Министерство науки и высшего образования Российской Федерации

УРАЛЬСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ
ИМЕНИ ПЕРВОГО ПРЕЗИДЕНТА РОССИИ Б. Н. ЕЛЬЦИНА

НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

ПРАВОВЫЕ АСПЕКТЫ ПРИМЕНЕНИЯ ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМ И МЕТОДЫ ИНТЕРПРЕТАЦИИ ИХ РАБОТЫ

Монография

НОВОСИБИРСК
2024

ББК 67.401.114
П685

Коллектив авторов:
К. И. Хальясмаа, А. И. Степанова, Е. Л. Зиновьева,
П. В. Матренин, А. И. Хальясмаа, С. А. Ерошенко

Рецензенты:
д-р техн. наук, профессор МИРЭА–РТУ *А. М. Романов*
д-р техн. наук, профессор ЛЭТИ *И. И. Холод*

П685 **Правовые аспекты применения интеллектуальных систем и методы интерпретации их работы**: монография / К. И. Хальясмаа, А. И. Степанова, Е. Л. Зиновьева и др.; под общ. ред. А. И. Хальясмаа. – Новосибирск: Изд-во НГТУ, 2024. – 191 с.

ISBN 978-5-7782-5328-5

В монографии рассматриваются правовые аспекты регулирования технологий искусственного интеллекта в России и за рубежом, включая вопросы ответственности и этики, а также способы повышения прозрачности работы интеллектуальных систем на основе методов объяснимого искусственного интеллекта. Описаны принципы наиболее распространенных методов. Приведен отраслевой пример использования системы поддержки принятия решений и объяснимого искусственного интеллекта в задаче краткосрочного прогнозирования генерации солнечной электрической станции с учетом метеорологических факторов.

Работа выполнена в рамках государственного задания при финансовой поддержке Министерства науки и высшего образования Российской Федерации (тема № FEUZ-2022-0030 Разработка интеллектуальной мультиагентной системы для моделирования глубоко интегрированных технологических систем в электроэнергетике).

ББК 67.401.114

ISBN 978-5-7782-5328-5

ОГЛАВЛЕНИЕ

СПИСОК СОКРАЩЕНИЙ

API – Application Programming Interface.
CAM – Class Activation Mapping.
CLEAR – CLass-Enhanced Attentive Response.
ESSR – Electricity Self-Sufficiency Rate.
FIM – Feature Importance Map.
Grad-CAM – Gradient-weighted Class Activation Mapping.
ICE – Individual Conditional Expectation Curves.
LIME – Local Interpretable Model-Agnostic Explanation.
LRP – Layer-wise Relevance Propagation.
LSTM – Long Short-Term Memory.
ROI – Region of Interest.
PCA – Principal Component Analysis.
PRM – Peak Response Maps.
SAR – Synthetic Aperture Radar.
SHAP – SHapley Additive exPlanations.
ST-LRP – Spatial–Temporal Layer-wise Relevance Propagation.
tSNE – T-distributed stochastic neighbor embedding.
VQA – Visual Question Answering.
WSLSTM – Weight-Sharing Long Short-Term Memory networks.
XAI – eXplainable Artificial Intelligence.
DSL – digital subscriber line.
ИИ – искусственный интеллект.
ЛПР – лицо, принимающее решение.

ТЕРМИНЫ И ОПРЕДЕЛЕНИЯ

Объяснимый искусственный интеллект – набор процессов и методов, которые позволяют пользователям понимать результаты, созданные алгоритмами машинного обучения, и доверять им.

Стохастическое вложение соседей с *t*-распределением – метод визуализации, который проецирует высокоразмерные данные в двух- или трехмерные пространства с использованием условных вероятностей для представления расстояний между точками данных и поиска сходств между ними.

Феноменология – исследование структур сознания того, как они переживаются с точки зрения первого лица.

Черный ящик – система, которая описывается только через входы и выходы без представления внутренней структуры модели и логики ее работы.

Ядерное сглаживание – локальный непараметрический способ оценки плотности случайной величины.

ВВЕДЕНИЕ

Согласно Национальной стратегии развития искусственного интеллекта до 2030 года, «искусственный интеллект – комплекс технологических решений, позволяющий имитировать когнитивные функции человека (включая поиск решений без заранее заданного алгоритма) и получать при выполнении конкретных задач результаты, сопоставимые с результатами интеллектуальной деятельности человека или превосходящие их». В этой стратегии подчеркивается важность технологий искусственного интеллекта для развития Российской Федерации и их значимость для достижения технологической независимости и конкурентоспособности страны.

Способность технологий искусственного интеллекта повышать производительность труда и находить решения, которые не способен найти человек, делает их актуальными во всех сферах экономики. В то же время до сих пор отсутствует ясное понимания того, как именно происходит процесс решения поставленной задачи и формируется конечный результат методами, наиболее эффективными с точки зрения точности решения задач. Это является одной из причин низкого уровня доверия к технологиям искусственного интеллекта и создает барьер на пути их внедрения.

Отдельный вопрос – это отставание нормативно-правовой базы от появления новых технологий искусственного интеллекта. Поэтому необходимо обобщать существующий отечественный и зарубежный опыт правового регулирования исследований в области искусственного интеллекта и применения интеллектуальных информационных систем. Важен вопрос разделения ответственности между всеми участниками процесса, от постановки задачи до конечного применения. Эти вопросы рассматриваются в первом разделе монографии.

Концепция объяснимого искусственного интеллекта является одним из перспективных направлений, позволяющим, с одной стороны,

повысить интерпретируемость результатов, полученных с помощью технологий искусственного интеллекта, с другой – определить причину ошибки при анализе нарушений, связанных с применением искусственного интеллекта. Методы объяснимого искусственного интеллекта описаны во втором разделе работы.

Третий раздел содержит практические разработки интеллектуальной информационной системы с интерпретаций ее результатов в задаче прогнозирования выработки фотоэлектрической станции и учетом метеорологических факторов.

1. АСПЕКТЫ ПРАВОВОГО РЕГУЛИРОВАНИЯ ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

1.1. ИСТОРИЯ РАЗВИТИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

На сегодняшний день единая трактовка понимания искусственного интеллекта отсутствует, поэтому существуют разнообразные подходы к пониманию данного термина. Рассмотрим два основных источника [1, 2]: в первом из них искусственный интеллект упоминается как некая система алгоритмов, которая использует ряд шагов для осуществления процесса преобразования входных данных в выходные данные. Примером такого процесса может быть получение изображений с помощью внешних камер наблюдения, которые затем будут обрабатываться для идентификации конкретных лиц.

Другим подходом является трактовка И.А. Филиповой искусственного интеллекта как способности выполнять некие творческие функции, обычно присущие лишь человеку [2].

Таким образом, в общих чертах искусственным интеллектом можно назвать научный подход к формированию различных технических систем, которые способны обрабатывать информацию и совершать те или иные действия, при обычных условиях характерные лишь для человека.

Несмотря на отсутствие единой трактовки термина, научное упоминание искусственного интеллекта было впервые представлено научному миру американцем Джоном Маккарти на Дармутском семинаре в 1956 году.

Это знаменательное событие – двухмесячная конференция, которая была проведена летом 1956 года в Дартмутском колледже и посвящена ИИ, сыграло ключевую роль для науки. Площадка дала возможность научным деятелям, интересующимся концепцией моделирования человеческого разума, познакомиться друг с другом и закрепила появление прорывной области науки, дав ей название – «Artificial Intelligence», или

«Искусственный интеллект». Главной целью конференции была дискуссионная тема: представляется ли возможным моделирование рассуждений, интеллекта и творческих процессов с помощью вычислительных машин?

По определению Джона Маккарти, искусственный интеллект – это наука и техника создания интеллектуальных машин, особенно интеллектуальных компьютерных программ, что связано с аналогичной задачей использования компьютеров для понимания человеческого интеллекта. Однако важно понимать, что ИИ не должен ограничиваться биологически наблюдаемыми методами [3].

Понятие «автоматон» происходит от греческого слова «самодвижущийся» и означает роботоподобную конструкцию, которая посредством механического привода способна выполнять разноплановые действия, такие как движение конечностями, повороты головы, открытие и закрытие век. История таких устройств уходит своими корнями в весьма отдаленное прошлое. Различные механизмы, функционал которых можно было бы отнести к сегодняшним роботам, упоминались еще до нашей эры. Одно из первых упоминаний о роботе, похожем на человека, относится к III веку до н. э.: инженер, механик и математик Филон Византийский (280 – 220 до н. э.) создал статую-служанку: механизм позволял ей наливать в чашу вино и разбавлять его водой.

Леонардо да Винчи (1452 – 1519 гг.) изобрел ряд подобных механизмов. Среди его изобретений – механический рыцарь. Механизм, расположенный внутри рыцарской брони, позволял системе повторять различные движения за человеком: автоматон мог присесть, встать, двигать руками, шеей. Точных данных о том, был ли этот механизм действительно собран, нет, потому как найдены были лишь чертежи устройства, однако его смогли реконструировать наши современники по чертежам изобретателя.

Постепенно происходило последовательное внедрение автоматонов в различные сферы жизни: имитация слуг для подачи чашек с чаем, воины-автоматы, стреляющие из лука, куклы для театра. Позднее – более сложные механизмы, которые могли играть на музыкальных инструментах и, например, выводить пером текст.

При этом инновационные разработки часто сопровождались обманом. Один из наиболее ярких примеров – «Шахматный турок». С виду действительно создавалось впечатление, что устройство обладает искусственным интеллектом и может, применяя свои аналитические

способности, с легкостью обыгрывать своих соперников. Одним из его известных соперников был даже Наполеон, однако на деле внутри такой машины находился человек, и чаще всего это были гроссмейстеры, такие, как, например, Иоганн Альгайер (сильнейший шахматист XIX века).

Говоря уже о более приближенном к нам времени, стоит упомянуть произведение Карела Чапека – пьесу под названием «R.U.R. – Россумские Универсальные Роботы». Именно благодаря этому известному чешскому писателю-фантасту XX века в лексиконе появилось слово «робот», впервые упомянутое на страницах его произведения в 1920 году. История повествует о фабрике по производству искусственных людей из плоти и крови, которые были названы чешским словом, обозначающим каторгу или тяжелую работу – «robota». Его искусственные люди могли мыслить, а в центре сюжета – уничтожение роботами всего человечества. Пьеса была несколько раз экранизирована и на самом деле стала поводом для различных фантазий на тему роботов.

Еще одна важная дата: 1946 год – рассказ «Хоровод» Айзека Азимова. Американский писатель, родом из Смоленской области, представил в нем три закона робототехники [4].

1. Робот не может причинить вред человеку или своим бездействием допустить, чтобы человеку был причинен вред.
2. Робот должен повиноваться всем приказам, которые дает человек, кроме тех случаев, когда эти приказы противоречат первому закону.
3. Робот должен заботиться о своей безопасности в той мере, в которой это не противоречит первому или второму закону.

Именно они стали основой для дальнейшей проработки взаимодействия с ИИ, ведь как раз спустя 10 лет после этого события было дано первое конкретное определение искусственного интеллекта, о котором упоминалось выше [5].

Что касается современных концепций понимания ИИ, они также частично опираются на три фундаментальных закона робототехники. Более глубокий и фундаментальный анализ работы искусственного интеллекта был зафиксирован на Асиломарской конференции, прошедшей в 2023 году и посвященной растущему интересу к искусственному интеллекту. На конференции были сформулированы 23 принципа, которые призваны оказать содействие в понимании подводных камней и дискуссионных моментов раскрытия потенциала использования ИИ.

По данным портала C-News, в 2023 году на примере США можно увидеть снижение показателей по созданию рабочих мест в сфере ИТ – лишь 700 новых рабочих мест на фоне 267 тысяч в предыдущем году, при фактически таком же количестве увольнений. Несложно догадаться, что происходит это на фоне повсеместной замены человеческого труда искусственным интеллектом. Такая трансформация требует в первую очередь активной проработки вопросов законодательства [6].

Однако новый виток развития искусственный интеллект получил чуть ранее – еще в начале 2000-х годов, когда роботы стали всё более привычным явлением на предприятиях, улицах города и даже в домах. Поскольку искусственный интеллект стал планомерно проникать в различные сферы нашей жизни, в обществе вновь остро встал вопрос об этической стороне искусственного интеллекта.

На государственном уровне стали подниматься вопросы о постепенном внедрении ИИ в важнейшие области социальной и экономической сферы, о чем свидетельствует Указ Президента Российской Федерации от 9 мая 2017 г. № 203 «О Стратегии развития информационного общества в Российской Федерации на 2017–2030 гг.», а также распоряжение Правительства Российской Федерации от 28 июля 2017 г. № 1632-р «Об утверждении программы “Цифровая экономика Российской Федерации”».

Что касается международного уровня обсуждения вопросов, касающихся искусственного интеллекта, можно отметить:

- первый симпозиум по робоэтике, состоявшийся в 2004 году в итальянском городе Сан-Ремо (Италия);
- принятие в Японии Всемирной декларации о роботах (2004).

Оба документа определили на начальном этапе основные вехи развития искусственного интеллекта, а более основательным документом была принятая в 2017 году Европейская хартия робототехники с рекомендациями для Европейской комиссии по нормам применения гражданского права в области робототехники [7–9].

Дискуссионные вопросы развития ИИ обусловлены повышенным вниманием к проблемам этики машинного обучения и робототехники.

Согласно концепции Асаро можно определить несколько направлений понимания робоэтики:

1) этические системы, которые являются составной частью конструкционных элементов робота;

2) этические нормы разработчиков роботов;
3) этика непосредственного использования роботов [10].

Что касается первой составляющей этических систем, предполагается, что заложенные в конструкцию робота алгоритмы при принятии важных с точки зрения этики решений должны делать это грамотно и последовательно. Этот этап подразумевает, что ключевые установки, включая решение задач, сопряженных с вопросами этических норм, должны закладываться на этапе создания. При этом допускается внесение точечных корректировок, не противоречащих нормальной работе ИИ, в том числе в рамках законодательства.

Вторая составляющая, этические нормы, важна с точки зрения самостоятельного определения разработчиками нравственных норм, поскольку именно она будет влиять на регуляцию деятельности ИИ. Необходимо, чтобы данные установки соответствовали общепринятым нормам морали. Важно разграничение ответственности между разработчиком, собственником и пользователем ИИ. Немаловажным аспектом является также проработка экспертами, в том числе на межгосударственном уровне, правовых документов, закладывающих базу для использования ИИ.

Третья составляющая, этика непосредственного использования роботов, должна не противоречить существующим законам и не допускать причинения вреда людям, с одной стороны, и имуществу, включая самих роботов, – с другой. Необходима прозрачность технологий искусственного интеллекта, вред от применения которых может быть нанесен не только их владельцу, но и другим людям.

Помимо глобального развития искусственного интеллекта, законодательство развивается также на локальном уровне. Развитие нейросетей, а также моделей глубокого обучения привело к принятию собственных этических норм десятками организаций, связанных с ИТ, в их числе, к примеру, IBM и компания Microsoft, опубликовавшая в 2016 году «10 законов для искусственного интеллекта» [11].

Что касается России, то в 2021 году Москва стала организатором первого международного форума «Этика искусственного интеллекта: начало доверия», по итогам которого был принят российский Кодекс этики в сфере искусственного интеллекта. В настоящий момент он носит лишь рекомендательный характер, но несколько крупных российских компаний присоединилось к кодексу, среди которых Сбербанк, Яндекс, МТС, Ростелеком и Газпромнефть [12, 13].

Развитие законодательства на локальном уровне породило концепцию AI Localism (local от латинского localis – местный), которая описывает структуру управления снизу вверх, потому как национальная политика не всегда согласуется с особенностями местного управления. Связано это прежде всего с тем, что стратегии развития искусственного интеллекта, как правило, очень глобальны, поэтому в конечном итоге всё сводится к конкретным действиям лиц, которые решают проблемы в пределах какой-то конкретной территории (города или поселения) в целях удовлетворения местных потребностей.

Основная идея концепции заключается в необходимости формирования правовой и регулирующей структуры так, чтобы из ИИ можно было извлечь максимальное количество пользы и минимизировать потенциальный вред [14].

Ниже приведены примеры применения AI Localism.

США, штат Орегон, Сэнди: местное самоуправление создало систему SandyNet для обеспечения высокоскоростного DSL и беспроводного подключения к Интернету по низкой цене. Подобные примеры есть в Испания (Барселона) и Южной Корее (Сеул).

Что касается правового регулирования с точки зрения защиты данных, можно рассмотреть следующие примеры.

- Постановление администрации города Нью-Йорк, которое регулирует сбор и использование данных граждан правительством и правоохранительными органами.
- Введенный в Сан-Франциско запрет для полиции и местных органов власти использовать технологию распознавания лиц для идентификации жителей.

Подобные меры направлены на решение проблем на местном уровне, связанных с беспокойством по поводу нарушения конфиденциальности личной жизни.

Однако у любого управления даже на локальном уровне есть свои недостатки, потому и концепция AI Localism не всегда приводит к правильному решению. Существуют примеры, когда местные усилия по регулированию и использованию ИИ нарушают общественные свободы и наносят вред общественным благам.

К примеру, широкую огласку получила общественная критика администрации района Харборфронт-Сентр в Торонто (Канада) за решение поручить дочерней компании Google, занимающейся городским планированием и инфраструктурой Sidewalk Labs, собирать инфор-

мацию о местных жителях с помощью датчиков и камер для «оптимизации» городской среды. В итоге в 2020 году проект всё же был свернут из-за отсутствия общественного доверия к его прозрачности, а также начавшейся пандемии COVID-19.

В настоящее время в политике управления всё чаще упоминается такое понятие, как концепция умного города, направленная на получение функциональных и экономических выгод. При этом в стремлении овладеть инновационными технологиями из виду часто упускают разные темпы развития самого ИИ и инструментов, обеспечивающих его работу.

Как следует подходить к пониманию ИИ: с технической или практико-этической точки зрения? Для того чтобы разобраться, как правильнее осуществлять правотворческий процесс, необходимо прежде всего определить, с какой стороны подходить к пониманию ИИ.

Принципы и права с точки зрения искусственного интеллекта можно разделить на две группы: нормативные и догматические.

Нормативные принципы подразумевают закрепление прав человека и целей устойчивого развития, отсутствие дискриминации и конфиденциальность личных данных, поскольку необходимо стремиться достичь управления ИИ на любом уровне.

Догматические принципы подразумевают неоспоримые особенности применения ИИ, закрепленные различными контролирующими органами с помощью систематических проверок и прозрачности использования ИИ.

Надежность соблюдаемых правовых принципов позволяет разработку и применение технологии искусственного интеллекта. Поскольку это всё еще развивающаяся и относительно молодая область, которая находится на стыке множества других областей, в том числе более зрелых, она впитывает уже устоявшиеся ценностные ориентиры. Хотя многие эти принципы формируют фундаментальную основу благодаря своей многогранности, они часто носят слишком общий характер и в силу этого непригодны к использованию в более узконаправленной тематике.

Возникает вопрос, как подойти к решению этой проблемы. Улучшить ситуацию можно с помощью индуктивного подхода, т. е. получения обобщающего знания на основе отдельных данных. Его принцип включает в себя анализ предписывающей документации по разработке технологий ИИ, в первую очередь того, что касается крупных компаний

международного уровня. Именно такой анализ привел к появлению 23 основных категорий, зафиксированных на Асиломарской конференции.

Такой подход позволяет более конструктивно подойти к проектированию каждой отдельной категории ИИ с учетом различных этических ценностей. Конкретные формулировки позволяют более четко применять принципы проектирования для улучшения ИИ.

На фоне всеобщего осознания необходимости правового регулирования искусственного интеллекта в общественных целях пробел в определении ответственности остается довольно весомым, т. е. правовая база не успевает за развитием технологий. Вопрос упирается прежде всего в человекоориентированный подход к пониманию ИИ, поскольку техническая система ответственности нести не может и она всегда ложится на конкретного человека, что создает неопределенность в случае конфликтных ситуаций. Это, в свою очередь, представляет собой дополнительные трудности в разработке объяснимого искусственного интеллекта.

Например, в США правительственные органы обращаются к местным законодательным органам для определения пределов использования ИИ для решения жалоб, связанных с возможностью нанесения вреда автоматизированными системами, что будет нарушением прав граждан в случае их нерегулирования.

Законы служат тем инструментом, который в государственном аппарате используется для осведомления граждан о возможных последствиях применения ИИ, в том числе для контроля посягательства на частную жизнь.

Однако существует проблема несогласованности законодательства на местном и национальном уровне об использовании искусственного интеллекта. Всё это напоминает гипотезу из широкоизвестного произведения «Алиса в стране чудес» писателя Льюиса Кэрролла. Алиса, участвуя в забеге, понимает, что ей нужно бежать вдвое быстрее, чем обычно, чтобы двигаться вперед. Эта метафора хорошо описывает и то, насколько законодательство в целом отстает от темпов развития ИИ.

Помимо того, что ИИ является технической системой, он также является частью социальной системы и культуры в целом. По сути, ИИ – это социотехническая система. Как следствие, прозрачность ИИ должна быть принята в качестве технико-социального вопроса. Прежде всего прозрачность позволяет определить цель использования системы ИИ,

т. е. объяснить, для каких целей разработана система, почему выбрана именно она, какие данные нужны для ее работы, кто применяет систему, как она создается, каковы условия использования этой системы.

Подотчетность систем и надзор за ними связаны в первую очередь с растущей обеспокоенностью о возможном негативном воздействии ИИ, т. е. о прозрачности используемых технологий. Так как в конечном итоге ответственным лицом будет в любом случае тот или иной человек, то необходимо четко понимать, каким конкретно образом работает искусственный интеллект. Ряд стран, например США, стали принимать конкретные шаги в отношении технологий, применяющихся в общественных местах, что послужило поводом для всеобщей кооперации усилий для развития стратегий алгоритмической прозрачности [15].

В 2020 году администрация г. Портленд, штат Орегон, ввела строгие правила в области распознавания лиц, приняв два постановления, запрещающих частным и государственным учреждениям использовать программное обеспечение наблюдения за людьми в общественных местах. Этот закон запрещает использовать средства распознавания лиц на видеозаписях, включая записи, собранные с камер полиции, записи камер отелей и аптек для выявления и преследования людей.

С другой стороны, существует ряд примеров из других стран, где ситуация совершенно обратная. К примеру, в Англии активно используются записи с нательных камер полиции для предотвращения и быстрого выявления должностных преступлений в форме коррупции. Что касается практики Российской Федерации, здесь использование должностными лицами материалов с камер в общественных местах допускается (метро, стадион).

Таким образом, можно сделать вывод о том, что до сих пор отсутствует единое понимание трактовки ИИ. В общих чертах искусственным интеллектом можно назвать научный подход к формированию различных машинных систем, которые способны выполнять те или иные действия, при обычных условиях характерные лишь для человека.

Ниже приведены основные вехи развития роботизированных и интеллектуальных систем.

- Автоматон – первая конструкция, откуда берет начало развитие робототехники.
- 1920 год – появление непосредственно слова «робот» в произведении чешского писателя.

- 1946 год – формулирование трех законов робототехники в литературе, которые легли в основу дальнейшей проработки правовой базы.
- 1956 год – термин «искусственный интеллект» был впервые упомянут в научном ключе.
- Активное внедрение технологий ИИ в повседневную жизнь с начала 2000-х годов породило понимание необходимости проработки правовой базы его использования. На межгосударственном и внутригосударственном уровне стали приниматься различные правовые документы касательно ИИ (Асиломарская конференция, локальные законодательства внутри стран).
- Общество пришло к осознанию большого разрыва между проработкой правовой базы и темпами развития технологий искусственного интеллекта.

