы в целом.

**Литература**


1. Финн В.К. Об интеллектуальном анализе данных // Новости искусственного интеллекта. 2004. № 3. С. 3–18.
2. Финн В.К. Искусственный интеллект: Идейная база и основной продукт // IX Национальная конференция «Искусственный интеллект-2004». 2004. Т. 1. С. 11–20.
3. Бирюков Д.Н., Ломако А.Г., Ростовцев Ю.Г. Облик антиципирующих систем предотвращения рисков реализации киберугроз // Труды СПИИРАН. 2015. № 2(39). С. 5–25.
4. Бирюков Д.Н., Ломако А.Г. Денотационная семантика контекстов знаний при онтологическом моделировании предметных областей конфликта // Труды СПИИРАН. 2015. № 5(42). С. 155–179.
5. Бирюков Д.Н., Ломако А.Г., Жолус Р.Б. Пополнение онтологических систем знаний на основе моделирования умозаключений с учетом семантики ролей // Труды СПИИРАН. 2016. № 4(47). С. 105–129. https://doi.org/10.15622/sp.47.6
6. Namatēvs I., Sudars K., Dobrājs A. Interpretability versus explainability: classification for understanding deep learning systems and models // Computer Assisted Methods in Engineering and Science. 2022. V. 29. N 4. P. 297–356. http://dx.doi.org/10.24423/cames.518
7. Gunning D. Explainable artificial intelligence (XAI). 2017. [Электронный ресурс]. URL: https://nsarchive.gwu.edu/sites/default/files/documents/5794867/National-Security-Archive-David-Gunning-DARPA.pdf (дата обращения: 21.10.2024).
8. Varshney K.R. Trustworthy machine learning and artificial intelligence // XRDS: Crossroads, The ACM Magazine for Students. 2019. V. 25. N 3. P. 26–29. https://doi.org/10.1145/3313109
9. Doshi-Velez F., Kim B., Towards a rigorous science of interpretable machine learning // arXiv. 2017. arXiv:1702.08608v2. https://doi.org/10.48550/arXiv.1702.08608
10. Yuan W., Liu P., Neubig G. Can we automate scientific reviewing? // arXiv. 2021. arXiv.2102.00176. https://doi.org/10.48550/arXiv.2102.00176
11. Arya V., Bellamy R.K.E., Chen P.-Yu., Dhurandhar A., Hind M., Hoffman S.C., Houde S., Liao V.Q., Luss R., Mojsilović A., et al. One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques // arXiv. 2019. arXiv.1909.03012. https://doi.org/10.48550/arXiv.1909.03012
12. Samek W., Wiegand T., Müller K.-R. Explainable artificial intelligence: Understanding, visualizing and interpreting deep


**References**


1. Finn V.K. On intelligent data analysis. *Novosti Iskusstvennogo Intellekta*, 2004, no. 3, pp. 3–18. (in Russian)
2. Finn V.K. Artificial Intelligence: The Idea Base and the Main Product. *Proc. of the 9th National Conference on Artificial Intelligence*, 2004, vol. 1, pp. 11–20. (in Russian)
3. Biryukov D.N., Lomako A.G., Rostovtsev Yu.G. The appearance of anticipating cyber threats risk prevention systems. *SPIIRAS Proceedings*, 2015, no. 2(39), pp. 5–25. (in Russian)
4. Biryukov D.N., Lomako A.G. Denotational semantics of knowledge contexts in ontological modeling of subject domains of the conflict. *SPIIRAS Proceedings*, 2015, no. 5(42), pp. 155–179. (in Russian)
5. Biryukov D.N., Lomako A.G., Zholus R.B. Ontological knowledge system completion based on modeling inferences taking into account role semantics. *SPIIRAS Proceedings*, 2016, no. 4(47), pp. 105–129. (in Russian). https://doi.org/10.15622/sp.47.6
6. Namatēvs I., Sudars K., Dobrājs A. Interpretability versus explainability: classification for understanding deep learning systems and models. *Computer Assisted Methods in Engineering and Science*, 2022, vol. 29, no. 4, pp. 297–356. http://dx.doi.org/10.24423/cames.518
7. Gunning D. *Explainable artificial intelligence (XAI)*, 2017. Available at: https://nsarchive.gwu.edu/sites/default/files/documents/5794867/National-Security-Archive-David-Gunning-DARPA.pdf (accessed: 21.10.2024).
8. Varshney K.R. Trustworthy machine learning and artificial intelligence. *XRDS: Crossroads, The ACM Magazine for Students*, 2019, vol. 25, no. 3, pp. 26–29. https://doi.org/10.1145/3313109
9. Doshi-Velez F., Kim B., Towards a rigorous science of interpretable machine learning. *arXiv*, 2017, arXiv:1702.08608v2. https://doi.org/10.48550/arXiv.1702.08608
10. Yuan W., Liu P., Neubig G. Can we automate scientific reviewing? *arXiv*, 2021, arXiv.2102.00176. https://doi.org/10.48550/arXiv.2102.00176
11. Arya V., Bellamy R.K.E., Chen P.-Yu., Dhurandhar A., Hind M., Hoffman S.C., Houde S., Liao V.Q., Luss R., Mojsilović A., et al. One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques. *arXiv*, 2019, arXiv.1909.03012. https://doi.org/10.48550/arXiv.1909.03012
12. Samek W., Wiegand T., Müller K.-R. Explainable artificial intelligence: Understanding, visualizing and interpreting deep