1.2. ПРИОРИТЕТ ЗАЩИТЫ ПРАВ И ИНТЕРЕСОВ ЧЕЛОВЕКА В УСЛОВИЯХ РАЗВИТИЯ ТЕХНОЛОГИЙ ИИ

Английский математик Алан Тьюринг сформулировал известный подход к определению того, обладает ли устройство искусственным интеллектом или нет. Устройство следует считать имеющим интеллект, если при общении с ним посредством анонимного канала связи нельзя понять, с кем идет беседа – с человеком или машиной [16].

В связи с этим первоначально правовое регулирование искусственного интеллекта обеспечивалось совокупностью технологических норм, обеспечивающих рациональное использование новых технологий.

Прогрессивное развитие технологии искусственного разума обусловило появление другого подхода к определению понятия «искусственный интеллект», в соответствии с которым отличительным критерием искусственного интеллекта служит его способность к автономным действиям и самосовершенствованию.

Таким образом, к признакам ИИ относятся:

- способность к принятию, обработке и передаче информации;
- способность к автономной работе;
- самообучение на основе анализа информации и приобретенного опыта;
- способность к принятию самостоятельных решений.

Рассмотрим тенденции развития инновационных технологий. В 2023 году компанией «Делойт», которая входит в «большую четверку» аудиторских компаний и является самой крупной профессиональной сетью по количеству сотрудников, был опубликован документ [17], в котором описываются важнейшие тенденции развития технологий, поскольку все они оказывают большое влияние на различные стороны жизни. Помимо общих направлений развития ИТ, таких как интеллектуальность систем, простота, в ежегодном отчете были сформулированы три дополнительные области развития ИТ, среди которых указаны бизнес-технологии, кибербезопасность и базовая модернизация. В общих чертах ИТ можно назвать инновациями в сфере взаимодействия, информации и вычислений.

Иммерсивность (от англ. immerse – погружать) – это свойство контента за счет применения различных, прежде всего аудиовизуальных, технологий погружать пользователя в содержание. В идеале погружение достигается за счет воздействия на все органы чувств человека, однако современные технологии еще не достигли такого охвата, поэтому в условиях сегодняшних реалий это визуальное и слуховое восприятие пользователем, в более редких случаях тактильное – вкус и запах в настоящий момент сильно ограничены. Интернет для предприятий – тенденция к тому, что взаимодействие с цифровым миром будет происходить с помощью технологий смешанной реальности [18]. Поскольку подход у компаний разный, часть организаций поддерживает бизнес-модели с «неограниченной реальностью», другие осуществляют это при помощи интересной среды для оптимизации процессов и совместной работы, а также обучения.

Открытость систем ИИ обеспечивает доверие к использованию таких систем, поскольку их возможности с каждым днем всё больше выходят за рамки примитивных вычислений. Кроме того, «приручение облачного хаоса» (обращение к метаоблакам или супероблакам для сокращения количества мультиоблачных сред) позволяет иметь доступ к таким инструментам, как хранилища данных.

Необходимость гибкости процессов вызвана устареванием технических навыков сотрудников и целесообразностью их замещения технологиями ИИ в долгосрочной перспективе для более эффективного использования ресурсов компаний.

Технологии блокчейна позволяют сделать системы более прозрачными с помощью разработки децентрализованных архитектур и экосистем.

Модернизация больших универсальных высокопроизводительных и отказоустойчивых серверов, использующихся в критически важных системах, необходима для замены устаревших базовых систем с целью обеспечения цифровой трансформации.

Пандемия COVID-19 изменила многие сферы жизни и перестроила работу различных отраслей [19].

1. *Генеративный искусственный интеллект* (*GenAI*) – алгоритмы ИИ, которые могут осуществлять поиск и обобщение, а также способны создавать принципиально новые результаты, основываясь на обучающей выборке данных. Он может использоваться в качестве инструмента для автоматизации некоторых процессов, например, создания однотипной документации, формулирования выводов по результатам исследования.

2. *Удаленная работа и совместная работа* часто основана на технологиях с поддержкой искусственного интеллекта. Например, чат-боты, которые работают на основе искусственного интеллекта и помогают обрабатывать большой поток вопросов от клиентов. Существуют также инструменты, используемые во время видеоконференции с поддержкой искусственного интеллекта – нужны они для оптимизации встреч и облегчения совместной работы в целом.

3. *Бесконтактные технологии*, такие как киоски самообслуживания с поддержкой искусственного интеллекта и чат-боты (стали очень востребованы для снижения учащенных случаев заражения).

4. *Принятие решений на основе данных* – ИИ применяется для сбора и анализа больших объемов данных, с которыми не способен работать человек, поэтому выводы и рекомендации ИИ могут в дальнейшем быть использованы для принятия более эффективных решений в различных сферах, как, например, разработка, обслуживание или продвижение продукта и т. д.

5. *Увеличение доступности инструментов и сервисов, поддерживаемых технологиями ИИ*:

- стоимость таких технологий снижается, что упрощает возможность их применения на предприятиях как малого, так и большого размера;
- увеличивается количество стартапов с использованием ИИ, которые внедряют инновационные технологии.

К тенденциям относится также развитие исследований в области Deep Technology («глубокие технологии»). Глубокие технологии –

применение инновационных разработок в области науки с целью создания коммерческой продукции. Рассмотрим основные направления.

- Генеративный ИИ, автономное вождение, защита данных, прикладной искусственный интеллект: обработка естественного языка, компьютерное зрение.
- EnergyTech: проекты альтернативной энергетики («зеленая» энергетика), ядерная энергетика, водородная энергетика, технологии хранения энергии, биоэнергетика, инновационное топливо.
- Распределенный реестр и распределенные вычисления: Кибербезопасность, блокчейн, Web 3.0.
- SpaceTech: спутниковые платформы, ракетоносители, орбитальные транспортные средства, телескопы, антенны, возобновляемые ракеты нового поколения.
- Новая мобильность: электромобили, автономный транспорт, БПЛА, новые типы двигателей и системы хранения энергии.
- Квантовые технологии: квантовые вычисления – квантовые коммуникации – квантовые сенсоры.
- AgroTech: ускоренная селекция и генотипирование, точное земледелие, специализированная с/х техника, автоматизированные фермы, биотехнологии в сельском хозяйстве.
- Новые материалы: наноматериалы, сверхпроводники, материалы с заданными свойствами, композитные материалы, графен, возобновляемые полимеры, антимикробный пластик, самовосстанавливающийся бетон, синтетические алмазы, 3D-печать, биоматериалы и др.
- Life Sciences: медицинские устройства, персонализированная медицина на основе ИИ, генетика, синтетическая биология, нейроинтерфейсы, кардиогеномика, ИИ для разработок в области химии и биотехнологий.
- Робототехника и сенсорика: промышленная и сервисная робототехника, коллаборативные роботы, Интернет вещей (IoT), экзоскелеты, электронная компонентная база.

Deep Tech представляет собой одну из наиболее привлекательных областей для инвестиций крупных компанийв. Это связано в первую очередь с тем, что такие технологии имеют долгосрочные конкурентные преимущества. Среди перечисленных направлений искусственный интеллект является той областью, которая привлекла наибольший объем инвестиций по состоянию на 2022 год [20] (рис. 1.1).

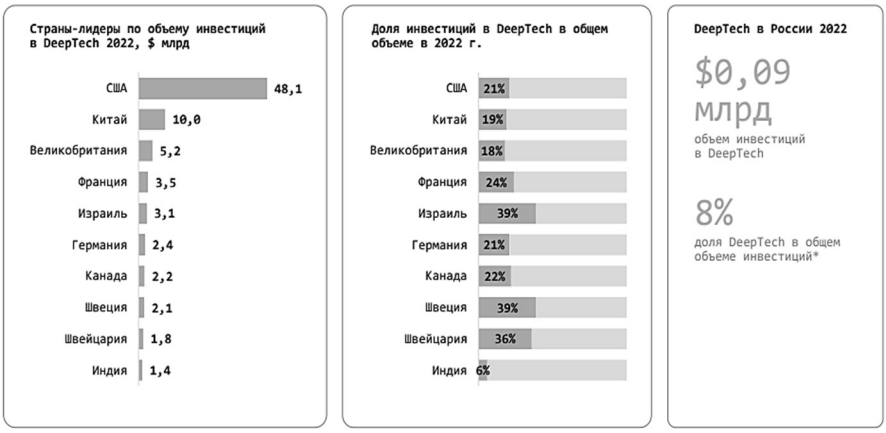


Рис. 1.1. Инвестиции в DeepTech

Однако несмотря на положительные стороны развития технологий, у них есть и минусы. Помимо обеспечения доступности технологий ИИ не стоит забывать и о вреде, который может наноситься этими технологиями и быть как интуитивно понятным, так и совершенно неизвестным. На сегодняшний день список этих риск-факторов может расширяться с приростом более мощных систем и включать в себя, например, незаконные источники заработка, дезинформацию и манипуляции с целью нарушения политических процессов, обход систем кибербезопасности и т. д. [21].

Некоторые риски возникают в силу технических ограничений и могут быть как предумышленными действиями, так и ошибками, именуемыми «галлюцинациями» или «конфабуляциями», что является частой проблемой работы генеративного ИИ.

Другой тип рисков связан больше с конкретной деятельностью людей, чем с работой самого ИИ. Фейковые материалы и намеренное распространение враждебных данных – лишь часть примеров с четким негативным посылом.

Следующая группа включает в себя взаимодействие человека и машины; простыми словами это можно описать как излишнее доверие к интеллектуальным системам, что влечет риски неточности принятых решений и падения квалификационных качеств персонала с течением времени.

Отдельным вопросом является использование ИИ в вооружении, так как это противоречит одному из основополагающих принципов о непричинении вреда человеку с помощью роботов.

Беспокойство общества вызывает и неконтролируемое развитие ИИ, которое может представлять угрозу существованию человечества.

Целесообразность составления единого большого списка рисков отсутствует, так как всеобъемлющим его все равно не сделать в силу слишком быстрых темпов развития ИИ, поэтому целесообразнее разделить их на общие категории. Это позволит оценить риски, так как они будут изменяться с течением времени в силу изменения технологий и внедрения в различные сферы. Такой подход будет способствовать формированию единой системы контроля рисков с учетом практического опыта тех или иных регионов с течением времени. Основной контролирующей площадкой указанных рисков может стать ООН.

Ниже приведена общая классификация рисков с точки зрения влияния ИИ.

1. *Физические лица*. Честь и достоинство человека; человеческий фактор (намеренные манипуляции, обман, подкуп, подстрекательство, вынесение судебных решений):

- жизнь человека, его безопасность (автономное оружие, беспилотные транспортные средства, обеспечение химической, биологической, радиологической и ядерной защиты);
- физическая и психическая неприкосновенность, здоровье и безопасность (диагностика, подстрекательства, нейротехнологии);
- гражданские права и свободы человека, например, справедливое судебное разбирательство, презумпция невиновности, свобода слова, право на частную жизнь и конфиденциальность;
- равные возможности для получения образования, работы, социальных выплат и финансовой стабильности в целом.

2. *Группы лиц*. Дискриминация, несправедливое обращение по половым, расовым и иным признакам:

- групповая изоляция;
- нарушение прав детей, пожилых людей и людей с ограниченными возможностями.

3. *Общество*. Международная и национальная безопасность (автономное оружие, дезинформация):

- демократия (выборность органов, доверие к власти);

- целостность информации (дезинформация, фейки, персонализированные новости, «информационные пузыри», снижение доверия к новостям);
- верховенство закона (функционирование и доверие к институтам, судебной системе в целом);
- обеспечение безопасности (военные и полицейские цели);
- культурное разнообразие и изменяющиеся человеческие отношения (фальшивые друзья).

4. *Экономика*. Неравные экономические возможности;
- концентрация мощностей;
- технологическая зависимость;
- распределение ресурсов;
- недостаточное/чрезмерное использование ИИ, техноцентричный подход;
- стабильность финансовых систем.

5. *Экосистемы*. Нагрузка на окружающую среду/климат/природные ресурсы.

Риски с точки зрения критических инфраструктур.

6. *Ценности и нормы* [22]:
- этические ценности;
- моральные ценности;
- социальные ценности;
- культурные ценности;
- правовые нормы.

Необходимо осознавать возможные негативные последствия использования ИИ. Для этого требуется обмен информацией для выявления проблем и управления мерами реагирования на них. Поскольку есть еще область непредвиденных последствий действий ИИ, это требует также всеобщей кооперации, а для этого нужно прежде всего обсуждать социально-политические проблемы на международном уровне.

Сверхскоростное улучшение технологий ИИ, в частности генеративного ИИ, сделало возможным создание различных аудио- и видеоматериалов, содержащих ложную информацию, что приводит к проблеме распространения дезинформации. В США уже активно вводятся ограничения на использование инструментов ИИ, поскольку изменяемые аудио- и видеоматериалы могут в действительности вводить людей

в заблуждение. Общественные исследования показывают, что людям становится всё сложнее отличать сгенерированную информацию от реальной, особенно в силу усовершенствования технологий.

Одним из предложенных способов решения проблемы является наглядная маркировка созданных ИИ материалов (например, текст или графика, наложенные на видеозаписи). Закон о маркировке ИИ 2023 года (AI Labeling Act of 2023) требует раскрывать информацию об использовании технологий ИИ.

В июле 2023 года семь ведущих технологических компаний США обязались разработать «надежные технические механизмы» для оповещения пользователей в случаях, когда технологии создаются ИИ [23].

Хотя мнение о том, что маркировка обеспечит полную эффективность использования ИИ, является все еще дискуссионным, всё-таки предполагается, что она поможет более четко разграничивать подлинный контент от неподлинного. Для понимания необходимости маркировки нужно знать ее основные цели.

1. Оповещение пользователей о том, что какая-то конкретная часть контента была создана или отредактирована ИИ.

2. Снижение вероятности того, что контент вводит в заблуждение или непосредственно обманывает пользователей.

Можно выделить ряд задач, которые следует решить для целей маркировки ИИ:

- определить, какие именно типы контента необходимо маркировать и какие инструменты использовать при масштабировании этой задачи;
- учесть, к каким выводам могут прийти пользователи, которые увидели как маркированные, так и немаркированные материалы, и какую реакцию это может вызвать с их стороны;
- проработать различные способы маркировки в контексте различных стран и групп населения, а также в зависимости от разных форматов представления информации.

Отдельным вопросом является использование ИИ в военных целях. Оно может и должно соответствовать нормам и концепциям международного права. Использование ИИ в военных задачах должно быть подотчетным для сокращения рисков, связанных с несчастными случаями или непреднамеренным нанесением вреда здоровью. Необходима четкая документальная фиксация ответственной разработки и использования потенциала военного ИИ [24].

В 2018 году Минобороны США заручилось поддержкой Google для машинного обучения беспилотников. Это позволило бы технике обрабатывать гигантские объемы визуальной информации и соответственно улучшить наведение ударов. Так появился проект Пентагона Maven. Информация просочилась в СМИ, и 4 тыс. сотрудников Google запротестовали. В конце концов компания отказалась от сотрудничества с Пентагоном и даже ввела запрет на разработку ИИ для военных [25].

Не менее важным вопросом является соблюдение расового и социального равноправия. К примеру, в США алгоритмы ИИ были применены в сфере медицины для оптимизации распределения пациентов для дальнейшего лечения, однако система отработала не так, как было запланировано. Программа относила темнокожих с тяжелыми заболеваниями и белых, но более здоровых, к одной группе, так как посчитала, что те, кто тратят больше средств на лечение, нуждаются в помощи больше. Система дает сбои и в плане распознавания лиц, что может приводить к задержанию невиновных по подозрению в том или ином деянии. Не зря в вопросы этики вовлечены и такие крупные организации, как ЮНЕСКО, выпустившая в 2021 году Рекомендацию по этике ИИ, а также ОЭСР и ее «Принципы искусственного интеллекта» 2019 года [26].

Подводя итоги рассмотренному в настоящем подразделе, надо отметить следующее.

Создание руководств по применению ИИ развивается за счет распространения сходных принципов в разных странах и вовлеченности крупных организаций. Международная кооперация – ключ к пониманию работы систем ИИ, поскольку это технология общего назначения и может широко использоваться в различных задачах. Классификация риск-факторов использования ИИ – определение направлений работы в области ИИ, а также способ понимания возможного ущерба, наносимого ИИ.

1.3. ОТВЕТСТВЕННОСТЬ ПРИ СОЗДАНИИ И ИСПОЛЬЗОВАНИИ ИИ

Ответственность – это некое обязательство лица, связанное с осознанием им того, что именно на него ложится бремя по устранению всех последствий, которые могут возникнуть в результате предпринятых действий или бездействия. Понятие ответственности тесно сопряжено

с корректным восприятием и отражением действительности (сознанием) и пониманием, что обязанность – долг, основанный на социальном положении лица (этика в философии как раз указывается в качестве учения об обязанности) [27].

Особенности ответственности

- Наличие лиц(а), перед которым будет нестись ответственность (государство, различные социальные группы).
- Законодательное подкрепление ответственности (четко фиксированные нормы, которые определяют, с одной стороны, ее границы, с другой – меры воздействия на ответственное лицо, а также инструменты наказания).
- Возможность ее принимать, вменять, отвергать, брать.

Объектами ответственности могут быть, например, люди, животные, окружающая среда, материальные и нематериальные блага.

Ответственность подразделяется:

- на приобретенную естественным или случайным образом (например, родителями);
- сознательно принятую в соответствии с социальным статусом (должностным лицом) или посредством принятия условий двустороннего соглашения (договор на оказание услуг).

Ответственность бывает естественная и контрактурная, согласно трактовке Х. Йонаса (немецкого и американского экзистенциалиста) [28].

В понятие «ответственность» включается также обязанность, которая накладывается на какое-либо лицо определенной частью общества. Обязательством также может быть принятая лицом ответственность из чувства личного долга. Таким образом, человек вправе принимать решения и совершать действия, но он должен осознавать степень своей ответственности за их последствия и понимать, что ее бремя остается на нем и не может перекладываться на других.

«Свобода одного заканчивается там, где начинается свобода другого» – М.А. Бакунин [29].

Одним из важных критериев вменения ответственности является определение вины, поскольку вина, согласно общепринятым юридическим нормам, входит в состав правонарушения (является субъективной стороной). Вина – это психическое отношение лица к совершенному деянию (действию или бездействию). Вина в российском праве подраз-

деляется на умысел и неосторожность. В правовых системах разных стран определения могут немного варьироваться. К примеру, в Китае формула вины включает такой признак, как осознание общественной опасности [30].

Прямой умысел означает, что лицо осознавало опасность и противоправность деяния, а также предвидело общественно опасные последствия, желало их наступления (например, взлом интернет-аккаунтов с целью получения экономической прибыли). Косвенный умысел – лицо осознавало опасность и противоправность деяния, а также предвидело общественно опасные последствия, НО не желало их наступления, однако допускало такую возможность либо относилось к этому безразлично (например, незаконное распространение личных данных вопреки ограничивающим нормам) [31].

В России появились новые примеры мошенничества с использованием искусственного интеллекта: сначала преступники намеренно взламывают аккаунт человека в соцсети, после чего пишут сообщение его знакомым с просьбой перевести деньги. Для того чтобы история выглядела для жертв более убедительно, мошенники используют ИИ, генерируя сообщения на основе уже имеющихся аудиосообщений владельца [32]. В таком случае виден прямой преступный умысел.

Согласно статистике судебного департамента Верховного суда, в России каждый год осуждают примерно 500 человек за преступления, которые тесно сопрягаются с неправомерным доступом к базам данных, а также продажей личной информации. По словам экспертов, чаще всего утечка персональных данных происходит из госкомпаний, банков и мобильных операторов. В таких случаях работники не осознают степень вреда своих действий, хотя такую информацию в дальнейшем могут использовать мошенники, например, для оформления кредитов онлайн, когда не требуется личное присутствие заемщика [33]. Это пример косвенного умысла.

Легкомыслие – лицо осознавало опасность и противоправность деяния, а также предвидело общественно опасные последствия, НО самонадеянно рассчитывало их избежать / предотвратить (например, превышение скорости на крутом повороте с расчетом на навыки управления транспортным средством, в итоге попадание машины в неконтролируемый занос).

Небрежность – лицо не осознавало опасности и противоправности деяния, а также не предвидело общественно опасных последствий, хотя

должно было (например, проведение медицинского процесса с нарушением правил привело к ухудшению состояния пациента) [34].

Например, прокуратура Детройта (Калифорния) в 2019 году предъявила два обвинения в непредумышленном убийстве водителю Tesla, ехавшему на автопилоте, когда он проехал на красный свет, врезался в другую машину, в которой погибли два человека. Компания при этом заявила, что автопилот и более сложная система «Полного самостоятельного вождения» не могут управлять автомобилем самостоятельно и что водители должны быть внимательны и готовы отреагировать в любой момент. Это является примером легкомыслия [35].

Суперкомпьютер IBM Watson дал небезопасные рекомендации по лечению онкологических больных: в 2012 году врачи онкологического центра Memorial Sloan Kettering в партнерстве с IBM обучили Watson диагностике по лечению пациентов. Однако впоследствии в документальном отчете компании было указано, что суперкомпьютер часто давал плохие советы. Например, однажды предложил больному раком с сильным кровотечением дать препарат, который мог усилить кровотечение [36].

Уровень ответственности зависит от степени разрушений, которые возникают вследствие совершения тех или иных действий (чем более серьезны последствия разрушительной деятельности, тем больше ответственность). Уровень ответственности зависит также от степени рисков при осуществлении той или иной деятельности (т. е. чем больше вероятность возникновения непредсказуемых событий, тем выше уровень ответственности). Ответственность чаще всего подразумевает принятие решений в соответствии с выбранной стратегией поведения.

Логика человека считается обычно предсказуемой, обусловленной принятой для достижения цели стратегией (резкая смена стратегии происходит только в форс-мажорных случаях). В противовес этому для ИИ более характерны случайные решения, так как он принимает их в зависимости от новых поступающих данных (принцип черного ящика). Стохастичность его работы делает взаимодействие с ним более опасным, а также не снимает ответственности за принятые решения и действия с человека [27].

При анализе ответственности важно учитывать влияние ИИ на финальный результат.

К слабому влиянию на финальный результат относится ИИ, который применяется в текстовых редакторах для исправления ошибок,

стилистики, в фотокамерах для автонастройки (ISO, выдержка, размер диафрагмы и т. д.), в аудиоредакторах для шумоподавления. ИИ в таких случаях лишь косвенно влияет на результат и почти не раскрывает своего творческого потенциала.

При использовании ИИ в качестве инструмента, влияющего на финальный результат, он проявляет свой творческий потенциал, но не в полной мере. Свобода творчества ограничена контекстом или небольшим количеством данных для анализа и обучения. Такой ИИ, например, изменяет изображения в графических редакторах.

В случае проявления творческого потенциала ИИ в полной мере финальный результат слабо предсказуем. Системы машинного интеллекта самообучены, а создание финального результата полностью непредсказуемо и является целью функционирования ИИ [37].