learning models // arXiv. 2017. arXiv.1708.08296. https://doi.org/10.48550/arXiv.1708.08296
13. Angelov P., Soares E. Towards explainable deep neural networks (xDNN) // Neural Net- works. 2020. V. 130. P. 185–194. https://doi.org/10.1016/j.neunet.2020.07.010
14. Oh S.J., Augustin M., Schiele B., Fritz M. Towards reverse-engineering black-box neural networks // arXiv. 2018. arXiv.1711.01768. https://doi.org/10.48550/arXiv.1711.01768
15. Rai A. Explainable AI: From black box to glass box // Journal of the Academy of Marketing Science. 2020. V. 48. N 1. P. 137–141. https://doi.org/10.1007/s11747-019-00710-5
16. Lipton Z.C. The mythos of model interpretability // arXiv. 2017. arXiv.1606.03490. https://doi.org/10.48550/arXiv.1606.03490
17. Montavon G., Samek W., Müller K.-R. Methods for interpreting and understanding deep neural networks // Digital Signal Processing. 2018. V. 73. P. 1–15. https://doi.org/10.1016/j.dsp.2017.10.011
18. Mascharka D., Tran P., Soklaski R., Majumdar A. Transparency by design: Closing the gap between performance and interpretability in visual reasoning // arXiv. 2018. arXiv:1803.05268. https://doi.org/10.48550/arXiv.1803.05268
19. Beaudouin V., Bloch I., Bounie D., Clémençon S., d'Alché-Buc F., Eagan J., Maxwell W., Mozharovskyi P., Parekh J. Flexible and context-specific AI explainability: A multidisciplinary approach // arXiv. 2020. arXiv:2003.07703v1. https://doi.org/10.48550/arXiv.2003.07703
20. Sokol K., Flach P. Explainability fact sheets: A framework for systematic assessment of explainable approaches // Proc. of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*'20). 2020. P. 56–67. https://doi.org/10.1145/3351095.3372870
21. Xu F., Uszkoreit H., Du Y., Fan W., Zhao D., Zhu J. Explainable AI: A brief survey on history, research areas, approaches and challenges // Lecture Notes in Computer Science. 2019. V. 11839. P. 563–574. https://doi.org/10.1007/978-3-030-32236-6_51
22. Thompson N.C., Greenwald K., Lee K., Manso G.F. The computational limits of deep learning // arXiv. 2020. arXiv:2007.05558. https://doi.org/10.48550/arXiv.2007.05558
23. DuSell B., Chiang D. Learning context-free languages with nondeterministic stack RNNs // Proc. of the 24th Conference on Computational Natural Language Learning. 2020. P. 507–519. https://doi.org/10.18653/v1/2020.conll-1.41
24. Flambeau J.K.F., Norbert T. Simplifying the explanation of deep neural networks with sufficient and necessary feature-sets: case of text classification // arXiv. 2020. arXiv:2010.03724v2. https://doi.org/10.48550/arXiv.2010.03724
25. Gunning D., Stefik M., Choi J., Miller T., Stumpf S., Yang G.-Z. XAI — Explainable artificial intelligence // Science Robotics. 2019. V. 4. N 37. P. eaay7120. https://doi.org/10.1126/scirobotics.aay7120
26. Gilpin L.H., Bau D., Yuan B.Z., Bajwa A., Specter M., Kagal L. Explaining explanations: an overview of interpretability of machine learning // arXiv. 2018. arXiv:1806.00069v3. https://doi.org/10.48550/arXiv.1806.00069
27. Alber M. Software and application patterns for explanation methods // arXiv. 2019. arXiv:1904.04734v1. https://doi.org/10.48550/arXiv.1904.04734
28. Zhao X., Banks A., Sharp J., Robu V., Flynn D., Fisher M., Huang X. A safety framework for critical systems utilising deep neural networks // arXiv. 2020. arXiv:2003.05311v3. https://doi.org/10.1007/978-3-030-54549-9_16
29. Weller A. Transparency: Motivations and challenges // Lecture Notes in Computer Science. 2019. V. 11700. P. 23–40. https://doi.org/10.1007/978-3-030-28954-6_2
30. Raghu M., Schmidt E. A survey of deep learning for scientific discovery // arXiv. 2020. arXiv:2003.11755v1. https://doi.org/10.48550/arXiv.2003.11755
31. Hendricks L.A., Rohrbach A., Schiele B., Darrell T., Akata Z. Generating visual explanations with natural language // Applied AI Letters. 2021. V. 2. N 4. P. e55. https://doi.org/10.1002/ail2.55
32. Kaplan J. McCandlish S., Henighan T., Brown T.B., Chess B., Child R., Gray S., Radford A., Wu J., Amodei D. Scaling laws for neural language models // arXiv. 2020. arXiv:2001.08361v1. https://doi.org/10.48550/arXiv.2001.08361
33. Towell G.G., Shavlik J.W. Extracting refined rules from knowledge-based neural networks // Machine Learning. 1993. V. 13. N 1. P. 71–101. https://doi.org/10.1007/bf00993103
34. Molnar C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Christoph Molnar, 2025. 392 p.