В ходе вменения ответственности возникает много вопросов. Кому следует вменять ответственность за действия машины в каждом конкретном случае? Производителю? Автору программы? Собственнику устройства? Разработчику системы ИИ? Непосредственно системе ИИ? Какой будет форма ответственности? Каковы инструменты возложения ответственности?

Юристами прорабатываются разноплановые подходы к определению правового статуса ИИ. Некоторые рассматривают в том числе абсолютное равенство ИИ в правах с человеком. Однако более приближенной к жизненным реалиям видится теория, согласно которой правовое положение ИИ будет максимально приближенным к юридическим лицам. С другой стороны, по мнению ряда юристов, такой статус ИИ может способствовать возложению на него юридической ответственности вместо реальных виновников [38].

Три основных подхода к ИИ с точки зрения его правового регулирования

1. Искусственный интеллект является субъектом права, по содержанию приближенным к физическому лицу, но есть также предложение ввести отдельный правовой статус «электронного лица».

2. Искусственный интеллект является объектом права и должен по статусу приравниваться к животным, т. е. является имуществом.

3. Искусственный интеллект является лишь техническим средством.

Рассмотрим пример из практики.

В России Минэкономразвития предложило механизм компенсации вреда, обусловленного ИИ. Поправки с целью апробации нововведений были предложены в закон «Об экспериментальных правовых режимах в сфере цифровых инноваций». В соответствии с этим предполагается, что участники экспериментальных правовых режимов (ЭПР) в области цифровых инноваций (цифровых песочниц) будут обязаны страховать ответственность за негативные последствия их применения. В случае наступления таких последствий анализом будет заниматься специальная комиссия при участии регуляторов представителей исполнительной власти [39].

Субъекты ЭПР будут нести обязательства по ведению реестра лиц, с которыми возникли правоотношения, содержащего информацию, касающуюся ответственности за использование решений, основанных на ИИ, а также информации о лицах, которые при нештатных ситуациях будут ответственны за возникающие последствия использования этих технологий. Участникам предпишут страховать гражданскую ответственность за причинение вреда жизни, здоровью или имуществу других лиц в результате использования ИИ.

Сравнение ИИ и человека. Правовое регулирование ИИ осложняется сомнениями о его возможности имитировать когнитивные (умственные) функции человека. Во многих исследованиях ИИ представлен как система, способная осуществлять автономный мыслительный процесс, рассуждать, принимать решения без подкрепления, однако в настоящее время о полном соответствии ИИ способностям человека говорить не приходится. Необходимо также учитывать, что само понятие «интеллект» может включать в себя множество разных видов интеллекта, таких как эмоциональный, вербальный, эстетический, социальный.

Как поступать с объектами интеллектуальной собственности? Люди всё чаще прибегают к помощи ИИ, поэтому рассмотрим пример. Рассмотрим прецедент победы искусственного интеллекта (нейросети DALL-E) в престижном фотоконкурсе Sony World Photography Awards 2023, примечательный тем, что лицо, которое использовало ИИ для генерирования фотографии, изъявило желание признать его соавтором указанной работы, к чему организаторы конкурса отнеслись скептически [40]. В России, например, автором произведения, согласно Гражданскому кодексу, может быть только человек.

Поэтому возникает вопрос: кого следует считать автором объекта интеллектуальной собственности?

Есть несколько возможных подходов к определению автора результата интеллектуальной деятельности:

- разработчик программного обеспечения;
- пользователь;
- собственник оборудования, используемого для запуска и функционирования программы;
- искусственный интеллект (ИИ).

При определении ответственности прежде всего следует обратить внимание на следующие факторы.

- Возник ли ущерб в ходе эксплуатации и придерживался ли пользователь инструкции по эксплуатации?
- Зафиксированы ли в системе какие-либо общие или конкретные ограничения по использованию ИИ и были ли они донесены до пользователя?

От последующего ответа зависит степень вины пользователя ИИ или собственника технологии при возникновении конфликтной ситуации.

Рассмотрим примеры из практики.

Оператор беспилотного летательного аппарата не учел погодных условий, и дрон летел слишком близко к высокому зданию. В результате из-за сильного ветра он потерял устойчивость и столкнулся с фасадом здания, что повредило как сам дрон, так и остекление здания, хотя в инструкции по эксплуатации дрона было указано, что устройство не следует использовать в ветреную погоду. Этот случай демонстрирует важность тщательного планирования и учета всех факторов перед полетом на дроне.

Если ущерб нанесен из-за некорректного обучения системы искусственного интеллекта, кого следует считать виновным: разработчика или поставщика данных, или заказчика?

В 2016 году 22-летний Ричард Ли хотел получить новый паспорт, заполнив форму и загрузив фото на сайте МВД Новой Зеландии. Сайт выдал ошибку: «Глаза человека на фото закрыты» и отклонил заявку Ли. Конечно, глаза его были открыты. Молодой человек имел узкий разрез глаз, который система не смогла распознать (и это случалось в 20 % подобных случаев, как потом выяснилось), потому что привыкла к облику европейского типа, который имеют большинство жителей страны [41].

Возможно ли отследить возникший ущерб до стадии разработки или производства системы ИИ, что позволит более точно определить ответственность автора или производителя и др.?

В 2018 году компания Toshiba создала цифрового двойника для оживленного направления Лондон – Кембридж британской железной дороги Greater Anglia. Перевозчику понадобилось новое расписание, учитывающее все пики и спады нагрузки на маршрут. Обычно для этого брали исторические данные, но в этот раз железнодорожники решили опираться на информацию от цифрового двойника. На нем железнодорожники испытывали изменения в расписании и искали способы повысить его точность, чтобы при этом не навредить реальным пассажирам, если принятые решения окажутся неэффективны.

Подводя итоги рассмотренному в настоящем разделе, можно отметить следующее.

Ответственность – неотъемлемая часть использования ИИ, закрепленная законодательно и имеющая механизмы контроля и воздействия на ответственное лицо.

В текущий момент сделать работу ИИ автономной нельзя, поскольку необходимо определять ответственность, используя понятие **вины**.

Существует несколько основных направлений подходов к правовому регулированию: ИИ как субъекта права, как объекта права, как технического средства.

1.4. ИИ В ЭЛЕКТРОЭНЕРГЕТИКЕ

Роль ИИ в электроэнергетике

Энергетика – одна из отраслей, в которой критически важно обеспечивать защиту энергосистем от технологических нарушений. В этой связи для поддержания нормальной работоспособности объектов энергетики и минимизации рисков для потребителей всё более активно используются информационные модели, которые позволяют выявлять эти риски. Такие модели строятся на основе аналитики и изучения существующих систем, находящихся в эксплуатации, а активным помощником в этом вопросе становятся системы ИИ, целью которых является снижение стоимости рисков возможных ошибок.

К основным задачам ИИ в энергетике можно отнести:

- задачи прогнозирования (применение ИИ с целью прогнозирования выработки и расходования электроэнергии, а также оптимизации режимов работы оборудования и т. д.);
- задачи повышения энергоэффективности (рациональное использование ресурсов);
- задача контроля процессов (мониторинг состояния оборудования, алгоритмы его функционирования, управление оборудованием) [42].

Примеры из практики

В США (штат Колорадо) местная энергетическая компания Xcel сотрудничает с компанией Itron для улучшения работы умных счетчиков с помощью Интернета вещей. Экосистема направлена на обеспечение более высокого уровня информированности о работе сетей, что поможет минимизировать риски сбоев и улучшить реагирование на непредвиденные ситуации [43].

Совместный проект компаний IBM и Министерства энергетики США под названием SunShot, в рамках которого самообучаемая программа дает возможность достоверно прогнозировать выработку электроэнергии возобновляемыми источниками (ветряными, солнечными и гидроэлектростанциями). Искусственный интеллект обрабатывает большое количество ретроспективных данных, а также собирает информацию о погодных условиях в режиме реального времени [44].

Компания DeepMind Technologies Ltd. (основана в Лондоне в 2010 году, однако к 2014 г. уже поглощена компанией Google) уменьшила расход электричества центром обработки данных Google на целых 40 %. Это было сделано с помощью оптимизации работы центра обученной нейросетью на основе данных большого количества сенсоров, размещенных на территории центра [45].

Почему человеку предпочитают ИИ?

- Колоссальная экономия временных ресурсов – большая скорость обследования оборудования и обработки данных.
- Постоянный процесс актуализации информации с учетом обновляющихся входных данных.
- Меньшие риски с точки зрения финансовых потерь в результате своевременного обнаружения дефектов в работе оборудования, а также оптимизации расходов на техническое обслуживание.

• Большие технические возможности для осуществления разного рода задач – вычислительные мощности, наличие дополнительного технического оснащения.

• Поддержание безопасности работы персонала (избежание рисков, связанных с работой с высоким напряжением, проведением работ на высоте).

Недостатки использования технических средств на основе ИИ

• Первоначальные временные и денежные затраты в такого рода средства могут оказаться достаточно большими.

• Неспрогнозированные поломки, которые могут требовать навыков высококвалифицированного персонала, а также быть дорогостоящими, что может тормозить рабочие процессы и вести к финансовым потерям.

• Замена человеческого труда на машинный остро ставит вопрос потери рабочих мест, а соответственно растущей безработицы [46].

ИСПОЛЬЗОВАНИЕ ИИ В ЭЛЕКТРОЭНЕРГЕТИКЕ

Цифровые подстанции

В энергетике ИИ используется в системах автоматизации управления технологическим процессом (АСУ ТП). Сюда входят системы SCADA, системы сбора, мониторинга и управления, контроллеры – всё, что необходимо для взаимодействия с оборудованием для ответного реагирования на возникающие отклонения в работе устройств. Основная трудность заключается в том, что реагирование должно быть очень быстрым, при этом необходимо опросить большое количество датчиков и сенсоров, обработать данные, поскольку несвоевременное принятие соответствующих мер может привести не только к повреждению оборудования, но и к системным авариям и обесточиванию [47].

Робототехнические комплексы

Киберфизической системой называется программно-аппаратный комплекс, представляющий собой единую систему робототехнических комплексов и вычислительной техники, которые позволяют реализовать автоматический сбор и обработку данных на объектах энергетической инфраструктуры. Такие комплексы эффективны для сбора большого количества данных и служат для эффективного управления техническим состоянием оборудования. Такая система поддержки принятия решений

позволяет улучшить точность определения состояния высоковольтного оборудования станций и подстанций, а также увеличить скорость сбора этих данных [48].

Беспилотные воздушные суда

Беспилотные воздушные суда (БВС) – летательные средства, помогающие обследовать ЛЭП, объекты энергетики, а также выполняющие ключевую роль при ликвидации аварийных ситуаций путем определения самого опасного объекта. Дополнительно осуществляют функцию картографирования. Важность БВС обусловлена нехваткой актуальной информации в организациях, труднодоступностью объектов энергетики, низким качеством работы подрядчиков. С помощью БВС автоматизируется сбор данных, что позволяет сократить количество аварий и отследить действия подрядчиков – происходит колоссальная экономия временных ресурсов.

Пример из практики

«Россети Центр Белгородэнерго» – одно из ответвлений крупнейшего электросетевого холдинга России – практикует применение БВС с 2019 года. По результатам проведенных работ были обследованы большие расстояния ЛЭП (1000 км и более), итогом чего стало выявление 300 дефектов. Дополнительным преимуществом применения БВС можно назвать отсутствие необходимости вывода ЛЭП в ремонт при осмотре, что способствует выявлению причин неисправности с меньшими потерями. БВС в подобных задачах активно используются также за рубежом.

К **рискам использования БВС** относится возникновение электрического пробоя через корпус БВС, а также возникновение коронного разряда из-за возмущения внешнего электрического поля.

Оба варианта увеличивают вероятность выхода из строя оборудования и даже могут привести к прекращению электроснабжения.

Существуют также риски столкновения с препятствиями. Поэтому БВС могут применяться для осмотра объектов лишь на удалении, затрудняя их использование, к примеру, на открытом распределительном устройстве.

Инспекционные робототехнические комплексы

Робототехнические системы используются для операционного и планового осмотра ЛЭП, ее диагностики, а также для обслуживания

подземных кабелей, подстанций и пр. Первым инспекционным роботом, который передвигается по фазным проводникам или на стальных тросах, стал разработанный в 2000 году в столице Японии робот – проект института Kansai Power Corporation и HiBot.

Сходная технология нашла свое отражение и в проекте другой компании – IREQ, которая разработала платформу LineScout, представляющую возможность перемещения по измененным в реальном времени конфигурациям. Подобная конструкция с функцией «переползания» по проводам была разработана и в Шанхае для ЛЭП напряжением 500 кВ. В нее также включена роботизированная рука, используемая для осмотра компонентов ЛЭП или иного оборудования в режиме реального времени. Питается робот от проводов либо от солнечных панелей.

Компания EPRI является создателем роботизированного комплекса под названием Ti, отличием которого является полностью автоматизированное устройство. В нем батареи подзаряжаются от проводников роботизированной платформы, эксплуатируются на 138 кВ ЛЭП. Применяется компанией American Electric Power на участке в 120 км, может проходить в среднем по 5 км в день. Оснащенный ИК-камерами комплекс проверяет провода и различные составляющие ЛЭП, расстояние между проводами и деревьями, способен обнаружить коронные и дуговые разряды.

Наземные робототехнические комплексы

Помимо инспектирования воздушных ЛЭП, существуют и наземные устройства подобного типа. Они имеют бóльшую маневренность и чаще всего используются для перемещения на энергетических объектах для выполнения ремонтных работ. Диагностика оборудования на подстанциях проводится с помощью роботизированных комплексов, в состав которых входят платформа и центр хранения и анализа данных. Совершенствование этой технологии является одной из наиболее востребованных тем исследования на сегодняшний день, поскольку такая технология способствует выявлению тепловых дефектов и неисправностей.

Техническое оснащение наземных робототехнических комплексов:

- модуль контроля движения комплекса;
- модуль контроля положения в пространстве;
- модуль энергоснабжения;
- блок питания;

- модуль осмотра;
- модуль коммуникации;
- диагностическая аппаратура для фотосъемки.

Проекты по использованию ИИ

В энергетическом комплексе многих стран активно поддерживаются и внедряются проекты с использованием технологий ИИ. Такие технологии способны, к примеру, предупреждать отказы оборудования, экономить ресурсы на техническое обслуживание и минимизировать риски в работе предприятия из-за сбоев. Например, в General Motors за счет использования ИИ на 5 % выросла эффективность ветровых турбин, при этом затраты на техобслуживание были сокращены на 20 %. С помощью ИИ высчитывают параметры, которые сложно определить невооруженным глазом, к примеру такие, как «индекс здоровья» (health index) оборудования [49].

Компьютерное зрение

Обнаружение дефектов при производстве оборудования в энергетической отрасли достигает почти 100 % точности для снижения затрат на его обслуживание. С помощью компьютерного зрения аккумулируются данные, а алгоритмы машинного обучения помогают обнаружить процент отклонения от заложенных стандартов качества. Таким образом, специалисты получают рекомендации по выполнению ремонта и сформированные списки брака [50].

Пример из практики

Углеродное волокно – материал, состоящий из тонких нитей, образованных преимущественно атомами углерода. Волокно поставляется в качестве сырья, в том числе для производства лопастей ветрогенераторов. В процессе производства на углеродных нитях могут появляться различные дефекты: обрыв, узел, ворс (пучок), отклонение толщины жгута и полотна по ширине или по высоте, посторонние включения (капли смолы аппрета, мусор и т. п.). Для декоративного применения качество волокна не особо важно, однако для производства дефектные нити не подойдут. Именно поэтому на производстве внедряются технологии компьютерного зрения для обеспечения точности детектирования [51].

Система 3D Vision

Система применяется на производственной линии для производства электрооборудования. Системой на основе изображений высокого расширения строятся полные 3D-модели компонентов и их соединительных контактов. При прохождении компонентами через завод-изготовитель технология компьютерного зрения создает 3D-модели, захватывая различные изображения. Эта технология широко применима в различных отраслях, в том числе таких областях, как электроника, добыча полезных ископаемых, энергетика (создание чертежей для печати деталей на 3D-принтере).

Профилактическое обслуживание

Нередко отраслевые процессы проходят в тяжелых погодных условиях, поэтому различные формы деградации материала, например коррозия, являются вполне привычными явлениями. Такая проблема требует пристального внимания, поскольку в случае ненадлежащего обслуживания оборудование быстро придет в негодность и выйдет из строя, что повлечет остановку производственных процессов. В этом случае компьютерное зрение может помочь контролировать оборудование на основе множественных показателей и сигнализировать о необходимости проведения превентивных мер по обслуживанию.

Охрана труда и безопасность

Компьютерное зрение применяется для обработки изображений в сравнении с имеющимися данными для выявления различных аномалий и предотвращения опасных ситуаций на производственных линиях и площадках, поскольку персонал часто работает в опасных условиях, где несоблюдение техники безопасности может привести к серьезным травмам и тяжелым последствиям. При аварии система способна оповестить персонал о возникновении аварии и о том, насколько она интенсивна, что поможет своевременно приостановить рабочие процессы и обеспечить безопасность сотрудников.

Пример из практики

Концерн «Росэнергоатом» применил автоматизированную систему видеоанализа на Кольской АЭС в Мурманской области, где компьютерное зрение способно определить 26 видов нарушений по 19 параметрам. С помощью видеокамер определяется, применяют ли сотрудники средства индивидуальной защиты во время работы. Данные обрабатываются

нейросетью, а в случае выявления нарушений передаются ответственным лицам для быстрого реагирования (ранее этот процесс осуществлялся вручную), что позволяет достичь точности в фиксации 95–98 % нарушений [52].

Помощь в управлении большими энергосистемами

К примеру, Китай использует ИИ для предиктивного управления сетью высокоскоростных железных дорог протяженностью 45 000 км. Центр ИИ-технологий в Пекине обрабатывает огромные объемы данных в режиме реального времени со всей страны и может с 95%-й точностью предсказывать возникновение нештатных ситуаций и предупреждать о них бригады техобслуживания не позже чем за 40 минут до их прогнозируемого возникновения. В результате таких предсказаний за 2023 год ни на одной из действующих высокоскоростных железнодорожных линий Китая не случилось ни одного инцидента, потребовавшего снижения скорости составов из-за серьезных проблем с путями, а количество мелких неисправностей путей сократилось на 80 % по сравнению с предыдущим годом (до ИИ) [53].

Коммуникация с потребителями

Российские энергокомпании активно внедряют ИИ при работе с потребителями. Например, АО «Мосэнергосбыт» использует интеллектуальных ассистентов в каналах коммуникации: применяется анализ аудиозаписи для повышения качества обслуживания, роботы-коллекторы осуществляют звонки для напоминания о задолженностях и оплате счетов. Речевой помощник помогает также собирать показания счетчиков, дает разъяснения по начислениям, у него можно узнать подробную информацию о тарифе, запросить установку и замену приборов учета, обратиться за услугами (всего у робота более 80 сценариев по 89 тематикам) [54].

Алгоритмами применяется анализ аудиозаписей, при которой обрабатываются множественные звонки операторов, по которым искусственный интеллект переводит звонки в текст и проводит анализ для улучшения клиентского опыта. Это позволяет выявить новые зоны автоматизации и повысить эффективность работы ассистентов в целом. Анализируется, к примеру, как неделовая лексика, так и паттерны приветствий, прощаний, вежливого общения, представления той или иной продукции для поднятия продаж. В свою очередь это позволяет более

оперативно реагировать на возникающие проблемы обслуживания, а также способствует оптимизации издержек [55].

Подведем краткие итоги.

ИИ в энергетике играет ключевую роль в защите энергосистем от технологических нарушений с учетом критической значимости ошибок отрасли. ИИ оказывает активное содействие человеку.

Достоинства ИИ заключаются в скорости обработки данных и их актуализации, возможности оптимизации финансовых потерь, широком спектре технологических возможностей, включая охрану труда.

Недостатки ИИ: большие вложения в технологии на первоначальном этапе, непрогнозируемые поломки и ошибки, потеря рабочих мест. При этом ведется постепенное **улучшение работы ИИ**.

1.5. РАЗВИТИЕ ИИ В РОССИИ И В МИРЕ

Бесспорно, ИИ оказывает огромное влияние на всех нас при этом существует в нашем обиходе уже не первый год. Связано это с тем, что его возможности развития и границы использования были ранее весьма туманны, сегодня же они значительно расширились и продолжают делать это дальше.

ИИ может быть использован как вспомогательный инструмент для расширения человеческих знаний, так и для развития экономического потенциала множества стран. Однако существует ряд рисков, которые тесно связаны с применением ИИ, а скоростные темпы его развития и непрозрачность в работе отходят от традиционных моделей регулирования общества [56].

Вопросы, связанные с рисками

Несомненно, всесторонние возможности и риски, связанные с использованием ИИ, служат предметом интереса общественности. С точки зрения международной стратегии планирования большую обеспокоенность вызывает доступ к данным, вычислительные возможности, средства, стимулирующие развитие ИИ, а также конкуренция в сфере ИИ. Немаловажен и тот факт, что даже при рациональном и этичном использовании ИИ возникает дисбаланс с точки зрения распределения выгод и рисков, что, в свою очередь, может привести мировое сообщество к еще большей поляризации [57].

Для чего необходимо регулировать ИИ?

Необходимость правового регулирования технологии ИИ обусловлена рисками его применения и стремлением обеспечить равный доступ к технологии ИИ.

Одним из ключевых показателей в этой связи является достижение целей устойчивого развития (ЦУР). Они представляют собой набор из 17 взаимосвязанных целей, разработанных в 2015 году Генеральной ассамблеей ООН в качестве так называемого плана достижения лучшего и более устойчивого будущего в равной мере для различных стран.

Таким образом, в промежуточном отчете по управлению ИИ ООН отражает необходимость глобальной консолидации стран и их широкого участия в обсуждениях по этой проблеме.

Именно Организация Объединенных Наций имеет главное уникальное преимущество – это орган с универсальным членством, выполняющий свою деятельность на основании Устава ООН, в функции которого входит учет многонационального разнообразия народов мира. Таким образом, можно сказать, что эта централизованная система необходима не только для обмена знаниями, но и для согласования принятых норм и принципов с целью их выполнения и формирования подотчетности. Именно то, что в орган входят люди разного пола, возраста и с разным уровнем образования, позволяет опираться на разные мнения как правительственных органов, гражданского общества в целом, так и более узких кругов, к примеру, научных. По итогу широкомасштабных дискуссий ООН был сделан общий вывод о дефиците мирового управления ИИ [58].

Рассмотрим, какие факторы могут помочь в активном внедрении и использовании ИИ.

ИИ имеет возможность влиять на доступ к знаниям и повышать эффективность работы на разных уровнях. Многообразие сфер влияния начиная от сельского хозяйства и заканчивая здравоохранением может обеспечивать экономическую стабильность на государственном уровне. Тем не менее наряду с позитивными последствиями применения ИИ возникают вопросы, связанные с тем, каким именно образом необходимо распределить эти блага среди людей, а также как управлять негативными последствиями, в том числе бороться с растущей безработицей и, самое главное, обеспечивать подотчетность действующих участников и возникающих новых субъектов процесса [59].

Основные моменты глобального управления ИИ

Главным принципом в области разработки, внедрения и использования ИИ является сбалансированность управленческого воздействия. Управление ИИ должно привести к увеличению количества людей, занимающихся этими вопросами и конечно же не должно привести к их отстранению от этих процессов. Равный и открытый доступ к технологиям предусмотрен для того, чтобы избежать расколов, связанных с применением ИИ как внутри стран, так и между ними.

Глобальные вопросы, требующие внимания

Ряд систем ИИ непрозрачен либо из-за своей конструктивной сложности, либо из-за намеренного сокрытия информации о их работе (коммерческой тайной) – всё это ведет к непониманию возникновения источников рисков и ответственности при управлении рисками. Несмотря на глобальный характер вопросов, связанных с ИИ, управление технологией является фрагментированным и территориальным, а национальные границы регулирования конфликтов могут приводить к еще бóльшим напряженностям. В этой связи необходимо подходить к регулированию вопросов ИИ стран-участников именно с точки зрения их конкретной ситуации, так как на территории и в условиях конкретных стран принятые меры могут быть попросту бессмысленными.

Наряду с техническими и политическими препятствиями существуют проблемы, связанные с управлением обществом в целом, так как необходимо учитывать человеческие и экологические издержки, поскольку человеческая жизнь и здоровье, а также вопросы окружающей среды имеют большое значение. Помимо нецелевого использования можно также выделить опасения, связанные с упущенными возможностями применения ИИ – отсутствием способности воспользоваться благами ИИ и поделиться технологиями из соображений излишней осторожности [60].

МЕЖДУНАРОДНАЯ ПРАКТИКА

Саммит Большой Семерки

В октябре 2023 года в Хиросиме был разработан и согласован документ, цель которого – разработка безопасного и защищенного ИИ. Основной целевой аудиторией в этом случае являются организации, занимающиеся разработкой и использующие наиболее современные

и передовые технологии ИИ. На встрече было выделено и согласовано 11 принципов.

1. Применять необходимые меры на всех этапах разработки современных систем ИИ, включая подготовку и внедрение, для выявления, оценки и снижения рисков на протяжении всего их жизненного цикла.

2. Обнаруживать и устранять уязвимости, а также реагировать на инциденты и случаи неправильного использования после внедрения, включая выход на рынок.

3. Открыто информировать о возможностях, ограничениях и приемлемых и неприемлемых способах использования современных систем ИИ, чтобы обеспечить необходимую прозрачность и повысить подотчетность.

4. Содействовать ответственному обмену информацией и сообщениям об инцидентах между организациями, занимающимися разработкой современных систем ИИ, включая взаимодействие с отраслью, государственными структурами, гражданским обществом и научным сообществом.

5. Разрабатывать и внедрять политику управления ИИ и рисками, основанную на подходе, ориентированном на риски, включая политику конфиденциальности и меры по снижению негативных последствий, особенно для организаций, создающих современные системы ИИ.

6. Инвестировать в надежные средства обеспечения безопасности, включая физическую безопасность, киберзащиту и защиту от внутренних угроз, и применять их на всех этапах жизненного цикла систем ИИ.

7. Создавать и внедрять надежные механизмы для аутентификации контента и определения его источника, когда это возможно, такие как водяные знаки или другие методы, позволяющие пользователям распознавать контент, созданный ИИ.

8. Определять приоритеты для исследований, направленных на снижение социальных рисков, а также рисков в области охраны труда и техники безопасности, и акцентировать внимание на эффективных мерах по смягчению последствий.

9. Уделять особое внимание разработке современных систем ИИ для решения наиболее значимых мировых проблем, таких как климатические изменения, глобальное здравоохранение и образование, но не ограничиваться только ими.

10. Поддерживать разработку и принятие международных технических стандартов.

11. Внедрять соответствующие меры для защиты данных и охраны персональной информации и интеллектуальной собственности.

Евросоюз

Страны-члены Евросоюза совместно с Европарламентом пришли к консенсусу, зафиксировав свои договоренности в Законе об ИИ. Основной смысл этого документа заключается в риск-ориентированном подходе, согласно которому ИИ рассматривается как средство, способное причинить вред обществу. Однако важным уточнением является то, что этот закон вступит в силу не ранее 2025 года из-за необходимых процедур его принятия в рамках ЕС. Предварительно согласованный документ закрепил следующие концепты.

1. Определение и область применения (например, закон об ИИ не будет касаться систем, предназначенных исключительно для военных или оборонительных нужд).

2. Классификация систем ИИ по категориям высоких рисков и запрещенных практик (например, закон об ИИ запрещает использовать технологии распознавания лиц и другие системы «удаленной биометрической идентификации» в реальном времени в общественных местах, а также распознавания эмоций и применять полиции системы предиктивной аналитики для предотвращения правонарушений).

3. Ограничения для правоохранительных органов (например, введен специальный порядок, позволяющий правоохранительным органам использовать высокорискованные системы ИИ в экстренных ситуациях).

4. Системы ИИ общего назначения и базовые модели (добавлены новые положения, учитывающие случаи, когда системы ИИ могут применяться для различных целей и когда технологии общего назначения впоследствии интегрируются в другие высокорискованные системы ИИ).

5. Новая структура управления (в рамках Еврокомиссии создается Управление ИИ, которое будет контролировать передовые модели ИИ, способствовать разработке стандартов и методов тестирования, а также обеспечивать соблюдение общих норм всеми государствами-членами).

6. Штрафы (за нарушения закона об ИИ предусмотрены штрафы в виде процента от годового оборота компании-нарушителя за предыдущий финансовый год или в фиксированной сумме).

7. Прозрачность и защита основных прав (закон об ИИ требует оценки воздействия высокорисковых систем ИИ на основные права перед их выходом на рынок).

8. Меры поддержки инноваций (уточнено, что нормативные «песочницы» в области ИИ должны также проводить тестирование инновационных систем ИИ в реальных условиях).

Кроме того, ЕС также принял закон о кибербезопасности цифровых продуктов – первый подобный документ в мировом правовом поле. В нем говорится о повышении уровня кибербезопасности цифровых продуктов предприятий в ЕС, с помощью введения общеобязательных предписаний для технических средств на разных уровнях. Это влечет за собой следующее требование: все продукты рынка ЕС должны быть кибербезопасными. Меры кибербезопасности должны будут внедряться на протяжении всего жизненного цикла продукта от начала его проектирования и разработки до вывода на рынок. Кроме того, производитель будет обязан осуществлять своевременные обновления безопасности для потребителей. Эти меры должны привести к осуществлению более информированного и безопасного выбора.

Мониторинг стандартов ИИ

Европейская организация AI Watch [61] проводит мониторинг уже имеющихся и находящихся в разработке стандартов и проверяет их на соответствие нововведениям. Эти требования включают ряд характеристик.

- Проверенный и качественный сбор данных.
- Наличие технической документации до выхода продукции на рынок.
- Наличие механизма автоматической записи событий.
- Прозрачность и доступность информации о системе ИИ для пользователей.
- Возможности контроля ИИ человеком.
- Точность, надежность и кибербезопасность.
- Наличие внутренних проверок систем ИИ.
- Наличие системы управления рисками.

Ряд пробелов в стандартах можно увидеть в требованиях к наличию данных «технической документации» и в «системах управления рисками». Ряд отклонений выявлен и с точки зрения детальной структуры

требований. К примеру, в технической документации может быть лишь один технический документ.

Ниже представлен ряд документов, принятых в 2022 году.

ISO/IEC TR 24027:2021. Предвзятость в системах ИИ и принятии решений с помощью ИИ

Этот стандарт описывает предвзятость в системах искусственного интеллекта, особенно в контексте принятия решений. В нем представлены методы для измерения и оценки предвзятости с целью выявления и устранения уязвимостей, связанных с этой проблемой. Стандарт охватывает все этапы жизненного цикла систем ИИ, включая сбор данных, обучение, переобучение, проектирование и тестирование.

ISO/IEC 23053:2022. Рамки для систем искусственного интеллекта (ИИ), использующих машинное обучение (МО)

Данный документ устанавливает рамочную структуру для систем ИИ и машинного обучения, описывающую общую архитектуру ИИ с применением технологий машинного обучения. В нем перечислены компоненты системы и их функции в экосистеме ИИ. Документ подходит для организаций любого типа и размера, включая государственные и частные компании, государственные учреждения и некоммерческие организации, которые внедряют или используют системы ИИ.

ISO/IEC TR 24368:2022. Искусственный интеллект – обзор этических и социальных проблем

Этот документ предлагает комплексный обзор этических и социальных вопросов, связанных с ИИ. Он включает информацию о принципах, процессах и методах в этой области; предназначен для технических специалистов, регулирующих органов, заинтересованных сторон и общества в целом; не направлен на продвижение какого-либо конкретного набора ценностей. Документ содержит также обзор международных стандартов, касающихся этических и социальных проблем, связанных с ИИ.

ISO/IEC 38507:2022. Управленческие аспекты использования искусственного интеллекта организациями

Этот документ служит руководством для членов руководящего органа организации по контролю за использованием ИИ с целью гарантировать его эффективное, результативное и приемлемое применение.

Он также предоставляет рекомендации для более широкой аудитории, включая высшее руководство, внешние компании, технических специалистов, ассоциации и профессиональные организации, государственные органы и политиков, а также внутренних и внешних поставщиков услуг (включая консультантов), и аудиторов. Документ охватывает управление как текущим, так и будущим использованием ИИ, а также последствия этого использования для самой организации. Он применим ко всем типам организаций, включая государственные и частные компании, государственные учреждения и некоммерческие организации.

Великобритания

1–2 ноября 2023 года в Великобритании прошел первый международный саммит по безопасному ИИ. Странами-участниками стали 28 стран, в их числе США, КНР, Индия, ЕС. Документ содержит тезисы, похожие на документ, зафиксированный в Хиросиме, и в первую очередь рассматривает риски информационной и биологической безопасности, а также распространение неверной информации. Декларация основана на Международном кодексе поведения и подкрепляет идеи международной кооперации в области определения и управления рисками ИИ. Предлагается включение в процесс различных организаций международного уровня. Предполагается, что саммит будет проводиться регулярно.

Ряд стран и компаний подписали совместные соглашения. Австралия, Великобритания, Канада, Германия, Италия, Сингапур, США, Франция, Южная Корея, Япония, Европейский союз, а также компании Amazon Web Services, Anthropic, Google, Google DeepMind, Inflection AI, Meta (признана в РФ экстремистской), Microsoft, Mistral AI и Open AI договорились о совместной апробации новых моделей ИИ до и после их вывода на рынок. Это было сделано для того, чтобы максимально оптимизировать риски, связанные с новейшими технологиями ИИ [5].

США

Осенью 2023 года Президент США подписал указ об ИИ. Это первое подобное действие американского правительства в области обеспечения безопасности, рекомендаций по охране труда в целом и гражданских

прав в этой сфере. Наряду с упомянутым выше указ президента обязывает разработчиков ИИ-алгоритмов предоставлять данные, полученные при тестировании продукции. Указ закрепляет следующие принципы.

1. Введение новых стандартов безопасности для ИИ: компании, работающие с ИИ, обязаны предоставлять федеральным властям результаты своих испытаний безопасности; министерству торговли поручено разработать рекомендации по внедрению водяных знаков и создать программу кибербезопасности, на основе которого будут разрабатываться инструменты ИИ для выявления уязвимостей в критически важном программном обеспечении.

2. Защита конфиденциальности пользователей путем создания рекомендаций для профильных ведомств, которые помогут оценивать методы обеспечения конфиденциальности в области ИИ.

3. Обеспечение справедливости и соблюдение гражданских прав через разработку рекомендаций для федеральных подрядчиков с целью предотвращения дискриминации при использовании ИИ-алгоритмов, а также создание рекомендаций для повышения роли ИИ в правосудии, включая его применение в вынесении приговоров, оценке рисков и прогнозировании преступности.

4. Защита потребителей: министерству здравоохранения и социальным службам поручено разработать программу для анализа потенциально опасных вариантов использования ИИ в сфере здравоохранения.

5. Исследование возможных последствий влияния ИИ на рынок труда и способов поддержки сотрудников со стороны федеральных властей.

6. Содействие инновациям и конкуренции путем увеличения грантов на исследования в таких областях, как изменение климата, а также обновление критериев для оценки иммигрантов, позволяющих высококвалифицированным из них оставаться в США.

7. Сотрудничество с международными партнерами для внедрения стандартов ИИ на глобальном уровне.

8. Разработка рекомендаций по использованию и закупкам ИИ для федеральных агентств, а также ускорение процесса найма квалифицированных специалистов в этой области правительством [62].

Китай

Китайское правительство также активно финансирует сферу развития ИИ, поскольку считает, что ИИ абсолютно необходим для достижения их целей по модернизации промышленности и превращению страны в технологическую сверхдержаву. Это подтверждается промышленной стратегией Китая под названием «Сделано в Китае-2025». План развития ИИ направлен на то, чтобы к 2030 году Китай стал мировым лидером в этой области.

Положительные стороны развития ИИ

- Большое количество данных в связи с тем, что это самая населенная страна мира – в ней насчитывается 800 миллионов пользователей Интернета.
- Разрешительные законы в области конфиденциальности и обмена данными слабы, что дает компаниям широкую свободу для их сбора и обработки.
- Большие денежные ресурсы – китайские технологические гиганты с большим объемом денежных средств, такие как Baidu, Alibaba и Tencent, стимулируют исследования и применение инноваций в области ИИ.
- Государственная поддержка – Китай активно включил ИИ в свои национальные стратегии (например, трехлетний план действий в области ИИ (2018–2020 гг.)).
- Координация частного сектора – Китай создал «Национальную команду» по ИИ из частных компаний, которые будут создавать ключевые платформы приложений ИИ.
- Преимущество ранней активности – на китайские фирмы приходится 20 % мировых компаний в области ИИ, что уже делает Китай вторым по величине рынком ИИ после США.

Проблемные моменты, которые необходимо преодолеть

- Низкое качество данных – качественные промышленные данные не всегда доступны из-за отсутствия инфраструктуры, в том числе датчиков, на крупных промышленных объектах.
- Этические вопросы – полиция и военные полагаются на технологии ИИ в вопросах общественной безопасности и наблюдения, что приводит к социальному недовольству.

• Ограниченность знаний – в университетах Китая недостаточно специализированных академических программ по ИИ, и не хватает квалифицированных кадров в области ИИ, как и во всем мире.

• Государственная поддержка раздроблена – несмотря на усилия правительства по ее укреплению, в большинстве случаев развитие ИИ по-прежнему стимулируют частные компании, а провинциальная политика не унифицирована.

• Отсутствие политической инфраструктуры. Надлежащая нормативная база и стандарты лицензирования по-прежнему отсутствуют, что препятствует скоординированному развитию.

• Зависимость от иностранных технологий – китайские инновации в области ИИ иногда отстают от иностранных аналогов, и Китай по-прежнему полагается на зарубежные критически важные технологии, такие как производство полупроводников [63].

Основные документы, принятые в Китае за последние несколько лет

• Шанхайские правила о содействии развитию индустрии ИИ, принятые в октябре 2020 года.

• Правила о содействии развитию индустрии искусственного интеллекта в особой экономической зоне Шэньчжэня, вступившие в силу 1 ноября 2022 года.

• Положения об алгоритмическом управлении рекомендациями служб Интернета, вступившие в силу 1 марта 2022 года (устанавливают структуру управления для регулирования систем рекомендаций).

• Положения об управлении глубоким синтезом служб Интернета, вступившие в силу 10 января 2023 года и применяются к дипфейковым результатам технологии ИИ.

Основной целью принятых документов является содействие здоровому развитию и стандартизированному применению генеративного ИИ, защита национальной безопасности и общественных интересов, а также защита законных прав и интересов граждан, юридических лиц и других организаций в соответствии с китайскими законами, такими как Закон о защите персональных данных.

Основные принципы, отраженные в перечисленных документах

• Уважение к «социальной морали и этике» Китая и соблюдение «основных социалистических ценностей».

• Эффективные меры должны быть задействованы в ходе разработки алгоритмов, выбора данных для обучения, создания и оптимизации моделей, предоставления услуг и других процессов для предотвращения дискриминации по таким факторам, как раса, этническая принадлежность, религиозные убеждения, национальность, регион, пол, возраст, профессия или состояние здоровья.

• Уважение прав интеллектуальной собственности, коммерческой этики и защита коммерческих секретов, а также запрет на использование в целях монополии и недобросовестной конкуренции.

• Уважение законных прав и интересов других лиц и запрет на создание угрозы их физическому и психологическому благополучию или нарушение их прав и интересов, включая имидж, репутацию, честь, частную жизнь и личную информацию

• Необходимость использовать эффективные меры для повышения прозрачности, точности и надежности услуг генеративного ИИ [64].

Анализ развития ИИ в России и мире позволяет сделать следующие выводы.

ИИ – не новая, но во многом неизведанная технология с быстрыми темпами совершенствования. Оказывает влияние на общий прогресс, а значит, требует всеобщей заинтересованности и кооперации со стороны мирового сообщества.

Механизмы управления ИИ – открытые, справедливые, эффективные механизмы управления, обмен информацией и сотрудничество на международном уровне.

Цели развития – ИИ не должен использоваться в целях манипулирования, распространения дезинформации, должен соответствовать всеобщим ценностям.

1.6. ВЛИЯНИЕ ТЕХНОЛОГИЙ ИИ НА ЭКОНОМИКУ

Экономические связи

Развитие искусственного интеллекта существенно повлияло на некоторые важнейшие области экономики, и на сегодняшний день с прогрессированием новейших технологий ИИ число затронутых им секторов экономики продолжает расти.

Таким образом, ИИ стимулирует процесс цифровой трансформации.

В настоящее время технологии ИИ переходят от ограниченного искусственного интеллекта к гибридному (основан на использовании

глубокого машинного обучения и во многом приближен к способностям человека, представляя собой модель нейросетевого человеко-машинного интеллекта), а значит, экспертные системы, основанные на нем, могут все чаще использоваться для управления крупными системами, что дает экономическое преимущество не только на уровне одного государства, но и в межгосударственном плане [65, 66].

Как ИИ влияет на экономику?

К примеру, в России, по официальным данным, необходимо довести уровень внедрения ИИ в экономику до 50 %, при этом Россия находится уже в десятке стран-лидеров по внедрению технологий ИИ, в 2023 году этот показатель составлял 31,5 %.

Поддержка подобных проектов серьезно финансируется правительством. Согласно дорожной карте развития ИИ России, правительство до 2030 года направит около 24,6 млрд рублей на развитие этих технологий. Что касается объема внебюджетного финансирования, то он ожидается в размере 112,6 млрд рублей.

Согласно показателям по внедрению технологий ИИ лидируют секторы ИТ и коммуникационный (уровень внедрения технологий искусственного интеллекта здесь доходит до 53 % компаний). Далее идут организации топливно-энергетического комплекса, промышленности, транспортные компании и медицинская сфера [67–69].

ИИ в цифрах

Наука – российскими исследователями в 2022 году было написано около 2,5 тысяч публикаций в рамках разнообразных конференций, что позволило подняться России на 11 место и оказаться в десятке по этому показателю.

Бизнес – согласно показателям на 2022 год рынок искусственного интеллекта составлял около 650 млрд рублей.

Государственная поддержка – государственное финансирование ИИ в стране по сравнению с предыдущими двумя годами увеличилось в три раза в 2022 году [70, 71].

Внедрение ИИ вызывает различные вопросы.

На уровне частных лиц

Не исчезнет ли с развитием технологий ИИ существующая работа, насколько она будет востребована в будущем? Какими навыками должен обладать современный специалист в своей сфере? Каким образом

будет преобразовываться рабочее время с учетом автоматизации процессов? Как развитие ИИ повлияет на увеличение пенсионного возраста? Как необходимо выстраивать процесс обучения, в том числе школьников и студентов? [65].

На уровне предприятий

Какие виды деятельности можно отнести к перспективным? Какие области предпринимательства наиболее выгодны с точки зрения вклада в развитие робототехники и ИИ? В каких сферах необходимо направлять средства для повышения квалификации сотрудников, какие способы обучения будут наиболее эффективными? Где замена рабочих роботами допустима, а где ИИ не сможет заменить человека? В каком случае человек не будет терять свое рабочее место, а будет продолжать взаимодействовать с роботами и ИИ? Какие новейшие отрасли могут возникнуть в будущем?

На уровне правительственных структур

Каким образом активное внедрение ИИ в различные сферы жизни будет отражаться на экономических показателях и какие проблемы могут в этой связи возникнуть в вопросах экономической политики? Какие приоритетные направления по изменению законов и механизмов регулирования необходимы, чтобы наиболее эффективно взаимодействовать обществу и интеллектуальным системам? Каким образом государству необходимо реформировать систему образования для ее соответствия современным реалиям жизни? Есть ли необходимость в существенной реформации налоговой системы для того, чтобы компенсировать усиливающееся неравенство из-за внедрения технологий ИИ?

История промышленных революций

Первая промышленная революция происходила в период XVIII – первой половины XIX в.

Активно трансформировалась аграрная экономика, проводилась механизация производства, что вызвало огромный подъем производительности труда. Была освоена добыча угля, создан паровой двигатель и множество других технологий в различных областях. Выросли города в связи с созданием большого количества фабрик и заводов, развивалось образование.

Вторая промышленная революция происходила во второй половине XIX и начале XX в. Электрификация позволила внедрить массовое производство, а главной инновацией стало применение конвейера, представленного Генри Фордом. Активно распространялось электричество, велась разработка таких источников энергии, как нефть и газ. Развивался транспорт, стал применяться телеграф, возникали и развивались новые отрасли: электроэнергетика, нефтехимическая промышленность, автомобилестроение.

Третья революция началась в 1960-х гг. XX столетия после окончания Второй мировой войны. Ознаменовалась появлением цифровых технологий, когда появилась электроника, телекоммуникация и компьютеры, а также промышленные роботы. Активное развитие получили космические технологии и исследования. Развивалась связь, технологии сбора и обработки информации, появились Интернет и мобильная связь.

Цифровизация стала фундаментом для четвертой промышленной революции, которая происходит в настоящее время. Ее можно охарактеризовать новыми технологиями, которые влияют на все отрасли в совокупности. Основными приоритетами является развитие науки, подготовка высококвалифицированных кадров, развитие передовых технологий и различных уникальных компетенций [74, 75].

Процесс экономического развития

Кратко процесс экономического развития можно описать словосочетанием «промышленная революция», но на деле это не единичное событие, а более сложное явление, происходившее в течение нескольких десятилетий. Важный аспект – она не была лишь промышленной (точнее сказать, не в первую очередь), но и включала в себя успехи сельского хозяйства, торговли, финансов, которые стали возможны в том числе благодаря изменениям в политической и институциональной сфере. Однако экономический прогресс после нее не был неуклонным и поступательным и потребовалось еще немало времени, чтобы уровень жизни людей стал соответствовать экономической производительности.

График (рис. 1.2) создан экономистом Брэдом ДеЛонгом и описывает события вплоть до 1 млн лет до нашей эры, но его нет смысла расширять, потому что видна будет лишь почти ровная линия, при этом последние, самые важные 200 лет станут практически невидимыми. Из графика видно, что с нулевого года по 1800 уровень ВВП на душу населения удвоился, что является само по себе неплохим показателем,

но основной прирост пришелся на последние года. После промышленной революции показатели выросли в несколько раз и ВВП служит важным критерием измерения и оценки событий, предшествующих появлению роботов и ИИ.