learning models. *arXiv*, 2017, arXiv.1708.08296. https://doi.org/10.48550/arXiv.1708.08296
13. Angelov P., Soares E. Towards explainable deep neural networks (xDNN). *Neural Net-works*, 2020, vol. 130, pp. 185–194. https://doi.org/10.1016/j.neunet.2020.07.010
14. Oh S.J., Augustin M., Schiele B., Fritz M. Towards reverse-engineering black-box neural networks. *arXiv*, 2018, arXiv.1711.01768. https://doi.org/10.48550/arXiv.1711.01768
15. Rai A. Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, 2020, vol. 48, no. 1, pp. 137–141. https://doi.org/10.1007/s11747-019-00710-5
16. Lipton Z.C. The mythos of model interpretability. *arXiv*, 2017, arXiv.1606.03490. https://doi.org/10.48550/arXiv.1606.03490
17. Montavon G., Samek W., Müller K.-R. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 2018, vol. 73, pp. 1–15. https://doi.org/10.1016/j.dsp.2017.10.011
18. Mascharka D., Tran P., Soklaski R., Majumdar A. Transparency by design: Closing the gap between performance and interpretability in visual reasoning. *arXiv*, 2018, arXiv:1803.05268. https://doi.org/10.48550/arXiv.1803.05268
19. Beaudouin V., Bloch I., Bounie D., Clémençon S., d'Alché-Buc F., Eagan J., Maxwell W., Mozharovskyi P., Parekh J. Flexible and context-specific AI explainability: A multidisciplinary approach. *arXiv*, 2020, arXiv:2003.07703v1. https://doi.org/10.48550/arXiv.2003.07703
20. Sokol K., Flach P. Explainability fact sheets: A framework for systematic assessment of explainable approaches. *Proc. of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*'20)*, 2020, pp. 56–67. https://doi.org/10.1145/3351095.3372870
21. Xu F., Uszkoreit H., Du Y., Fan W., Zhao D., Zhu J. Explainable AI: A brief survey on history, research areas, approaches and challenges. *Lecture Notes in Computer Science*, 2019, vol. 11839, pp. 563–574. https://doi.org/10.1007/978-3-030-32236-6_51
22. Thompson N.C., Greenwald K., Lee K., Manso G.F. The computational limits of deep learning. *arXiv*, 2020, arXiv:2007.05558. https://doi.org/10.48550/arXiv.2007.05558
23. DuSell B., Chiang D. Learning context-free languages with nondeterministic stack RNNs. *Proc. of the 24th Conference on Computational Natural Language Learning*, 2020, pp. 507–519. https://doi.org/10.18653/v1/2020.conll-1.41
24. Flambeau J.K.F., Norbert T. Simplifying the explanation of deep neural networks with sufficient and necessary feature-sets: case of text classification. *arXiv*. 2020, arXiv:2010.03724v2. https://doi.org/10.48550/arXiv.2010.03724
25. Gunning D., Stefik M., Choi J., Miller T., Stumpf S., Yang G.-Z. XAI — Explainable artificial intelligence. *Science Robotics*, 2019, vol. 4, no. 37, pp. eaay7120. https://doi.org/10.1126/scirobotics.aay7120
26. Gilpin L.H., Bau D., Yuan B.Z., Bajwa A., Specter M., Kagal L. Explaining explanations: an overview of interpretability of machine learning. *arXiv*, 2018, arXiv:1806.00069v3. https://doi.org/10.48550/arXiv.1806.00069
27. Alber M. Software and application patterns for explanation methods. *arXiv*, 2019, arXiv:1904.04734v1. https://doi.org/10.48550/arXiv.1904.04734
28. Zhao X., Banks A., Sharp J., Robu V., Flynn D., Fisher M., Huang X. A safety framework for critical systems utilising deep neural networks. *arXiv*, 2020, arXiv:2003.05311v3. https://doi.org/10.1007/978-3-030-54549-9_16
29. Weller A. Transparency: Motivations and challenges. *Lecture Notes in Computer Science*, 2019, vol. 11700, pp. 23–40. https://doi.org/10.1007/978-3-030-28954-6_2
30. Raghu M., Schmidt E. A survey of deep learning for scientific discovery. *arXiv*, 2020, arXiv:2003.11755v1. https://doi.org/10.48550/arXiv.2003.11755
31. Hendricks L.A., Rohrbach A., Schiele B., Darrell T., Akata Z. Generating visual explanations with natural language. *Applied AI Letters*, 2021, vol. 2, no. 4, pp. e55. https://doi.org/10.1002/ail2.55
32. Kaplan J. McCandlish S., Henighan T., Brown T.B., Chess B., Child R., Gray S., Radford A., Wu J., Amodei D. Scaling laws for neural language models. *arXiv*, 2020, arXiv:2001.08361v1. https://doi.org/10.48550/arXiv.2001.08361
33. Towell G.G., Shavlik J.W. Extracting refined rules from knowledge-based neural networks. *Machine Learning*, 1993, vol. 13, no. 1, pp. 71–101. https://doi.org/10.1007/bf00993103