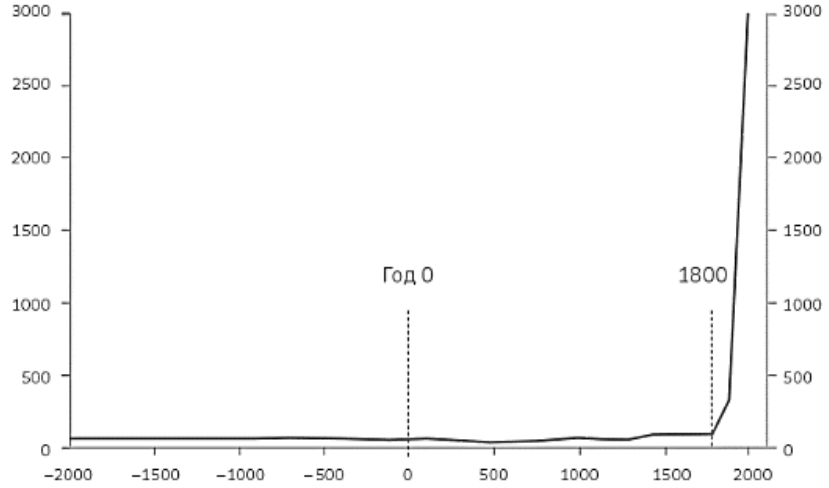


Рис. 1.2. Мировой ВВП на душу населения с 2000 г до н. э. по настоящее время (уровень 1800 г. н. э. условно принят за 100)

Структура экономики

Структура экономики тесно связана с распределением продуктов производства. Для того чтобы технологические новшества одного сектора влияли на производительность экономики в целом, рабочую силу, которая высвобождается в наиболее активно развивающихся отраслях, нужно продуктивно направлять в другие сферы экономики. С другой стороны, технический прогресс не всегда является условием экономического прогресса. Необходимо иметь ресурсы для освоения новых методов производства, в том числе для выпуска инструментов и оборудования, которые необходимы для технического прогресса.

«Самым печальным аспектом современной жизни является то, что наука накапливает знания быстрее, чем общество приобретает мудрость». Айзек Азимов [72].

«Рост производительности – это еще не всё; однако по большому счету это почти что всё». Пол Кругман [73].

Немаловажен и демографический фактор, означающий, что пропорционально росту производства растет и численность населения. Это может, с одной стороны, положительно отражаться на появлении большого количества рабочих рук, а соответственно и увеличении производства, но, с другой, разделение одинакового количества капитала и ресурсов среди большего количества населения не увеличивает производительности труда. Высокие темпы прироста населения на начальном этапе связан с количеством детей, которые с точки зрения экономики являются непродуктивной частью общества. Наконец, экономика сама по себе нестабильна, так как включает в себя множество сопутствующих факторов, как, например, болезни, стихийные бедствия, войны и т. д.

Замена человека машиной – вред или польза?

Стоит отметить, что с исторической точки зрения процесс сокращения рабочих мест в определенных сферах является естественным. К примеру, в XX веке в связи с механизацией транспортных средств повсеместно сокращалось количество извозчиков (в 1915 году количество лошадей в США составляло около 26 млн, а на сегодняшний день эта цифра достигает лишь 10 млн).

Еще одним примером может служить профессия оператора телефонной и телеграфной связи, где число работников на протяжении 100 лет постоянно увеличивалось, а потом резко сократилась с появлением автоматических коммутаторов, а впоследствии уже Интернета и мобильной связи [76].

Тем не менее, несмотря на исчезновение ряда рабочих мест, вместо них появляются новые. Сферы, в которых происходили наиболее яркие изменения с точки зрения производительности труда, нередко приводили к увеличению занятости как раз в этих областях. К примеру, введение компанией «Ford» конвейера на автомобильном производстве привело к тому, что в 1909 году спустя двадцать лет на сборку одного автомобиля уходило менее 50 часов по сравнению с изначальными показателями в более чем 400 рабочих часов. Однако, несмотря на возросшую скорость сборки, количество рабочих мест в автомобильной промышленности резко увеличилось, а эффективность производства повлияла на снижение цен и увеличение спроса на автомобили.

Несмотря на большие суммы денег, вложенные в развитие роботов-помощников, некоторые, на первый взгляд простейшие, задачи остаются неподвластными ИИ. Например, до сих пор не удалось создать

машину, чтобы в достаточной степени быстро сложить полотенце. В 2019 году были анонсированы продажи робота FoldiMate, предназначенного для глажения и аккуратного складывания одежды, но отзывов пользователей об их реальной эффективности пока еще нет [77]. Сюда же относится задача автоматического завязывания шнурков.

Сингапурские исследователи попробовали обучить промышленного робота собирать стул из магазина IKEA, где такая мебель продается в запакованном виде и готова к сборке. Им это удалось, однако для сборки одного стула потребовалось два робота, запрограммированных конкретно для этой задачи, выполнивших ее за 20 минут [78].

ИИ непросто справляется с неоднозначными с точки зрения логики, либо ошибочно сформулированными задачами. К примеру, инструкция на шампуне: «Намылить, смыть, повторить». Для большинства людей будет очевидно, что такую последовательность необходимо повторить дважды. Но ИИ может интерпретировать ее в качестве бесконечного повторения.

Более сложным примером являются случаи, когда высокотехнологичные роботы совершают ошибки. Хирургический робот Da Vinci применялся для проведения операции по удалении опухоли толстой кишки, но машина дала сбой, повредила внутренние органы пациентки, что в конечном итоге привело к летальному исходу [79]. Кроме того, стоимость хирургических роботов довольно велика, и, по данным журнала Healthcare Quality, при хирургическом вмешательстве с участием роботов Da Vinci зафиксированы 174 травмы и 71 смертельный случай.

Какова цена технологий ИИ?

Несмотря на существование различных мнений о будущем развитии технологий ИИ, преимуществах или недостатках их использования, есть экономическая статистика. Во-первых, роботы (к ним здесь относятся в широком смысле слова и программные роботы тоже) – это не бесплатные наемные работники, а только единицы капитального оборудования. Хотя им не начисляется заработная плата, пособия и пенсия, бесплатными их назвать нельзя. Существуют затраты на их строительство и обслуживание, как и на финансирование их развития. Следующий важный аспект – наличие правильного программного обеспечения, которое тоже не возникает бесплатно. Если даже не разрабатывать его постоянно, то необходимо его регулярно обновлять.

Каким образом складывается цена на оборудование?

Использование робота требует больших вложений в основной капитал. Вложения будут зависеть от ряда факторов, которые определяют то, насколько целесообразными являются промышленные инвестиции. В этот перечень входят:

- затраты на оборудование;
- затраты на техническое обслуживание;
- инвестиции и страхование;
- амортизация оборудования.

С усовершенствованием ПО и конструктивных особенностей робота их устаревшие версии могут утрачивать первоначальную ценность.

Например, технологии звукозаписи – когда-то использовались грамофонные пластинки на 78 об/мин, но вскоре им на смену пришли пластинки на 45 об/мин и 33 об/мин. Через некоторое время более востребованными стали магнитные ленты, а их, в свою очередь, заменили компакт-диски. На сегодняшний день и они потеряли свою актуальность, так как есть множество возможностей скачать музыку из Интернета, использовать флешки, телефоны и т. д.

В связи с тем, что роботы имеют начальную стоимость, у людей есть возможность конкурировать с ними даже в том случае, если роботы будут более эффективными: чем выше стоимость роботов, тем больше конкурентных преимуществ у человека. Существуют также широкие возможности для того, чтобы применять наиболее эффективную с технической точки зрения комбинацию человек + роботы (искусственный интеллект). Однако поскольку рынок чаще всего определяется лучшим соотношением цены и предложения, есть прямая корреляция стоимости человеческого труда и расходов на применение робота.

Пример

В период Великой рецессии (2008–2013) существенно снизилась прибыль компаний, но тем не менее многие работодатели в Британии не сокращали сотрудников. В целом их можно было и уволить, а высвободившиеся средства направлять на модернизацию компьютерного оборудования и ПО. Однако работа людей за небольшую зарплату оказалась все-таки более выгодной [80].

Глобальное влияние новых технологий на экономику

В большинстве случаев мировая экономика, как и экономика отдельно взятых государств, довольно непросто приспособляется к большим технологическим изменениям. Наблюдается резкий прогресс экономических показателей, но отдельные люди, общественные группы, а порой даже регионы и страны не способны легко переключаться на только что возникшие новые виды деятельности. Таким образом, технологическая революция помогает ускорить процессы, расширить доступ к данным, но при этом возникает проблема неравномерного распределения преимуществ использования технологий ИИ.

В последние годы есть основания полагать, что распределение доходов и богатства становится более неравномерным. По словам голландского историка Рутгера Брегмана, в США «пропасть между богатыми и бедными сейчас сделалась шире, чем в Древнем Риме, несмотря на то что экономика последнего целиком базировалась на рабском труде» [81].

Основными факторами такого неравномерного распределения являются глобализация и технологические изменения (глобализация добавила ресурсы рабочей силы таких гигантов, как Китай, а технологический прогресс позволил осуществить экономию ресурсов).

Благодаря цифровым технологиям люди получили доступ к лучшим поставщикам услуг на мировом рынке. Цифровые технологии экономят ресурсы при масштабировании бизнеса, что позволяет уничтожить конкурентов и увеличить прибыль. Компания Amazon занимает почти 75 % рынка электронных книг, Netflix является самым используемым сервисом просмотра видеофильмов, а Google оккупировал 90 % рынка поисковой рекламы в мире. В цифровой среде масштабы известности несравнимы с более ранними эпохами (к примеру, успех книги Дж. К. Роулинг о Гарри Поттере) [82].

Территориальное неравенство

Роботы и ИИ могут использоваться на любой территории. Поэтому существует теория о том, что территориальные преимущества не будут значительно влиять на экономическую деятельность, так как промышленность и население будут мигрировать туда, где земля и услуги дешевле. Однако по итогам четвертой промышленной революции, несмотря на то что люди работают и общаются на большом расстоянии друг от друга, рассеивания экономической активности не произошло.

До недавнего времени предполагалось, что использование роботов и искусственного интеллекта равномерно распределится по разным странам, но в реальности это маловероятно. Прежде всего, вопрос упирается в задачи, которые ставят перед собой отдельные страны: они достаточно разные, соответственно у каждой страны есть свой собственный путь экономического развития, по которому она идет. Если посмотреть на историю промышленных революций, то она существенно изменила баланс международных сил, соответственно и новая революция может сделать то же самое.

К примеру, первая промышленная революция привела к росту значимости Великобритании в качестве передовой экономики, потому что она индустриализировалась первой и какое-то время до тех пор, пока Германия и США не обогнали ее, пользовалась «преимуществом первопроходца». Возникает вопрос: какие страны сейчас готовы занять лидирующие позиции с точки зрения применения технологий ИИ для преобразования экономики?

Так как масштабное применение ИИ наиболее ярко выражено в производстве и потреблении (хотя эти два направления не всегда находятся в тесной взаимосвязи друг с другом), то вероятнее всего в числе передовых стран, осуществляющих производство на базе искусственного интеллекта, будут США, Китай и Индия.

Разные пути национального развития

Если распространение роботов и широкое использование технологий ИИ призваны увеличить производительность, то экономические показатели различных стран могут измениться в зависимости от того, насколько в их экономику будут интегрироваться новые технологии, станут ли роботы и системы ИИ облагаться налогом и регулироваться на государственном уровне. Сюда входят и стратегии развития стран: одни могут стремиться получать дивиденды от внедрения ИИ, а другие использовать ИИ для увеличения и расширения производства, соответственно экономический перевес будут иметь вторые.

Следующая группа при большем производстве продуктов и росте доходов увеличит также оборонные мощности, а соответственно может возникнуть риск агрессивного потенциала.

То же самое происходит при разном налогообложении и регулировании систем ИИ и роботизированных технологий.

Кроме того, различные страны вкладывают в искусственный интеллект и очень различающиеся суммы. В силу подобных особенностей развивающимся странам становится всё сложнее стать лидерами экономического развития (если раньше они выигрывали на дешевом рабочем труде, то сейчас нередко можно экономить большое количество ресурсов за счет использования ИИ).

Государствам необходимо прорабатывать программы, связанные с решением вопросов безработицы. Странам стоит не препятствовать, а поощрять развитие технологий ИИ, так же как и исследования в этой области, потому что именно опасения, что технологии могут негативно повлиять на экономику и общество, являются главным аргументом против стимулирования развития. Соответственно людям и компаниям следует делать собственный свободный выбор. Государствам не стоит вносить целенаправленные изменения в налоговую систему для остановки развития технологий (налог на роботов). К примеру, в парламенте ЕС был предложен налог на роботов на том основании, что «взимание налога с работы, выполняемой роботом, или платы за использование и обслуживание робота следует рассматривать в контексте финансирования поддержки и переподготовки безработных, чьи рабочие места были сокращены или ликвидированы» [83].

Подведем краткие итоги.

Государствам необходимо разрабатывать соответствующие стратегии для решения вопросов занятости.

Странам стоит поощрять развитие технологий и исследований в этой области. Непонимание технологий вредит развитию.

1.7. РАЗВИТИЕ ТЕХНОЛОГИЙ ИИ И КОНКУРЕНЦИЯ В МАСШТАБАХ МИРОВОЙ ЭКОНОМИКИ

Модернизацию современной экономики можно описать следующим образом: она представляет собой развитие на макроэкономическом, институциональном, социально-цивилизационном, внешнеэкономическом, структурно-техническом, ресурсном и политэкономическом уровне. С точки зрения ее реструктуризации предполагается рост численности отраслей, базирующихся на современных технологиях, что и служит основной целью цифровизации экономики [84].

Все эти изменения структуры экономики преследуют одну цель – повысить ее глобальную конкурентоспособность. Появление новых

информационных технологий, как, например, Интернет, облачных и больших данных, технологий ИИ, привнесло большую долю изменений в традиционные отрасли экономики и в понимание промышленного развития и бизнес-моделей. Как раз цифровизация стала центральной темой в преобразовании и модернизации экономики и глобальной конкуренции в целом, поскольку наблюдается всеобщее внедрение технологий ИИ в различные секторы реальной экономики [85].

В ближайшее время развитие технологий связи 5G, искусственного интеллекта, промышленного Интернета повлияет на ускорение внедрения инноваций. Крупные и средние предприятия будут работать с платформами передачи данных для интеграции цифровой экономики с реальными секторами экономики и всеобщего ускорения цифровой трансформации в различных сферах. Соответственно можно сказать, что цифровые технологии занимают важную позицию с точки зрения построения устойчивых связей в реальных секторах экономики.

Поскольку новые отрасли ИИ активно растут, развиваются и включают в себя такие компоненты, как интеллектуальное производство, автомобильная электроника, Интернет вещей и 5G, техническая оснащенность стала основой для развития информационных технологий. Именно это является центральным компонентом цифровизации, а соответственно крайне важно занимать лидирующие позиции по технической оснащенности для осуществления цифровой трансформации.

Поскольку ажиотаж вокруг искусственного интеллекта усиливается, различные поставщики услуг стремятся прорекламировать и продвинуть свои услуги и продукты, в которых применяются технологии ИИ. Однако чаще всего то, что называется ИИ, является лишь одним из компонентов ИИ [86].

Машинное обучение

Для работы ИИ необходимо специализированное аппаратное и программное обеспечение, поскольку ни один из языков программирования, даже самый распространенный, например Python или Java, не является сам по себе ИИ. Соответственно одним из вариантов может быть применение машинного обучения, когда системы ИИ работают, потребляя большие объемы обучающих данных, после чего выявляют некие закономерности и впоследствии применяют их для решения задач. Машинное обучение позволяет создавать алгоритмы на основе анализа

данных, минимизируя участие человека в процессе разработки алгоритмов и выявлении закономерностей.

Глубокое обучение

Для наиболее сложных задач машинного обучения применяется глубокое обучение, позволяющее обрабатывать сложные форматы данных, такие как естественную речь, видеозаписи. Соответственно глубокое обучение (Deep Learning, DL) подразумевает самостоятельное выстраивание многослойных правил в формате нейронной сети на основе данных, полученных в процессе обучения. В общих чертах глубокое обучение можно представить в виде огромной формулы, которая включает в себя множество составляющих и позволяет решать задачи на основании уже имеющегося опыта [87, 88].

Создание ИИ

При создании технологий ИИ важна не только математическая составляющая и знание языков программирования, но и аппаратная составляющая устройств, которая лежит в основе ИИ. Это подразумевает наличие специальных вычислительных компонентов, с помощью которых происходит реализация технологий ИИ. Они осуществляют сложные вычисления, которые нужны для определения закономерностей, принятия решений и обработки данных [89].

Для более понятного описания можно представить пирамиду, где нижний уровень – это аппаратное обеспечение, средний уровень – это фреймворки и библиотеки, а верхний – прикладное ПО. Графические процессоры GPU: первоначально они предусматривались для отображения (рендеринга) графики. Так как они хорошо справляются с параллельной обработкой данных, то они эффективны для обучения моделей. Блоки тензорной обработки TPU: они созданы специально для убыстрения процесса вычислений ИИ, а наибольшую эффективность демонстрируют в задачах глубокого обучения. Блоки нейронной обработки NPU способны решать задачи, которые затрагивают нейронные сети и являются некой имитацией нейронных связей, которые есть в человеческом мозгу. Они всё анализируют и обрабатывают большие объемы данных, давая возможность обучаться и выстраивать прогнозы.

Еще более мощным инструментом являются тензорные вычислительные блоки (TPU), разработанные в инновационном центре Google. Главной задачей TPU стало повышение эффективности нагрузок ИИ, в большинстве связанных с нейронными сетями. Главное преимущество –

их энергоэффективность из-за потребления меньшего количества энергии в сравнении со стандартными CPU (центральный процессор) и GPU.

Нейронный процессор (NPU) – отдельный класс сопроцессоров и микропроцессоров, который используется для того, чтобы ускорить работу алгоритмов искусственных нейронных сетей.

Интегральные схемы, ориентированные на конкретные приложения (ASIC), – это микросхемы, которые разрабатываются для определенных задач при вычислительных процессах ИИ.

Полевые программируемые вентильные массивы (ППВМ) (Field Programmable Gate Arrays) – более конструктивно сложная и развитая программируемая логическая интегральная схема (ПЛИС), включающая блоки как для стандартных операций, например, обработка аудио- и видеоданных, так и для создания цифровых схем.

Нейроморфные чипы, способные быстро реагировать на изменяющиеся события в реальном времени, которые как раз требуют быстрой реакции.

ИЗВЕСТНЫЕ МИРОВЫЕ ПОСТАВЩИКИ АППАРАТНЫХ ТЕХНОЛОГИЙ

В связи с упомянутым выше существует некий сформировавшийся перечень мировых поставщиков аппаратных средств ИИ [90–98].

1. NVIDIA

Компания основана в 1993 году в США и является ведущей в разработке графических процессоров и систем на чипе. Производство чипов является основным источником дохода компании, чистая прибыль которой в 2023 году составила $29,8 млрд (к примеру, еще в 2018 году чистая прибыль компании составляла 3,047 млрд долларов). Капитализация самой компании достигает $1,7 млрд, поэтому ее можно назвать одной из самых дорогих компаний Америки. При этом графические процессоры компании не только предназначены для работы с графикой, но и осуществляют процесс вычислений.

2. Advanced Micro Devices (AMD)

Американский производитель интегральных микросхем, центральных процессоров (CPU), графических процессоров (GPU), адаптеров и чипсетов, основанный в 1969 году, занимающий передовую позицию

на мировом рынке. Начиная с 2009 года размещает микроэлектронное производство на мощностях других компаний за неимением своих мощностей, изначально арендуя их у GlobalFoundries Inc. (GFS), а с 2018 года – у TSMC. Является одним из основных конкурентов NVIDIA: в 2023 году получил чистую прибыль 854 млн долларов.

3. Graphcore Limited

Компания основана в 2016 году, занимается разработкой ускорителей для ИИ и машинного обучения. Заходя на рынок, анонсировала передовую цепочку графических инструментов для ИИ, называющуюся Poplar Software Stack, а в 2017 году представила свой первый чип. Базируется в Великобритании. Позиция компании довольно шаткая, поскольку ей оказывается трудно соревноваться с такими гигантами отрасли, как NVIDIA, поэтому она вынуждена сворачивать свои масштабы и закрывать, к примеру, производство в Китае.

4. Cerebras

Американская компания, основанная в 2015 году, работает с ИИ. Имеет офисы в Кремниевой долине, Сан-Диего, Торонто и Токио. Получила широкую известность благодаря разработанной системе Wafer Scale Engine (WSE), устанавливаемой на процессоре, целью которой является предоставление упрощенного программируемого ресурса ИИ кластерного масштаба без сложной настройки, что помогает сэкономить трудозатраты специалистов. Создает компьютерные системы для сложных приложений ИИ, работающих на основе глубокого обучения.

5. Semiconductor Manufacturing International Corporation (SMIC)

Одна из крупнейших компаний Китая, занимающаяся производством микроэлектроники. Основана в 2000 году со штаб-квартирой в Шанхае, имеет передовое производство чипов, а также реализует проекты Huawei, Qualcomm, Broadcom, Texas Instruments. Ведет активное расширение мощностей: несмотря на санкции США, стала производить микрочипы по 5-нм технологическому процессу, показала высокие технологические возможности.

6. Google Cloud TPU (Google)

Специально созданный ускоряющий чип, который был представлен компанией в 2016 году. Используется в машинном обучении и поддерживает такие продукты Google, как «Переводчик», «Фотографии»,

«Поиск», «Ассистент». Существует также технология Edge TPU, которая появилась в ответ на растущий спрос на облачные модели ИИ и предназначена для периферийных устройств, как, например, смартфоны, планшеты, устройства IoT.

7. Hua Hong Semiconductor

Китайский производитель полупроводников, расположенный в Шанхае и основанный в 1996 году. Имеет три фабрики по производству двухсотмиллиметровых пластин, а также еще две по производству трехсотмиллиметровых пластин через свою дочернюю компанию (HLMC). Ведет активное расширение мощностей; основная цель – развитие собственной отрасли интегральных схем на фоне санкций.

РАСПРЕДЕЛЕНИЕ ДОХОДОВ И РЕГУЛИРОВАНИЕ ЭКОНОМИКИ

Ранее уже затрагивались особенности распространения роботов и то, как ИИ может влиять на макроэкономику. Поэтому одним из важнейших факторов является распределение доходов и регулирование экономики государством. Опираясь на опыт экономической революции прошлого, можно предположить, что в ходе глобальной цифровизации экономики произойдет значительный всплеск производительности в связи с заменой человека роботами и упрощением его рутинных задач. Соответственно компенсирующие меры государств должны оказывать балансирующее воздействие на общемировую экономику, чтобы избежать узаконенного растущего неравенства на фоне разного распределения экономических мощностей.

Таким образом, экономия рабочей силы происходит благодаря компьютерам и различным инструментам, которые связаны с ними, поэтому производители, контролирующие рынки, для создания оборудования для ИИ могут оказывать все большее влияние на других экономических игроков. Происходит монополизация рынков, а это значит, что при одинаковых условиях увеличивается прибыль ограниченного числа контролирующих сил и, как следствие, возрастает финансовое неравенство, а в дополнение к этому еще и усложняется решение глобальных вопросов мировой политики. Соответственно при отсутствии должного вмешательства государства преимущество от цифровой революции получают лишь более крупные компании.