35. Kim S., Jeong M., Ko B.C. Interpretation and simplification of deep forest // arXiv. 2020. arXiv:2001.04721v4. https://doi.org/10.48550/arXiv.2001.04721
36. Nam W.-J., Gur S., Choi J., Wolf L., Lee S.-W. Relative attributing propagation: interpreting the comparative contributions of individual units in deep neural networks // arXiv. 2019. arXiv:1904.00605v4. https://doi.org/10.48550/arXiv.1904.00605
37. Oramas J.M., Wang K., Tuytelaars T. Visual explanation by interpretation: Improving visual feedback capabilities of deep neural networks // arXiv. 2019. arXiv:1712.06302v3. https://doi.org/10.48550/arXiv.1712.06302
38. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead // arXiv. 2019. arXiv:1811.10154v3. https://doi.org/10.48550/arXiv.1811.10154
39. Samek W., Montavon G., Vedaldi A., Hansen L.K., Müller K.-R. Explainable AI: interpreting, explaining and visualizing deep learning // Lecture Notes in Computer Science. 2019. V. 11700. 439 p. https://doi.org/10.1007/978-3-030-28954-6
40. Hansen L.K., Rieger L. Interpretability in intelligent systems – A new concept? // Lecture Notes in Computer Science. 2019. V. 11700. P. 41–49. https://doi.org/10.1007/978-3-030-28954-6_3
41. Liao Q.V., Gruen D., Miller S. Questioning the AI: Informing design practices for explainable AI user experiences // arXiv. 2020. arXiv:2001.02478v2. https://doi.org/10.48550/arXiv.2001.02478
42. Holzinger A., Langs G., Denk H., Zatlouk K., Müller H. Causability and explainability of artificial intelligence in medicine // Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2019. V. 9. N 4. P. e1312. https://doi.org/10.1002/widm.1312
43. Miller T. Explanation in artificial intelligence: insights from the social sciences // Artificial Intelligence. 2019. V. 267. P. 1–38. https://doi.org/10.1016/j.artint.2018.07.007
44. Kulesza T., Burnett M., Wong W., Stumpf S. Principles of explanatory debugging to personalize interactive machine learning // Proc. of the 20th International Conference on Intelligent User Interfaces (IUI '15). 2015. P. 126–137. https://doi.org/10.1145/2678025.2701399
45. Tintarev N. Explaining recommendations // Lecture Notes in Computer Science. 2007. V. 4511. P. 470–474. https://doi.org/10.1007/978-3-540-73078-1_67
46. Chrysostomou G., Alertas N. Improving the faithfulness of attention-based explanations with task-specific information for text classification // arXiv. 2021. arXiv:2105.02657v2. https://doi.org/10.48550/arXiv.2105.02657
47. Vilone G., Longo L. Explainable artificial intelligence: A systematic review // arXiv. 2020. arXiv:2006.00093v3. https://doi.org/10.48550/arXiv.2006.00093
48. Papenmeier A., Englebienne G., Seifert C. How model accuracy and explanation fidelity influence user trust // arXiv. 2019. arXiv:1907.12652v1. https://doi.org/10.48550/arXiv.1907.12652
49. Harutyunyan H. Achille A., Paolini G., Majumder O., Ravichandran A., Bhotika R., Soatto S. Estimating informativeness of samples with smooth unique information // arXiv. 2021. arXiv:2101.06640v1. https://doi.org/10.48550/arXiv.2101.06640
50. Liu S., Wang X., Liu M., Zhu J. Towards better analysis of machine learning models: a visual analytics perspective // Visual Informatics. 2017. V. 1. N 1. P. 48–56. https://doi.org/10.1016/j.visinf.2017.01.006
51. Arrieta A.B., Díaz-Rodríguez N., Del Ser J., Bennetot A., Tabik S., Barbado A., García S., Gil-López S., Molina D., Benjamins R., Chatila R., Herrera F. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI // Information Fusion. 2020. V. 58. P. 82–115. https://doi.org/10.1016/j.inffus.2019.12.012
52. Girshick R., Donahue J., Darrell T., Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation // Proc. of the IEEE Conference on Computer Vision and Pattern Recognition. 2014. P. 580–587. https://doi.org/10.1109/CVPR.2014.81
53. Ancona M., Ceolini E., Özitreli C., Gross M. Towards better understanding of gradient-based attribution methods for deep neural networks // arXiv. 2018. arXiv:1711.06104v4. https://doi.org/10.48550/arXiv.1711.06104
54. Rumelhart D.E., Hinton G.E., Williams R.J. Learning internal representations by error propagation // Readings in Cognitive Science: A Perspective from Psychology and Artificial Intelligence. 2013. P. 399–421.
55. Kindermans P.-J., Hooker S., Adebayo J., Alber M., Schütt K.T., Dähne S., Erhan D., Kim B. The (Un) reliability of saliency