На основе рассмотренных выше компаний, которые активно развиваются в сфере производства аппаратных технологий для ИИ, видна

тенденция, что бóльшая часть из них сосредоточена в ограниченном количестве стран. Чаще всего это компании, размещенные на территории США, либо американские компании, распределяющие свои производственные мощности на территориях других стран, таких, как, например, страны Азии. В отчете TechInsights сказано, что в Топ-25 крупнейших поставщиков полупроводниковой продукции по итогам 2023 года вошли 13 компаний со штаб-квартирой в США, еще несколько предприятий базируется в Европе, Японии, на Тайване, в Южной Корее и Китае [99–101].

Что касается полупроводниковых материалов, то, к примеру, в 2022 году поставки кремниевых пластин, необходимых для производства чипов, на мировом рынке поднялись до показателей 14,713 млрд квадратных дюймов. В том же году увеличился спрос на кремниевые пластины для автомобильного и промышленного сектора, Интернет вещей, системы связи 5G. Как раз кремниевые пластины служат в качестве основного компонента для большей части составляющих полупроводников, перечень использования которых очень широк: персональные компьютеры, мобильные гаджеты, серверы, бытовая электроника, индустриальное оборудование, средства связи и пр. Начиная с 2022 года резко выросла также цена на сырье для производства чипов [102]. Соответственно любые внешнеполитические отношения между государствами моментально начинают оказывать влияние на торговые и экономические отношения.

К примеру, в Российской Федерации в 2023 году Министерство промышленности и торговли выделило более 2,5 млрд рублей на разработку технологий по выпуску кабелей и чипов внутри страны, поскольку эта отрасль оказалась в дефиците из-за сформировавшейся геополитической ситуации и санкций со стороны США и Европы. В связи с обострением конкуренции на глобальном рынке государство предпринимает меры для обеспечения бесперебойного доступа к микроэлектронике и налаживания производственных мощностей внутри страны [103].

В 2023 году США уменьшили количество экспортируемых микрочипов ориентировочно на 20 % в сравнении с 2022 годом (этот показатель является минимальным с момента мирового кризиса в 2009). По информации «РИА Новости», в течение 2022 года поставки микрочипов на мировые рынки из США сократились примерно на 15 %,

а в 2023 году – уже на 19,2 %. Здесь ключевую роль сыграла затоваренность рынка, неблагоприятная макроэкономическая ситуация и напряженные отношения с Китаем. В связи с действующими ограничениями Китай уменьшил количество закупаемых микрочипов на 51 % (в 2021 году доля экспорта в Китай составляла 33 %, тогда как в 2023 году опустилась до 13 %).

Ряд других крупных покупателей также сократил импорт микрочипов из США. В 2023 году Тайвань сократил закупки на 1 %, Малайзия на 19 %, Израиль на 32 %, Япония на 46 %. Экспорт упал также в таких странах, как ОАЭ, ЮАР, Швейцария и Турция. Закупки смещаются из Китая в Таиланд, Вьетнам, Индию и Камбоджу. В 2022 году в США был выпущен указ о том, что запрещено использовать чипы, произведенные вне страны. Кроме того, крупным производителям полупроводников, финансируемым государством, было запрещено строить новые заводы на территории КНР минимум на 10 лет (только расширять уже работающие) [104].

В 2024 году Китай внедрил новые правила по отказу от микропроцессоров Intel и AMD, которые со временем будут исключены из применения в государственных ПК и серверах, поскольку идет активное замещение иностранных технологий на собственные. Госзакупки направлены также на то, чтобы отойти от операционных систем Windows, Microsoft и программного обеспечения, затрагивающих базы данных иностранного производства с приоритизацией внутренних возможностей страны. В 2023 году Китай ввел ограничения на экспорт металлов для производства процессоров. Все эти меры предпринимаются для снижения зависимости страны от иностранных изделий [105].

В 2024 году правительство Индии инвестировало $15,2 млрд в заводы по производству полупроводников, подтвердив предложение Tata Group о строительстве первого в стране крупного предприятия по производству микросхем. Индия является перспективной страной с точки зрения развертывания производств электроники, так как она – одна из самых быстроразвивающихся экономик мира. Однако это осложняется отсутствием в стране полупроводниковых производств. В 2022 году анонсировалось вложение $30 млрд в преобразование технологического сектора государства и формирование цепочек поставок полупроводников [106].

Дальнейшее экономическое развитие

Вполне возможно, что только некоторое количество стран станет производить продукты и услуги, полностью базирующиеся на ИИ. Тем не менее это не означает, что такими продуктами или услугами не будут пользоваться другие государства. Этот процесс сравним с компьютеризацией, так как далеко не все страны занимаются производством компьютеров; в разработке ПО доминируют опять же США, однако компьютеры используют во всем мире. Мало того, если какая-то из стран вдруг откажется применять компьютеры, поскольку она их не производит, то она автоматически окажется в последних рядах с точки зрения позиционирования в мировой экономике.

Разработка технологий ИИ, скорее всего, пойдет по сходному пути, это будет означать, что если какая-то страна не производит системы искусственного интеллекта или алгоритмы, не занимается развитием глубокого обучения либо физических объектов (роботов), то она может их только применять. Наоборот, если этого не делать, то вероятнее всего значимость национальной экономики страны на мировом рынке существенно снизится. В силу этого в проигрыше окажутся те страны, которые не смогли в силу различных обстоятельств продвинуться в технологическом развитии.

Подведем краткие итоги.

Роботизация и внедрение ИИ влияют на общемировую экономику, а также ведут к неравенству баланса сил (появляются более крупные игроки, которые могут в большей степени влиять на экономику других стран).

Экономическое стимулирование ИИ и производство аппаратной составляющей и оборудования – определяющие факторы цифровой революции.

Внутригосударственное развитие технологий – важный аспект регулирования технологий ИИ в связи с монополизацией рынка.

1.8. ИНДЕКС ГОТОВНОСТИ ИИ

Цифровизация в качестве катализатора развития технологий

Амбициозные цели цифровизации на государственном и политическом уровне еще больше способствуют массовому использованию технологий ИИ. Компании и организации из разных стран, а также отраслей начинают все чаще полагаться на решения ИИ, вкладывают ресурсы

в его развитие, чтобы повысить эффективность работы компании в целом, расширить возможности работы сотрудников, улучшить обслуживание клиентов и т. д. Как ни странно, как только дело доходит до использования потенциала ИИ в полной мере и вопрос затрагивает собственные инвестиции компании, необходимые для проведения исследований и совершенствования процессов, большинство компаний опасаются вкладываться в развитие в ИИ.

Что такое индекс зрелости ИИ

Индекс зрелости является средством, с помощью которого можно регулировать процессы разработки и внедрения технологий искусственного интеллекта в различных отраслях экономики, в социальной сфере и на уровне государственного управления. Индекс зрелости можно определить как некую оценку, которая показывает, насколько внедрены и освоены возможности, связанные с применением искусственного интеллекта в той или иной отрасли. Такие показатели, в свою очередь, помогают определить приоритетные направления дальнейшего развития отрасли для достижения высокой производительности компании или организации с помощью применения ИИ [107].

Специализированные исследования

В 2021 году компания Accenture проводила масштабное исследование, в котором приняли участие около 2000 компаний из 15 стран (крупнейших с точки зрения рыночной капитализации), в числе которых США, страны Европы, а также Индия. Компании оценивались с точки зрения того, как они пользуются инструментами ИИ, и выяснилось, что лишь 12 % из них входят в число лидеров и являются так называемыми AI Achievers, т. е. применяют у себя эти инструменты с максимальной эффективностью. Что касается энергетической отрасли, там процент достигает и вовсе показателей в 10 % [108].

Оценка индекса зрелости

Кстати, Accenture разработала систему оценки индекса зрелости, использовав технологии ИИ, – в ходе опроса были применены модели машинного обучения для обработки массива данных. Специалисты пришли к выводу, что индекс зрелости ИИ определяется тем, насколько большой объем данных может обрабатываться технологиями ИИ и стратегией развития компании в целом. Здесь также играют роль такие возможности, как, например, использование облачных

платформ и инструментов, платформ обработки данных и управления процессами.

Какова формула успеха?

Итак, самыми успешными являются те организации, которые применяют интегрированный подход для масштабной поддержки технологий ИИ. Руководители таких организаций воспринимают ИИ в качестве стратегического приоритета для своего успешного развития. Компании осуществляют постоянные вложения в развитие ИИ, чтобы впоследствии получить еще больше выгоды от таких инвестиций. Такие компании на 25 % чаще внедряют пилотные проекты. Проводится обучение сотрудников, которым не хватает базовых навыков для масштабирования ИИ. Разработчики активно совершенствуют обработку данных.

Коллаборация ИИ и человека

Организаторы, внедряющие ИИ, также получают финансовое стимулирование. При наличии четкой стратегии, ориентированной на использование технологий ИИ, а также внешнего финансирования следующей ступенью становится обязательное обучение и переквалификация сотрудников. Таким образом сотрудники проще воспринимают ИИ, что способствует хорошей совместной работе ИИ и человека. По статистике, в 44 % таких организаций есть сотрудники с высоким уровнем квалификации.

Успешные компании также прорабатывают стратегии привлечения талантливых работников, чтобы не отставать от современных тенденций. Наряду с наймом на работу это может представлять собой партнерство или развитие специализированных компаний для выполнения определенных функций, к примеру, специалистов по обработке данных или поведенческому анализу, социологов и других специалистов, которых можно привлечь к совместной работе. К примеру, в 2018 году компания Exelon основала Академию аналитики, что позволило обучить сотрудников для выполнения более квалифицированной работы.

Примеры из практики

Ведущая нефтегазовая компания в Юго-Восточной Азии создала геймифицированную платформу, на которой обучаются сотрудники, чтобы расширить свои знания в области искусственного интеллекта.

Компания создала также облачный сервис, который анализирует производительность, оценивает различные данные о сотрудниках в промежутке последних десяти лет, чтобы программа позволяла рекомендовать работников, которые больше всего подходят под различные цифровые роли. Такое нововведение значительно упростило работу кадров, сократило предвзятость руководства к сотрудникам при принятии решений о продвижении их по карьерной лестнице.

О развитии метавселенных

Всемирный экономический форум и компания Accenture уже в 2024 году подготовили отчет о развитии промышленных метавселенных (пространствах, которые включают в себя физическую и цифровую реальность). Отчет был сформирован в ходе наблюдений более 150 экспертов – представителей различных слоев общества: научной сферы, международных организаций, предпринимательства, а также государственных органов и технологического сектора. Авторами отчета было исследовано более 600 примеров использования технологий, которые относятся к концепции метавселенной и были внедрены крупными компаниями в 10 отраслях экономики [109].

При проведении исследования авторы ставили цель описать то, что происходит с метавселенными на текущий момент, а также рассмотреть направления их развития. После анализа сфер промышленного производства, автомобилестроения, городов и их инфраструктуры, энергетики и здравоохранения эксперты пришли к следующим заключениям. В 2030 году рынок метавселенных достигнет $900 млрд, а мировой доход от индустриальных метавселенных составит $100 млрд. Основу технологий промышленных метавселенных составляют цифровые двойники.

Немаловажными будут пространственные вычисления (Spatial Computing, так называемая расширенная реальность), искусственный интеллект, в том числе генеративный ИИ, технологии Web3 и блокчейна. Помимо этого, в отчете упоминаются робототехника, IoT, облачные и периферийные вычисления, а также технологии связи 5G и 6G. Исходя из аналитики отчета, авторы сформулировали мнение, что самыми продуктивными станут компании, которые будут планомерно использовать все технологии, которые входят в метавселенные (к примеру, не только генеративный ИИ, но и все в совокупности). Несмотря на уменьшающийся интерес к вложениям в стартапы метавселенных,

крупные компании, как, например, Netflix, Apple и Microsoft, продолжают активно инвестировать в решения пространственного взаимодействия (Spatial Experiences).

Азия, Латинская Америка, Ближний Восток и Африка наиболее активны в развитии промышленных и корпоративных метавселенных (Industrial and Enterprise Metaverse). Как раз на них приходится наибольшая доля проектов – 37 %, тогда как на США приходится 36 %, а Европу – 27 %. Что касается количества патентов, здесь лидируют США, тем не менее Китай обладает самым большим количеством патентов по корпоративному моделированию (Enterprise Simulation) и цифровым двойникам. Если говорить об успешном развитии промышленных метавселенных, необходимо решить вопросы цифровой безопасности, а также правового регулирования негативных воздействий на окружающую среду.

Периферийные вычисления

Еще одной аналитической работой центра Accenture стало рассмотрение подходов к внедрению периферийных вычислений, так как в связи с использованием технологий искусственного интеллекта данные превращаются в основной источник цифровой трансформации. Был проведен опрос среди 2100 руководителей высшего звена в 18 отраслях промышленности, затронувший 16 стран. Выяснилось, что 65 % компаний так или иначе используют в работе периферийные вычисления. Специалистами было выделено несколько подходов к применению технологий ИИ, а именно: специализированный, тактический, интегрированный, суперинтегрированный.

Периферийные вычисления можно определить как перенос вычислений на периферию сети, т. е. туда, где они максимально близко находятся к пользователям и устройствам. Их целью является использование преимуществ растущего числа интеллектуальных устройств, работающих в сети, которые поддерживают более сложную обработку данных в точке сбора. Система периферийных вычислений позволяет анализировать данные в режиме реального времени или наиболее близком к реальному, что дает возможность быстро получать бизнес-аналитические данные. Многие компании, несмотря на высокие затраты на эти технологии, используют специализированный подход в управлении устройствами и активно применяют Интернет вещей [110].

Специализированный подход

Менее 30 % опрошенных утверждают, что облачные платыормы используются ими для управления бизнес-операциями. Лишь 11 % применяют их для инновационной деятельности. Эта целевая группа менее результативна с точки зрения применения технологий периферийных выполнений, и ей необходимо будет еще многое сделать, чтобы полномасштабно применить периферийные вычисления. 28 % организаций, относящихся к специализированному подходу, планируют полностью интегрировать периферийные вычисления с облаком, притом около половины из этих организаций в ближайшие три года.

Тактический подход

В этом случае организации используют периферийные вычисления для того, чтобы выполнять определенные задачи компании в более сжатые сроки. Такие компании нередко отстают от внедрения технологий для корпоративных функций, поскольку внедряют их только в определенных областях, а значит, ограничивают их тиражирование. Из-за проблем с масштабированием такие компании добиваются меньших результатов. Лишь 28 % тактических компаний полностью или частично интегрировали свои стратегии в облачные. Согласно опросу, 20 % компаний придерживаются тактического подхода. В течение ближайших трех лет эти компании планируют интегрировать периферийные вычисления в свои процессы при помощи модернизации собственных возможностей.

Интегрированный подход

Такой подход означает интеграцию периферийных вычислений с цифровым ядром, облаком, данными ИИ, приложениями и платформами. Это позволит организациям масштабироваться в ускоренном режиме. К примеру, компания Inventec – тайваньский производитель электроники – комбинирует облачные технологии, ИИ, а также цифровые двойники. Компания внедрила концепцию «умного производства» на своих шести предприятиях, сочетая технологии компьютерного зрения и цифровых двойников. Компания внедрила также Интернет вещей для автоматизации и взаимодействия на линии сборки. За счет этой технологии компания сократила количество ошибок в процессе производства и оптимизировала время, уходящее на проверки. Среди компаний, использующих этот подход, почти 50 % уже применяют цифровые

двойники, а 79 % планируют полностью интегрировать периферийные вычисления в облако.

Суперинтегрированный подход

Небольшая группа компаний составляет лишь 6 % от общего числа использующих технологии периферийных вычислений. Для них периферийные вычисления – это уже не отдельная технология, а более полное внедрение ИИ в продукт или технологию на основе цифрового ядра и общей бизнес-стратегии. Здесь создается бизнес-лидер по внедрению передовых технологий. Расширяется сотрудничество ИТ и бизнеса, активно инвестируются средства в ИИ и машинное обучение, привлекаются партнеры для поддержки периферийных вычислений.

Организации, которые выбрали такой подход, становятся более успешными по сравнению с предыдущими рассмотренными подходами. Согласно исследованиям, у них в 4 раза больше шансов существенно ускорить инновационное развитие, в 9 раз больше шансов повысить эффективность и практически в 7 раз сократить издержки. Такой подход можно рассмотреть на примере компании Tesla, которая создала свой личный набор микросхем для производства суперкомпьютера, обучающего системы ИИ для автомобилей.

ГОСУДАРСТВЕННОЕ СТИМУЛИРОВАНИЕ

Вложения в ИИ со стороны государства продолжают расти, поэтому далее представлено государственное участие в проектах в сфере ИИ более детально.

Эксперты ОЭСР выделяют следующие варианты.

- **Инвестор.** Государство финансирует разработку и содействует внедрению новых технологий.
- **Заказчик.** Государство осуществляет закупки цифровых продуктов или участвует в разработке новых программных решений через механизм государственно-частного партнерства (ГЧП).
- **Регулятор.** Государство, следуя научно-техническому прогрессу, своевременно обновляет соответствующую нормативно-правовую базу.
- **Стандартизатор.** Государство организует разработку национальных стандартов с привлечением всех заинтересованных сторон, проводит оценку их соответствия современному уровню технического развития.

- **Владелец данных.** Государственные органы хранят и обрабатывают огромные массивы данных, обеспечивают их безопасность и целостность.
- **Поставщик услуг.** Государственные цифровые платформы взаимодействуют с гражданами, активно используя технологии ИИ.

Эксперты Всемирного банка (ВБ) отмечают несколько перспективных сфер применения технологий ИИ в системе государственного управления:

- обработка обращений граждан;
- контроль соблюдения законодательства и оценка рисков;
- финансовый контроль бюджетных расходов;
- оптимизация внутрикорпоративных операционных процессов;
- персонализированное предоставление услуг на основе анализа цифрового профиля гражданина;
- эффективное распределение ресурсов и помощь в принятии решений.

Рассмотрим понятие индекса готовности к внедрению ИИ в России.

В 2023 году РАНХиГС, взявшая на себя обязательства по сбору и замерам индекса зрелости ИИ в 55 федеральных органах исполнительной власти, провела аналогичные исследования. Сделано это было в соответствии с дорожной картой «Развитие высокотехнологичного направления “Искусственный интеллект” на период до 2030 г.», утвержденной правительством в конце 2022 года. Ведомствам присваивались четыре уровня зрелости: начальный, базовый, прогрессивный и лидерский. Сама по себе методика вычисления индекса зрелости была разработана в рамках федерального проекта «Искусственный интеллект» нацпрограммы «Цифровая экономика РФ» [111].

Цель вычисления этого индекса на уровне правительства заключается в том, что он должен отразить, насколько «автономно» оказание услуг населению, а помимо этого степень интеграции, обмена данными и их обработки на уровне взаимодействия между различными ведомствами, а также препятствия к широкому использованию технологий ИИ. Значение индекса будет оказывать влияние на выделяемые бюджеты согласно планам цифровизации федеральных органов исполнительной власти. Это позволит более конструктивно оценивать необходимость инвестирования в технологии ИИ, выделения грантов, а также

определять влияние каждого отдельного органа на конкретные секторы экономики, за которые он отвечает.

Что оценивается?

Данные – оценивается, насколько быстро можно получить доступ к тому или иному государственному органу, если делается соответствующий запрос.

Управление – происходит оценка того, как подходит конкретный орган к применению технологий ИИ.

Культура и кадры – оценивается потенциал персонала, который работает с данными и ИИ, а также меры по развитию таких навыков у сотрудников.

Процессы – происходит оценка численной составляющей подразделений органов исполнительной власти, которые выступают в качестве заказчиков внедрения технологий ИИ.

Продукты – оценивается совокупность решений, подкрепленных технологиями ИИ, которые были внедрены.

Модели и инструменты – происходит оценка того, насколько быстро реализуется ИИ-проект в процессе от изначальной постановки задачи до непосредственной реализации.

Инфраструктура – оценивается степень технической оснащенности, а именно то, насколько доступны вычислительные мощности, а также сами инструменты разработки, чтобы реализовать ИИ-проекты.

Как изменились показатели?

Поскольку исследование индекса готовности проводится на ежегодной основе, есть возможность сравнить показатели с 2021 годом. На сегодняшний день использование ИИ в приоритетных отраслях экономики Российской Федерации и секторах социальной сферы составляет 32 %, что в сравнении с 2021 годом улучшает показатель использования ИИ в стране на 16 %.

Лидирующие сферы деятельности по направлению «Использование ИИ» в 2023 году:

- финансовые услуги;
- сектор ИКТ;
- здравоохранение;
- торговля;
- СМИ.

Факторы, замедляющие процесс

Однако существует ряд факторов, препятствующих развитию и применению ИИ в организациях: недостаток персонала, недостаточное количество данных и осведомленность исполнителей и руководителей о возможностях применения ИИ.

Что касается финансовых трудностей и кадрового дефицита, то они снизились на 25 % в сравнении с 2021 годом. Тем не менее на сегодняшний день только 34 % организаций, участвовавших в опросе, имеют необходимый кадровый ресурс в области ИИ, а если говорить об отраслях, которые только начинают применять ИИ, то этот показатель составляет и вовсе 25 %, что свидетельствует о необходимости увеличения количества таких специалистов на 20–30 %.

Показатели по использованию ИИ на региональном и муниципальном уровне крайне низкие и объясняются отсутствием кадров, а также инфраструктуры и в целом базы данных для применения ИИ [107, 112].

Лидерами использования ИИ являются Ханты-Мансийский АО и Московская область, где уровень применения технологий ИИ превысил планку в 50 %, однако при этом в 29 субъектах уровень использования ИИ менее 10 %. Средний показатель применения технологий ИИ на муниципальном уровне управления 6,3 %, хотя разница с лидирующими городами при этом довольно значительная. Например, Казань имеет показатель 37,5 %. Общий показатель применимости технологий ИИ в системе государственного управления на региональном уровне 13,4 %.

В большинстве сфер деятельности различные организации и компании, которые до сих пор не используют технологии ИИ, уже запланировали внедрить ИИ-решения в свои процессы. Бóльшая часть организаций, которые на текущий момент применяют ИИ, – это крупные и средние. По статистике каждая вторая крупная организация и каждая третья средняя используют ИИ. Однако в малых организациях этот показатель составляет уже 14 %. Три основных барьера использования ИИ в 2023 году, как и в 2021 году, остались прежними. При этом влияние финансовых ограничений, дефицита специалистов, работающих с ИИ, а также недостаточная осведомленность о возможностях применения ИИ и отсутствие цифровой инфраструктуры снизились.

Подведем краткие итоги.

Индекс зрелости – оценка, показывающая уровень внедрения и освоения возможностей, связанных с применением ИИ в отрасли. Помогает определить ее дальнейшее развитие.

Специальные исследования – позволили создать классификацию организаций и выявить лидеров по отраслям, а также более слабые звенья на основе оценки экономических показателей.

Конкретные формулировки – перечень качеств, которыми должны обладать различные организации для достижения максимальной продуктивности. Приведены конкретные примеры внедрения ИИ различными компаниями, тенденции развития ИИ, а также использование государством показателей индекса готовности ИИ.

2. АНАЛИЗ ВОЗМОЖНОСТИ ПРИМЕНЕНИЯ МЕТОДОВ ОБЪЯСНИМОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В ЭЛЕКТРОЭНЕРГЕТИКЕ

2.1. ПОНЯТИЕ ОБЪЯСНИМОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

2.1.1. ПРИНЦИПЫ ОБЪЯСНИМОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

В обзоре [113], написанном в 1986 году, впервые были рассмотрены положения, касающиеся объяснения и прозрачности результатов, получаемых с помощью компьютерных технологий. В обзоре введены научно обоснованные принципы проектирования систем, основанные на *феноменологии*, которые заключались в том, что система должна отражать то, как структурировано ментальное представление пользователя об области ее применения.

При разработке информационных систем с помощью детерминированных методов разработчики могли оценить правильность их работы на основании проверки структур данных и логики. Методы машинного обучения (особенно методы глубокого машинного обучения) относятся к недетерминированным методам, результаты которых невозможно однозначно объяснить с использованием понятных пользователю терминов и подходов. Другими словами, эти методы представляют собой так называемый *черный ящик*, систему, которая описывается только через входы и выходы без представления внутренней структуры модели и логики ее работы. Для преодоления данной проблемы в настоящее время разрабатываются методы объяснимого искусственного интеллекта (XAI – eXplainable Artificial Intelligence). Под объяснимым искусственным интеллектом может пониматься набор процессов и методов,

которые позволяют пользователям понимать результаты, созданные алгоритмами машинного обучения, и доверять им [114].

В отчете института NIST [115] представлены основные принципы систем объяснимого искусственного интеллекта:

- объяснимость – система должна представлять пользователю доказательства принимаемым решениям;
- значимость – система должна предоставлять пользователю понятные ему объяснения;
- точность – система должна корректно отражать связь между выходами и входами модели;
- пределы знаний – система должна работать только для области, для которой она была разработана, и предлагать результаты только при достаточной уверенности в результатах объяснения.

Принципы позволяют определять контекстуальные факторы, которые следует учитывать при объяснении. Эти принципы зависят от взаимодействия системы с пользователем, так как основная функция объяснения заключается в создании интерпретаций, соответствующих ментальным моделям [116].

Поэтому подход к объяснимости проектируется относительно ментального представления пользователей, для которых объяснение предназначено. Этот факт учитывается в черновом стандарте IEEE P7001 [117], который, как ожидается, станет стандартом Европейского союза в рамках законодательства об искусственном интеллекте, объектом которого является объяснимость с точки зрения заинтересованных сторон. Под заинтересованными сторонами понимают пользователей системы, общественность (людей, которые могут пострадать от внедрения системы) и экспертов. Относительно пользователей системы (в том числе ее разработчиков) объяснимость может быть определена:

- через документацию, которая описывает обучающие данные;
- системную информацию, которая может потребоваться для отладки системы;
- описание среды запуска системы;
- документацию сценариев «что если»;
- объяснение работы модели.

Объяснение работы модели зависит:

- от уровня детализации:
  - краткий (например, оповещение);
  - обширный (например, отчет системной информации);

- формата (визуальный, графический, голосовой);
- степени взаимодействия с человеком:
  - декларативные объяснения (предоставление объяснения без дальнейшего взаимодействия пользователя с системой (например, вывод обоснования по одобрению кредитов));
  - одностороннее взаимодействие (предоставление объяснения на основе изменяющегося запроса в систему (например, построение визуализации в зависимости от факторов));
  - двустороннее взаимодействие (предоставление различных объяснений на основании информации, полученных после уточнения запросов пользователя).

Для общественности объяснимость систем может обеспечиваться:

- документацией, описывающей воздействие системы на окружение, а также данные, связанные с общественностью и используемые в системе;
- открытой политикой управления системой.

Для экспертов, которые могут быть представлены как аудиторы, специалисты в предметной области, консультанты, объяснение может описываться:

- системами и способами управления ею;
- методами проверки работы системы;
- стандартами по обеспечению безопасности и риск-менеджменту;
- ведением журнала действий системы.

К проблемам внедрения и развития направления объяснимого искусственного интеллекта относят [114]:

- ограниченность интерпретируемости моделей из-за непредсказуемых, невоспроизводимых результатов и согласованности входов моделей с ее входами;
- невозможность согласованного объяснения результатов моделей для групп или подмножеств данных (в настоящее время возможно локальное объяснение для одного экземпляра данных или глобальное объяснение для всего набора данных);
- возможную противоречивость в объяснении различных методов интерпретируемости для различных областей применения.

Для преодоления указанных ограничений разрабатываются новые методы улучшения интерпретируемости моделей (использование информации из промежуточных слоев глубоких нейронных сетей,

агрегация метрик интерпретируемости различных моделей), проектируются состязательные модели для количественной оценки степени объяснимости.

2.1.2. РАЗВИТИЕ НАПРАВЛЕНИЯ ОБЪЯСНИМОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Интерес к объяснимому искусственному интеллекту вырос за последние годы, так как внедрение таких методов делает возможным решение ряда проблем, связанных:

- со снижением доверия пользователей к интеллектуальным системам;
- с повышением вероятности неправомерного использования систем;
- необходимостью учета требований различных заинтересованных лиц к объяснению результатов моделей машинного обучения;
- потерей контроля над моделями из-за их ограниченной способности корректировать свое поведение в проблемных (редких) случаях.

В разделе 2.1.1 рассматривался черновой проект IEEE P7001, целью которого является установление измеримых, проверяемых уровней прозрачности для автономных систем в различных режимах их работы. Прозрачность (объяснимость) является одним из общих принципов, изложенных в стандарте IEEE, который гласит: «*Основа конкретного решения автономной и интеллектуальной системы всегда должна быть обнаруживаемой*» [116]. Рабочая группа, создавшая черновой проект IEEE P7001, была создана на основании рекомендаций, содержащихся в стандарте IEEE [118].

В соответствии с документом [118] принцип прозрачности (объяснимости) должен быть согласован с принципами уважения прав человека, ответственности разработчиков систем и их подотчетности, минимизации рисков неправильного использования систем.

В экспериментальных исследованиях (в том числе и при развитии направления объяснимого искусственного интеллекта) руководствуются принципами, представленными в отчете Белмонта [119]:

- уважения к людям (признание автономии людей и их права на знание о потенциальных рисках и преимуществах любой технологии);
- невредоносности (необходимость учитывать влияния результатов моделей искусственного интеллекта на предубеждения людей);

- справедливости применения результатов моделей искусственного интеллекта.

Подходы, инициативы, проблемы в области объяснимого искусственного интеллекта, программы развития компаний отражаются в технических документах (white paper) и манифестах, а инициативные группы разрабатывают проекты стандартов.

В статье [120] представлен манифест, который призывает к консенсусу по развитию объяснимости систем на основе искусственного интеллекта. Манифест включает в себя следующие принципы:

- необходимость создания объяснений для новых типов искусственного интеллекта (например, для генеративных моделей, распределенного и совместного обучения);
- улучшение существующих методов объяснения;
- необходимость разъяснения использования концепции объяснимого искусственного интеллекта пользователям систем;
- создание структуры методов объяснимого искусственного интеллекта;
- ориентацию объяснений на пользователей (создание объяснений, понятных человеку);
- поддержку многомерности объяснимости;
- корректировку методов объяснения в зависимости от заинтересованных сторон;
- уменьшение негативного воздействия объяснимого искусственного интеллекта (например, повышение точности объяснений);
- улучшение общественного воздействия объяснимого искусственного интеллекта (например, установление оригинальности данных, созданных с помощью искусственного интеллекта, и обнаружение плагиата).

Компания Google создала технический документ [121], который предназначен для разработчиков моделей машинного обучения и специалистов по данным. Этот документ был направлен на упрощение разработки моделей, улучшение моделей, обнаружение аномалий; объяснение поведения моделей заинтересованным сторонам. В техническом документе приводится описание методов объяснимого искусственного интеллекта, случаи их использования и ограничения. Основные положения, представленные в этом документе, были применены в Vertex Explainable AI [122].

Компания Siemens разработала технический документ, который описывает внедрение объяснимого искусственного интеллекта в жизненный цикл любого решения [123]. Документ описывает возможности применения объяснимости в технических системах для обеспечения осмысленного взаимодействия между искусственным интеллектом и людьми на всех этапах жизненного цикла продукта.

Компания Atos в техническом документе [124] рассматривает комплексный продукт Atos XAI Framework, который позволяет раскрыть интерфейсы прикладного программирования (Application Programming Interface (API)). Применение различных методов объяснимого искусственного интеллекта позволяет получать объяснение результатов работы моделей машинного обучения с различным уровнем детализации и декомпозиции системы на уровне локального и глобального типа объяснения.

Компания Intel применяет методы объяснимого искусственного интеллекта для анализа предиктивного поведения моделей TensorFlow и PyTorch [125].

Компании различных направлений рассматривают особенности применения объяснимого искусственного интеллекта в зависимости от предметной области:

- компания Erocsson [126] описывает особенности применения методов XAI в системах управления средствами коммуникации;
- компании Quantum XAI [127] и SideShift.ai [128] рассматривают особенности применения XAI для аудита смарт-контрактов в системах блокчейн;
- компания Amelia [129] анализирует применение XAI для финансового сектора;
- компания Cofotge [130] описывает применение XAI для решений в области страхования;
- компания NVIDA [131] рассматривает применение XAI для управления кредитными рисками.

Компании Microsoft [132], IBM [133] в своих технических документах и манифестах раскрывают применение объяснимого искусственного интеллекта для соответствия требованиям этики.

Кроме компаний и официальных инициативных групп, положения объяснимого искусственного интеллекта внедряются отдельными инициативными группами, к которым можно отнести Deeploy [134] и xAI [135].

Чат-бот компании Deeploy позволяет различным заинтересованным сторонам реализовать гибкое одностороннее взаимодействие с моделью искусственного интеллекта через интерфейс на основе естественного языка. Пользователи могут получать персонализированное объяснение на основании только запрашиваемой информации [134].

В июле 2023 года была создала компания xAI, целью которой является разработка и внедрение принципов объяснимости в различные сферы жизнедеятельности. В частности, компанией были разработаны оптимизатор Adam, методы пакетной нормализации, нормализации слоев и обнаружения состязательных примеров. Кроме того, компанией были внедрены методы Transformer-XL, Memorizing Transformer, μTransfer и SimCLR. Положения объяснимого искусственного интеллекта легли в основу продуктов AlphaStar, AlphaCode, Inception, Minerva, GPT-3.5 и GPT-4 [135].

В Российской Федерации развитие принципов объяснимого искусственного интеллекта косвенно заложено в указе Президента Российской Федерации от 10.10.2019 № 490 «О развитии искусственного интеллекта в Российской Федерации». Так, развитие и использование технологий искусственного интеллекта должно быть реализовано за счет [136]:

- повышения эффективности процессов планирования, прогнозирования и принятия управленческих решений;
- повышения безопасности сотрудников;
- повышения лояльности и удовлетворенности потребителей.

Как было рассмотрено выше, эти вопросы напрямую зависят от соблюдения принципов объяснимости. В ГОСТ Р 59276–2020 рассматриваются вопросы обеспечения доверия пользователей к системам искусственного интеллекта [137]. В этом стандарте показано, что недостаточная объяснимость может привести к снижению качества системы искусственного интеллекта на этапах разработки и эксплуатации систем. Для повышения доверия разработчиков и пользователей предлагается использовать алгоритмы, позволяющие системе представлять пользователю объяснимые, предсказуемые решения.

Для повышения компетенций в области объяснимого искусственного интеллекта Министерством науки и высшего образования в 2021 году была доработана модель компетенций. В модель была добавлена профессиональная компетенция студента «*Способен создавать*

и применять методы объяснимого искусственного интеллекта для создания интерпретируемых интеллектуальных систем» [138].

Внедрением принципов объяснимого искусственного интеллекта занимается стартап «Сохраняем жизни». Командой была разработана новая концепция образно-логических нейронных сетей, где каждый нейрон осуществляет акты элементарных рассуждений, суммируя которые вся сеть может решать сложные задачи с предоставлением необходимых объяснений. Концепция была внедрена для задач распознавания и классификации объектов на земле и в воздухе, обработки семантической информации, обнаружения полезных ископаемых на основании данных дистанционного зондирования Земли [139].

Стартап Glagol [140] позволяет вести коммуникации с пользователями при учете множества параметров. Искусственный интеллект позволяет ускорить процессы найма в отделе кадров и подбора поставщиков в отделе закупок, а применение объяснимого искусственного интеллекта позволяет объяснить решения, принимаемые голосовыми помощниками.

Проект РНФ [141] направлен на обеспечение киберустойчивости систем очистных сооружений с использованием методов объяснимого искусственного интеллекта. Ожидается, что применение данных методов будет способствовать объяснению причин выявленных кибератак и аномалий.

2.2. КЛАССИФИКАЦИЯ МЕТОДОВ ОБЪЯСНИМОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

В статье [142] представлена классификация методов объяснимого искусственного интеллекта (рис. 2.1). Выделяют четыре типа классификации: по масштабу, типу данных, назначению интерпретации и типу алгоритма.

Классификация по типу алгоритма базируется на ограничениях на применение метода: метод может быть зависящий и не зависящий от модели. Преимуществом методов интерпретации, не зависящих от модели, перед методами, зависящими от нее, является их гибкость [143]. Разработчики могут использовать любую модель машинного обучения, если методы интерпретации могут быть применены к любой модели. Так как обычно при решении задачи сравниваются результаты нескольких моделей, с точки зрения интерпретируемости проще работать с объ-

яснениями, не зависящими от модели, поскольку один и тот же метод может быть использован для любого типа модели.

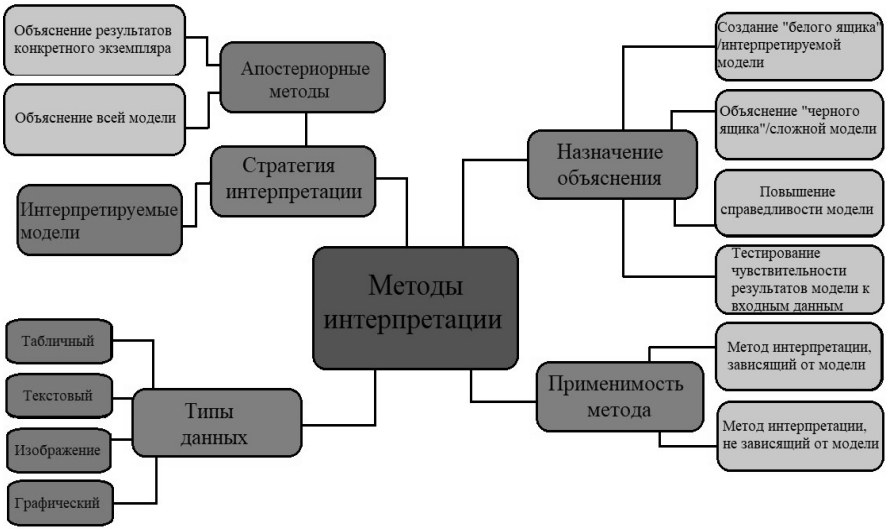


Рис. 2.1. Классификация методов объяснимого искусственного интеллекта

Классификация по типу входных данных описывает тип данных, к которому могут быть применены методы. Наиболее распространенными типами данных являются табличные, графические, текстовые, изображения [142]. В [144] представлена подобная классификация, но по типу выходных данных.

Методы можно классифицировать по цели создания и применения. На основании этой классификации можно выделить четыре категории [142]:

- методы создания моделей белого ящика (интерпретируемых моделей);
- методы объяснения сложных моделей (моделей «черного ящика);
- методы повышения справедливости модели, которые снижают вероятность проявления дискриминации данных из определенных выборок;
- методы анализа чувствительности прогнозов модели к входным данным.

Важным типом классификации является классификация по стратегии интерпретации. Методы могут быть интерпретируемые и апостериорные.

Интерпретируемые модели не требуют отдельных методов объяснения, так как их структура может быть интерпретируема человеком. К этому типу методов относятся методы [142, 143, 145]:

- линейная регрессия;
- логическая регрессия;
- дерево решений (за исключением деревьев решений большого размера);
- ансамблевые модели;
- *k*-ближайших соседей.

Апостериорные модели могут быть разделены на *локальные* и *глобальные*. Если метод дает объяснение только для конкретного экземпляра входных данных, то он называется *локальным*. Если метод объясняет всю модель, то он называется *глобальным*. Схемы данных стратегий интерпретации представлены на рис. 2.2.

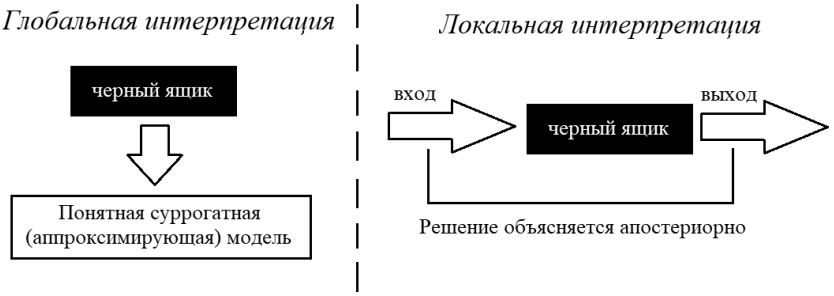


Рис. 2.2. Схемы стратегий интерпретаций:
а – глобальная; *б* – локальная

Глобальная интерпретация основана на объяснения работы модели в целом. Если существует модель $f(X)$, являющаяся решением неизвестной зависимости $y = m(x)$, то суррогатная (замещающая) модель, которая создается при глобальной интерпретации, по значениям X обучается предсказывать не y, а выход основной модели $f(X)$. Глобальные методы часто выражаются в виде ожидаемых значений на основе распределения данных. Например, график частичной зависимости, который является графиком эффекта признака, представляет собой ожидаемое предсказание

при маргинализации всех остальных признаков. Поскольку глобальные методы интерпретации описывают среднее поведение, они особенно полезны, когда разработчик модели хочет понять общие механизмы в данных или отладить модель. К ним относят [143]:

- метод взаимодействия признаков (feature interaction (H-statistic)), который количественно оценивает, в какой степени предсказание является результатом совместных эффектов признаков;
- функциональное разложение (unctional decomposition), которое разлагает сложную функцию предсказания на более мелкие части;
- оценку важности признака при перестановке (permutation feature importance), которая основана на учете потерь при перестановке признака;
- глобальные замещающие модели (global surrogate models), которые заменяют исходную модель более простой моделью для интерпретации.

Локальная интерпретация основана на объяснении результатов модели y на основании конкретного экземпляра данных, т. е. локальные методы строят объяснение, которое отвечает на вопрос «почему модель $f(X)$ для вектора X приняла значение y?» Локальные методы применяются при необходимости объяснения определенного экземпляра без анализа работы всей модели. К методам локальной интерпретации относят [143]:

- кривые условного ожидания индивидуальных наблюдений (individual conditional expectation curves (ICE)), которые описывают, как изменение признака влияет на прогноз;
- ограниченные правила (якоря) (scoped rules (anchors)), которые описывают, какие значения фиксируют прогноз (являются якорями прогноза);
- контрафактические объяснения (counterfactual explanations), определяющие, какие признаки нужно изменить, чтобы достичь желаемого прогноза;
- локально интерпретируемое не зависящее от модели объяснение (Local Interpretable Model-Agnostic Explanation (LIME)), которое заменяет сложную модель локально интерпретируемой замещающей моделью;
- значения Шепли (Shapley values), которые распределяют прогноз между отдельными признаками;

- SHAP (SHapley Additive exPlanations), представляющий метод вычисления значений Шепли с глобальной интерпретацией на основе их комбинаций по всем данным.

LIME, значения Шепли и SHAP являются методами атрибуции, т. е. в них прогноз отдельного экземпляра описывается суммой эффектов различных признаков. Другие методы являются экземплярно-ориентированными, т. е. основанными на предоставлении информации в виде релевантных или контрастных примеров. Например, для кредитного решения контрафактическое объяснение может указать, что *если бы заявитель увеличил свой доход на 10 000, то кредит был бы одобрен*. При применении ограничивающих правил для кредитного решения условие одобрения может быть сформулировано так: *если возраст клиента больше* 40 *лет и его кредитный рейтинг выше* 750, *то вероятность одобрения кредита составляет более* 90 %. В этом примере правило использует два ключевых условия.

В книге [143] приводятся также методы интерпретации для нейронных сетей, которые объясняют отдельные прогнозы и упрощают нейронные сети. Прогнозы нейронной сети получаются многочисленным прохождением входных данных через множество слоев умножения с весами и через нелинейные преобразования. Поэтому для объяснения результатов, полученных с помощью нейронных сетей, необходимо применять объяснение. К результатам могут быть применены методы, не зависящие от модели, но есть две причины, по которым вводятся специальные методы для нейронных сетей:

- нейронные сети изучают признаки и концепции в своих скрытых слоях, поэтому нужны специальные инструменты для их обнаружения;
- использование градиента для реализации методов объяснения более эффективно с вычислительной точки зрения, чем применение методов, которые не учитывают устройства модели.

К специализированным методам объяснения нейронных сетей относят [143]:

- изученные признаки (Learned Features), которые отражают, какие признаки были изучены нейронной сетью;
- атрибуцию пикселей или карты значимости (Pixel Attribution (Saliency Maps)), которые показывают, как каждый пиксель повлиял на прогноз;
- концепцию (concepts), которая показывает, какие абстрактные концепции были изучены и предложены нейронной сетью;

- состязательные примеры (Adversarial Examples), которые изучают, как можно изменить поведение нейронной сети, т. е. «обмануть» ее;
- влиятельные экземпляры (Influential Instances), которые показывают, насколько большим влиянием обладал каждый экземпляр обучающих данных для определенного прогноза.

2.3. МЕТОДЫ ЛОКАЛЬНОЙ ИНТЕРПРЕТАЦИИ

В настоящем разделе рассматриваются все методы локальной интерпретации, перечисленные выше, кроме значений Шепли и SHAP.

2.3.1. КРИВЫЕ УСЛОВНОГО ОЖИДАНИЯ ИНДИВИДУАЛЬНЫХ НАБЛЮДЕНИЙ

Метод кривых условного ожидания индивидуальных наблюдений (ICE) отображает одну линию для каждого экземпляра, показывая, как предсказание для этого экземпляра изменится при изменении признака [146].

Параметры для одной линии (одного экземпляра) можно вычислить, оставив все остальные признаки неизменными, создав варианты этого экземпляра путем замены значения признака на множество других и получения предсказаний с использованием модели для новых созданных экземпляров. Результатом является набор точек для экземпляра со значениями признака из сетки и соответствующими предсказаниями [143]: для каждого экземпляра в $\left\{\left(x_S^{(i)}, x_C^{(i)}\right)\right\}_{i=1}^{N}$ кривая $f_S^{(i)}$ строится относительно $x_S^{(i)}$, в то время пока $x_C^{(i)}$ остаются фиксированными.

Другими словами, ICE позволяет проанализировать, как выглядит среднее отношение между признаком и предсказанием. Это работает хорошо только в том случае, если взаимодействия между признаками, для которых рассчитывается график частичной зависимости, и остальными признаками слабы.

Для того чтобы облегчить анализ различий между кривыми ICE отдельных экземпляров, можно центрировать кривые в определенной точке признака и отображать только разницу между предсказаниями

относительно этой точки. Полученный график называется центрированным графиков *c*-ICE. Новые кривые определяются как

$$f_{\text{cent}}^{(i)} = f^{(i)} - \bar{1} f\left(x^a,\, x_C^{(i)}\right), \tag{2.1}$$

где $\bar{1}$ – вектор единиц с соответствующим числом измерений; f – подобранная модель; x^a – опорная точка.

Центрирование может быть полезно, если необходимо проанализировать изменение прогноза при фиксированной точке диапазона признаков, а не абсолютное изменение прогнозируемого значения.