34. Molnar C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Christoph Molnar, 2025, 392 p.
35. Kim S., Jeong M., Ko B.C. Interpretation and simplification of deep forest. *arXiv*, 2020, arXiv:2001.04721v4. https://doi.org/10.48550/arXiv.2001.04721
36. Nam W.-J., Gur S., Choi J., Wolf L., Lee S.-W. Relative attributing propagation: interpreting the comparative contributions of individual units in deep neural networks. *arXiv*, 2019, arXiv:1904.00605v4. https://doi.org/10.48550/arXiv.1904.00605
37. Oramas J.M., Wang K., Tuytelaars T. Visual explanation by interpretation: Improving visual feedback capabilities of deep neural networks. *arXiv*, 2019, arXiv:1712.06302v3. https://doi.org/10.48550/arXiv.1712.06302
38. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *arXiv*, 2019, arXiv:1811.10154v3. https://doi.org/10.48550/arXiv.1811.10154
39. Samek W., Montavon G., Vedaldi A., Hansen L.K., Müller K.-R. Explainable AI: interpreting, explaining and visualizing deep learning. *Lecture Notes in Computer Science*, 2019, vol. 11700. 439 p. https://doi.org/10.1007/978-3-030-28954-6
40. Hansen L.K., Rieger L. Interpretability in intelligent systems – A new concept? *Lecture Notes in Computer Science*, 2019, vol. 11700, pp. 41–49. https://doi.org/10.1007/978-3-030-28954-6_3
41. Liao Q.V., Gruen D., Miller S. Questioning the AI: Informing design practices for explainable AI user experiences. *arXiv*, 2020, arXiv:2001.02478v2. https://doi.org/10.48550/arXiv.2001.02478
42. Holzinger A., Langs G., Denk H., Zatlouk K., Müller H. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2019, vol. 9, no. 4, pp. e1312. https://doi.org/10.1002/widm.1312
43. Miller T. Explanation in artificial intelligence: insights from the social sciences. *Artificial Intelligence*, 2019, vol. 267, pp. 1–38. https://doi.org/10.1016/j.artint.2018.07.007
44. Kulesza T., Burnett M., Wong W., Stumpf S. Principles of explanatory debugging to personalize interactive machine learning. *Proc. of the 20th International Conference on Intelligent User Interfaces (IUI '15)*, 2015, pp. 126–137. https://doi.org/10.1145/2678025.2701399
45. Tintarev N. Explaining recommendations. *Lecture Notes in Computer Science*, 2007, vol. 4511, pp. 470–474. https://doi.org/10.1007/978-3-540-73078-1_67
46. Chrysostomou G., Alertas N. Improving the faithfulness of attention-based explanations with task-specific information for text classification. *arXiv*, 2021, arXiv:2105.02657v2. https://doi.org/10.48550/arXiv.2105.02657
47. Vilone G., Longo L. Explainable artificial intelligence: A systematic review. *arXiv*, 2020, arXiv:2006.00093v3. https://doi.org/10.48550/arXiv.2006.00093
48. Papenmeier A., Englebienne G., Seifert C. How model accuracy and explanation fidelity influence user trust. *arXiv*, 2019, arXiv:1907.12652v1. https://doi.org/10.48550/arXiv.1907.12652
49. Harutyunyan H. Achille A., Paolini G., Majumder O., Ravichandran A., Bhotika R., Soatto S. Estimating informativeness of samples with smooth unique information. *arXiv*, 2021, arXiv:2101.06640v1. https://doi.org/10.48550/arXiv.2101.06640
50. Liu S., Wang X., Liu M., Zhu J. Towards better analysis of machine learning models: a visual analytics perspective. *Visual Informatics*, 2017, vol. 1, no. 1, pp. 48–56. https://doi.org/10.1016/j.visinf.2017.01.006
51. Arrieta A.B., Díaz-Rodríguez N., Del Ser J., Bennetot A., Tabik S., Barbado A., García S., Gil-López S., Molina D., Benjamins R., Chatila R., Herrera F. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 2020, vol. 58, pp. 82–115. https://doi.org/10.1016/j.inffus.2019.12.012
52. Girshick R., Donahue J., Darrell T., Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587. https://doi.org/10.1109/CVPR.2014.81
53. Ancona M., Ceolini E., Özitreli C., Gross M. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv*, 2018, arXiv:1711.06104v4. https://doi.org/10.48550/arXiv.1711.06104
54. Rumelhart D.E., Hinton G.E., Williams R.J. Learning internal representations by error propagation. *Readings in Cognitive Science: A Perspective from Psychology and Artificial Intelligence*, 2013, pp. 399–421.

methods // Lecture Notes in Computer Science. 2019. V. 11700. P. 267–280. https://doi.org/10.1007/978-3-030-28954-6_14
56. Roscher R., Bohn B., Duarte M.F., Garcke J. Explainable machine learning for scientific insights and discoveries // arXiv. 2020. arXiv:1905.08883v3. https://doi.org/10.48550/arXiv.1905.08883

55. Kindermans P.-J., Hooker S., Adebayo J., Alber M., Schütt K.T., Dähne S., Erhan D., Kim B. The (Un) reliability of saliency methods. *Lecture Notes in Computer Science*, 2019, vol. 11700, pp. 267–280. https://doi.org/10.1007/978-3-030-28954-6_14
56. Roscher R., Bohn B., Duarte M.F., Garcke J. Explainable machine learning for scientific insights and discoveries. *arXiv*, 2020, arXiv:1905.08883v3. https://doi.org/10.48550/arXiv.1905.08883

**Авторы**

**Бирюков Денис Николаевич** — доктор технических наук, профессор, начальник кафедры, Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург, 197198, Российская Федерация, SC 57188163400, https://orcid.org/0000-0003-1300-2470, Biryukov.D.N@yandex.ru

**Дудкин Андрей Сергеевич** — кандидат технических наук, доцент, заместитель начальника кафедры, Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург, 197198, Российская Федерация, SC 57211979130, https://orcid.org/0000-0003-0283-9048, andry-ll@mail.ru

**Denis N. Biryukov** — D.Sc., Professor, Head of Department, Mozhaisky Military Aerospace Academy, Saint Petersburg, 197198, Russian Federation, SC 57188163400, https://orcid.org/0000-0003-1300-2470, Biryukov.D.N@yandex.ru

**Andrey S. Dudkin** — PhD, Associate Professor, Deputy Head of Department, Mozhaisky Military Aerospace Academy, Saint Petersburg, 197198, Russian Federation, SC 57211979130, https://orcid.org/0000-0003-0283-9048, andry-ll@mail.ru

*Статья поступила в редакцию 03.02.2025*
*Одобрена после рецензирования 11.04.2025*
*Принята к печати 26.05.2025*

*Received 03.02.2025*
*Approved after reviewing 11.04.2025*
*Accepted 26.05.2025*