На рис. 2.3 представлено сравнение между кривыми ICE (рис. 2.3, *а*) и центрированными кривыми *c*-ICE (рис. 2.3, *б*), построенными для объяснения результатов в задаче прогноза аренды велосипедов. Анализ кривых ICE для этого эксперимента позволяет заключить, что нет очевидной связи между признаками и прогнозным значением прокатов велосипедов. Это означает, что для данного датасета возможно использование глобальной интерпретации. При применении *c*-ICE можно проанализировать изменение прогноза по сравнению с прогнозом при соответствующем значении признака на его наблюдаемом минимуме.

Для анализа изменений в каком-либо направлении и нахождения гетерогенных отношений можно построить график производной (*d*-ICE). С графиком производной ICE легко заметить диапазоны значений признака, где предсказания модели меняются для (по крайней мере, некоторых) экземпляров. Если между анализируемым признаком x_S и x_C нет связи, то функция *d*-ICE может быть выражена как

$$f(x) = f(x_S,\, x_C) = g(x_S) + h(x_C), \quad \frac{\partial f(x)}{\partial x_S} = g'(x_S)\,. \tag{2.2}$$

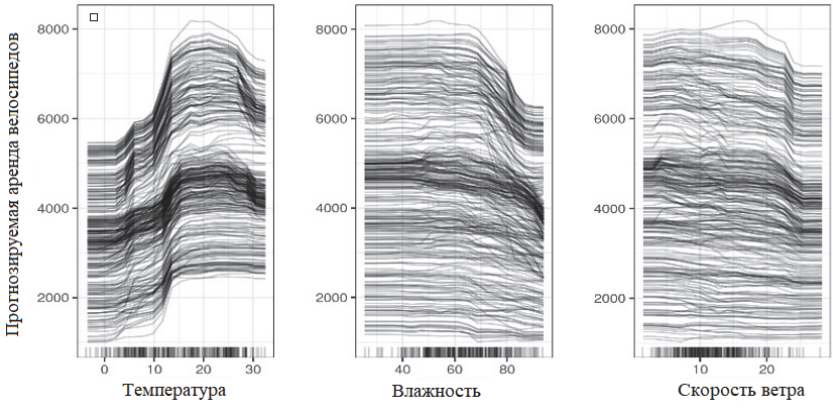
Достоинства метода:

- понятность графиков (одна линия представляет прогноз для одного случая);
- возможность детектирования неоднородных связей.

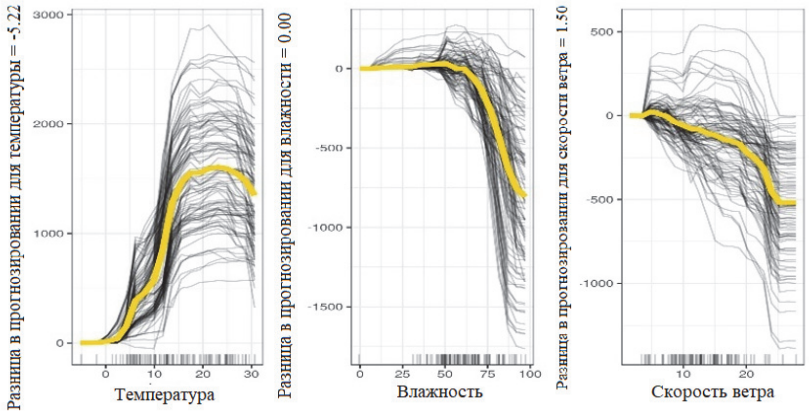
Недостатки метода ICE и его модификаций:

- невозможность анализа влияния на результат более чем одного признака;
- перегруженность графиков при множестве кривых;

- корреляция рассматриваемых признаков может привести к недопустимости некоторых точек на графике с точки зрения совместного распределения признаков.



а



б

Рис. 2.3. Графики прогнозируемых прокатов велосипедов по погодным условиям:

а – ICE; *б* – *c*-ICE

Метод ICE находит применение в различных сферах. В статье [147] ICE и SHAP применяются для мониторинга технического состояния плотин. Применение ICE в такой задаче позволяет точно определить

поведение параметров при различных трендах изменения состояния плотин без сильной локальной изменчивости по сравнению с результатами SHAP.

В статье [148] *c*-ICE применяется для визуализации зависимости вероятности появления рака от возраста для каждого экземпляра в выборке данных.

В статье [149] ICE применяется для визуализации зависимости наличия древесной биомассы от содержания органического углерода в почве.

Метод ICE может быть реализован в программных пакетах R iml, condvis, ICEboc49, pdp [150]. В Python графики частичной зависимости встроены в scikit-learn, начиная с версии 0.24.0 [151].

2.3.2. ЛОКАЛЬНО ИНТЕРПРЕТИРУЕМОЕ, НЕ ЗАВИСЯЩЕЕ ОТ МОДЕЛИ ОБЪЯСНЕНИЕ (LIME)

Локальные интерпретируемые не зависящие от модели объяснения (LIME) были представлены M. T. Ribeiro, S. Singh, C. Guestrin в [152]. Суррогатные модели обучаются для аппроксимации прогнозов базовой модели «черного ящика». Вместо обучения глобальной суррогатной модели LIME фокусируется на обучении локальных суррогатных моделей для объяснения отдельных прогнозов. LIME проверяет, что происходит с прогнозами при подаче разных экземпляров из датасета в модель «черного ящика». LIME создает новый набор данных, состоящий из экземпляров и соответствующих прогнозов модели «черного ящика». Затем на этом наборе данных LIME обучает интерпретируемую модель, которая в дальнейшем взвешивает близость имеющихся экземпляров к подаваемому для объяснения результата. LIME может быть выражен как [143]

$$\text{explanation}(x) = \arg\min_{g \in G} L(f,\ g,\ \pi_x) + \Omega(g). \qquad (2.3)$$

Модель объяснения экземпляра x – это модель g (например, модель регрессии), которая минимизирует потери L (например, среднее квадратичное отклонение) и измеряет, насколько объяснение близко к прогнозу исходной модели f. При этом сложность модели объяснения $\Omega(g)$ поддерживается на низком уровне искусственным ограничением количества учитываемых признаков. G определяется как семейство возможных объяснений (например, все возможные модели регрессии). Мера

близости π_x определяет, насколько велика окрестность вокруг экземпляра x, которая рассматривается для объяснения.

Преимущества LIME:

- возможность обработки не только табличных, но и текстовых данных и изображений;
- относительная легкость понимания за счет использования для объяснения интерпретируемых моделей;
- возможность анализа точности (надежности) интерпретации в окрестности экземпляра;
- возможность использования для объяснения признаков, которые не были учтены при обучении (например, модель «черного ящика» может быть обучена на компонентах анализа главных компонентов ответов на опрос, а LIME может быть обучен на исходных ответах и вопросах), что позволяет повысить понятность объяснения для пользователя.

Недостатки LIME:

- необходимость настройки модели;
- необходимость проверки правильности объяснения из-за возможного неправильного выбора соседей для генерации объяснения;
- необходимость определения сложности модели LIME заранее;
- нестабильность объяснений;
- возможность манипуляции объяснениями.

Таким образом, метод LIME на данный момент необходимо применять с большой осторожностью из-за серьезных вопросов к его безопасности [153].

Метод LIME достаточно распространен, несмотря на рассмотренные недостатки. В статье [154] LIME используется для объяснения влияния различных параметров оборудования на проект электротехнического комплекса дома. В настоящем исследовании используется разряженная линейная регрессия. В качестве признаков рассматриваются емкости аккумуляторных батарей, параметры возобновляемых источников энергии и тепловые нагрузки.

В статье [155] LIME применяется для анализа энергетической эффективности зданий. По результатам исследования LIME определил температуры нагрева и охлаждения как наиболее критические характеристики, влияющие на энергетическую эффективность зданий.

В статье [156] LIME используется для объяснения прогноза нагрузки электроэнергетической системы и влияния даты, ретроспективы

нагрузки и метеорологических параметров на величину электропотребления. Авторы отмечают высокую нестабильность работы LIME, которая, несмотря на объяснение работы алгоритмов, не позволяет повысить доверие пользователей.

LIME может быть реализован с использованием библиотеки Lime на Python [157], пакетов Lime [158] и iml [150] на R.

2.3.3. ОГРАНИЧЕННЫЕ ПРАВИЛА (ЯКОРЯ)

Метод якоря объясняет отдельные прогнозы любой модели «черного ящика», находя правило принятия решения, которое достаточно «закрепляет» прогноз: правило «закрепляет» прогноз, если изменения в других значениях признаков не влияют на него. Якоря используют методы обучения с подкреплением (решение задачи однорукого бандита) в сочетании с алгоритмом поиска графа при сохранении возможности восстановления из локальных оптимумов [159]. Алгоритм был предложен теми же разработчиками, которые предложили LIME, однако вместо суррогатных моделей в этом методе формируются простые правила вида ЕСЛИ–ТО. Все правила имеют свою область применения и описываются ограничениями, связанными с функцией аппроксимации. Так как для каждого экземпляра создаются и оцениваются соседи или возмущения для применения правил, то это позволяет подходу быть модельно-независимым [143]. Якорь может быть выражен как

$$\mathrm{E}_{D_x(z|A)}\left[1_{f(x)=f(z)}\right] \geq \tau, \quad A(x)=1, \tag{2.4}$$

где x – объясняемый экземпляр; A – набор предикатов (правил, якорей); f – объясняемая модель, $D_x(\cdot \mid A)$ – распределение соседей x, соответствующих A; $0 \leq \tau \leq 1$ – порог точности.

Выражение может быть объяснено следующим образом: для экземпляра x необходимо найти правило (якорь) A так, чтобы тот же класс, что и у x, предсказывался, по крайней мере, для τ соседей, для которых применяется это правило. Точность правила определяется оценкой соседей и возмущений в соответствии с $D_x(z \mid A)$ с использованием предоставленной модели машинного обучения, обозначаемой функцией $1_{f(x)=f(z)}$.

Сравнение метода якоря с LIME на примере объяснения сложного бинарного классификатора («+» и «–») с использованием двух образцовых экземпляров показывает, что у результатов LIME не определяется их точность, так как в LIME изучается только линейная граница решения, заданная в пространстве *D*. В этом же случае метод якорей создает объяснения, адаптированные к поведению анализируемой модели и четко отображающие границы применения правил. Сравнение результатов применения методов показано на рис. 2.4 [159].

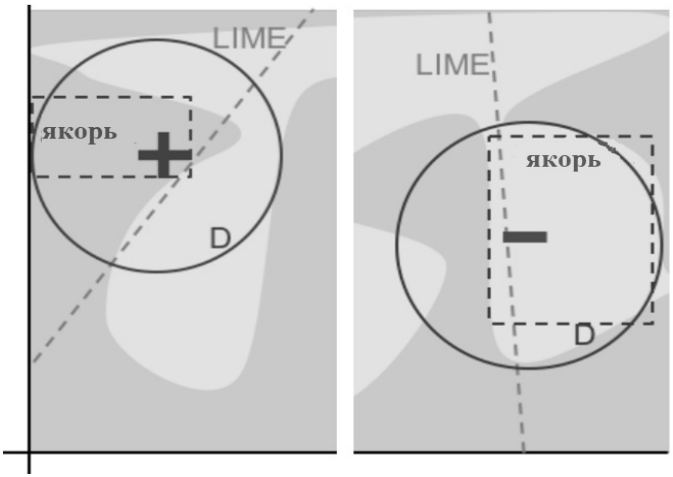


Рис. 2.4. Сравнение решений метода LIME и метода якоря

В [160] представлена разработка динамического якоря для динамических изображений: модель использует семантические признаки для прогнозирования оптимизированных форм якорей в местах, где объекты могут существовать на картах признаков. Предсказанные карты форм преобразуются на основании смещения местоположения.

Достоинства метода:

- легкая интерпретация объяснения по сравнению с LIME;
- установление важности правила (включая покрытие);
- высокая точность при нелинейности экземпляров (метод использует обучение с подкреплением вместо суррогатной модели, что позволяет избежать недообучения модели);
- возможность распараллеливания процесса с использованием многорукого бандита.

Недостатки метода:

- необходимость высокой точности настройки из-за того, что большинство объяснений базируется на возмущениях;
- необходимость введения дискретизации для повышения обобщающих свойств модели;
- реализация библиотек только для работы с табличными данными.

В статье [161] методы якоря и LIME используются для объяснения результатов распознавания кибератаки. По результатам экспериментов был сделан вывод о том, что оба метода дают точные объяснения, однако объяснение, получаемое с помощью метода LIME, базируется на зависимых факторах в отличие от метода якоря. В статье [162] отмечается преимущество применения метода якоря для объяснения результатов обнаружения кибератак, заключающееся в надежности и безопасности объяснения.

В статье [163] также показано, что метод якоря дает более стабильные результаты по сравнению с LIME при объяснении результатов прогнозирования появления болезней сердца.

В настоящее время доступны две реализации:

- пакет ancor, реализованный на Python (этот пакет был интегрирован компанией Alibi) [164];
- пакет ancor, реализованный на Java с интерфейсом R [165].

2.3.4. КОНТРАФАКТИЧЕСКИЕ ОБЪЯСНЕНИЯ

Контрафактические объяснения описывают причинно-следственную связь в форме: «*Если бы X* (*причина*) *не произошел, то Y* (*следствие*) *бы не произошел»*. В интерпретируемом машинном обучении контрафактические объяснения могут использоваться для объяснения прогнозов отдельных экземпляров. «*Событие*» – это прогнозируемый результат, «*причины*» – это конкретные значения признаков, которые были введены в модель и «вызвали» определенный прогноз. Отображаемая в виде графика связь между входными данными и прогнозом очень проста: значения признаков «вызывают» прогноз. Необходимо отметить, что связь между признаками и результатом может быть не причинно-следственной. Моделирование контрафактических объяснений выполняется на основе анализа того, как изменение значения признаков меняет выходной результат. Контрафактическое объяснение описывает наименьшее изменение значений признаков, которое изменяет предсказание на предопределенный результат.

В [166] для получения контрафактических объяснений предложено проводить минимизацию потерь:

$$L(x,\, x',\, y',\, \lambda) = \lambda (f(x') - y')^2 + d(x,\, x'), \tag{2.5}$$

где λ – параметр, уравновешивающий значения двух слагаемых; $(f(x') - y')^2$ – квадратичное расстояние между прогнозом модели для контрафактуального значения x' и желаемым результатом y', который пользователь должен определить заранее; $d(x,\, x')$ – расстояние между объясняемым результатом x и конрафактуальным x', рассчитываемое по следующему выражению:

$$d(x,\, x') = \sum_{j=1}^{p} \frac{\left| x_j - x'_j \right|}{MAD_j}, \tag{2.6}$$

где j – признак; $\left| x_j - x'_j \right|$ – манхэттенское расстояние; MAD_j – медианное абсолютное отклонение каждого признака.

Такой метод имеет некоторые недостатки. Во-первых, не создает контрафактуальных решений с несколькими изменениями признаков и вероятными значениями признаков. Другими словами, расчет d не представляет разреженных решений, так как увеличение десяти признаков на единицу даст то же расстояние до x, что и увеличение одного признака на десять. Во-вторых, нереалистичные комбинации признаков не штрафуются. В-третьих, метод плохо справляется с категориальными признаками. Авторы метода [156] предложили его отдельно для каждой комбинации значений категориальных признаков, но при большом количестве признаков это приведет к «комбинаторному взрыву»: например, шесть категориальных признаков с десятью уникальными уровнями приведут к необходимости проводить один миллион запусков [143].

В [167] предлагается решение, позволяющее избежать влияния указанных недостатков. Авторы предлагают также минимизировать ошибку модели по четырем параметрам:

$$L\left(x,\, x',\, y',\, X^{\text{obs}}\right) =$$

$$= \left(o_1(f(x'),\, y'),\, o_2(x,\, x'),\, o_3(x,\, x'),\, o_4(x',\, X^{\text{obs}})\right); \tag{2.7}$$

$$o_1(f(x'), y') = \begin{cases} 0, & f(x') \in y', \\ \inf\limits_{y' \in y'} \left| f(x') - y' \right|, & f(x') \notin y'; \end{cases} \tag{2.8}$$

$$o_2(x, x') = \frac{1}{p} \sum_{j=1}^{p} \delta_G \left(x_j, x'_j \right); \tag{2.9}$$

$$o_3(x, x') = \left\| x - x' \right\|_0 = \sum_{j=1}^{p} \mathrm{I}_{x'_j \neq x_j}; \tag{2.10}$$

$$o_2\left(x', X^{\mathrm{obs}}\right) = \frac{1}{p} \sum_{j=1}^{p} \delta_G \left(x'_j, x^{\mathrm{obs}} \right), \tag{2.11}$$

где o_1 – первая цель, отражающая, что предсказание контрафактуального x' должно быть близко к желаемому y', и рассчитываемая по манхэттенской метрике в соответствии с формулой (2.8); o_2 – вторая цель, отражающая, что x' может быть похоже на x, и определяющая расстояние между данными величинами как расстояние Гауэра в соответствии с (2.9); o_3 – третья цель, отражающая, что x' может быть похоже на x, и учитывающая количество измененных признаков в соответствии с (2.10); o_4 – четвертая цель, отражающая, что контрафактические значения x' должны иметь вероятные значения/комбинации признаков из датасета X^{obs}, и рассчитываемая в соответствии с (2.11); где p – количество признаков; δ_G – значение, определяющее тип x_j (числовое, категориальное).

Преимущества метода:

- относительная простота интерпретации (особенно в сравнении с LIME);
- возможность обобщения результатов с помощью контрафактуальных экземпляров или с помощью разницы между объясняемым и конрафактуальными экземплярами;
- высокая безопасность модели (применение метода не требует раскрытия модели и данных);

- относительная простота реализации;
- возможность работы не только с системами на базе машинного обучения.

Недостатком модели является возможность применения множества контрафактуальных объяснений, которые необходимо анализировать.

В статье [168] показано, что контрафактуальные объяснения находят наиболее широкое применение в системах, основанных на определенных и однозначных правилах (например, объяснение решения о предоставлении кредита). В статье [169] показывается пример применения контрафактуальных объяснений для системы оценки осанки человека. Авторы используют формализированные признаки, на основании которых можно выделить однозначные правила.

Контрафактуальное объяснение может быть реализовано:

- многоцелевым методом [170];
- пакетами Alibi на языке Python, которые реализуют простой и расширенные методы контрафактуального объяснения [171];
- алгоритмами MACE [172], разнообразного контрафактуального объяснения [173].

2.4. АДДИТИВНОЕ ОБЪЯСНЕНИЕ ШЕПЛИ

В настоящем разделе рассматриваются два метода: значения Шепли и аддитивное объяснение Шепли (SHAP). SHAP основывается на методе вычисления Шепли, однако может быть применен также для глобальной интерпретации, основанной на комбинациях значений Шепли по всем данным [143].

2.4.1. ЗНАЧЕНИЯ ШЕПЛИ

Значения Шепли были предложены Л. Шепли [174]. Метод расчета значений Шепли относится к теории игр и описывает распределение выплат между игроками в зависимости от их вклада в общий результат. Игроки сотрудничают в коалиции и получают определенную прибыль от этого.

Если применять такую методику к машинному обучению, то под термином *игра* может пониматься задача прогнозирования для одного экземпляра набора данных. Под термином *выигрыш* понимается фактический прогноз для этого экземпляра за вычетом среднего прогноза

для всех экземпляров. Под *игроками* понимаются значения признаков экземпляра, которые подаются на вход модели для получения прогноза выигрыша.

При применении этого метода необходимо разделять термины:

- *значение признака* определяется как число или категория;
- *значение Шепли* определяется как вклад признака в прогноз;
- *функция значения* определяется как функция выплат игрокам.

Значение Шепли рассчитывается в соответствии с формулой

$$\varphi_j(\text{val}) = \sum_{S \subseteq \{1,\ldots,p\} \setminus \{j\}} \frac{|S|!\left(p - |S| - 1\right)!}{p!} \left(\text{val}(S \cup \{j\}) - \text{val}(S)\right), \quad (2.12)$$

где S – подмножество признаков, используемых в модели; j – признак, влияние которого необходимо объяснить; p – количество признаков.

Интерпретация значения Шепли для j-го признака может быть сформулирована следующим образом: как значение j-го признака внесло вклад φ_j в прогноз конкретного случая по сравнению со средним прогнозом для всего набора данных? Все возможные наборы значений признаков должны быть оценены с j-м признаком в наборе данных и без него.

Метод значений Шепли удовлетворяет свойствам:

- эффективности (вклады признаков суммируются с разницей прогноза и среднего значения):

$$\sum_{i=1}^{p} \varphi_j = f(x) - E_X(f(X)); \quad (2.13)$$

- симметрии (вклады двух значений признаков j и k должны быть одинаковыми, если они вносят одинаковый вклад в прогноз):

$$\varphi_j = \varphi_k, \ \text{val}(S \cup \{j\}) = \text{val}(S \cup \{k\}), \ S \subseteq \{1, \ldots, p\} \setminus \{j, k\}; \quad (2.14)$$

- фиктивности (значение Шепли для признака j должно быть равно 0, если признак не вносит вклада в прогноз вне зависимости от того, к какой коалиции он добавляется):

$$\varphi_j = 0, \ \text{val}(S \cup \{j\}) = \text{val}(S), \ S \subseteq \{1, \ldots, p\}; \quad (2.15)$$

• аддитивности $\left($для игры с комбинированными выплатами $\mathrm{val} + \mathrm{val}^{+}$ значение Шепли должно быть рассчитано как $\varphi_j + \varphi_j^{+}\right)$.

При этом, так как все признаки в модели связаны друг с другом, необходимо понимать, что значение Шепли – это средний вклад значения признака в прогноз в разных коалициях, а не разница в прогнозе после удаления признака из модели [143].

Преимущества данного метода:

• аддитивность и справедливость распределения прогноза между всеми используемыми признаками, что делает значение Шепли единственным методом, совместимым *с правом на объяснение* (например, в LIME прогноз может не быть распределен между всеми признаками);

• возможность контрастных объяснений (возможность сравнения результатов между подмножествами, между разными точками данных);

• наличие теоретической основы из теории игр.

Недостатки значений Шепли:

• необходимость использования больших вычислительных мощностей из-за экспоненциального роста количества коалиций;

• нежелательность применения при поиске разряженных объяснений (объяснений с малым количеством признаков);

• необходимость доступа для расчета значений к данным, а не только к функции прогнозирования;

• большое влияние нереалистичных экземпляров данных на результат объяснения.

Применение метода расчета значений Шепли нашло широкое распространение в энергетике. В [175] предлагается модель прогнозирования нагрузки на основании алгоритма распределения весов длительной краткосрочной памяти (WSLSTM) и значений Шепли. Значения Шепли в данной статье использовались для выбора наиболее влияющих на прогноз признаков. Этот метод был выбран на основании того, что он может оценить нелинейную связь между признаками.

В статьях [176, 177] значения Шепли применяются для обеспечения устойчивой торговли электроэнергией между микросетями и коммунальными сетями. В исследованиях изучалась совместная работа микросетей и коммунальной сети. Авторы этих статей предложили методы, обеспечивающие устойчивую торговлю на основе индивидуальных вкладов в снижение ежедневных затрат на генерацию в случае, когда

генерация распределяется между микросетями и коммунальными сетями. Значения Шепли были применены для расчета платежей за электроэнергию между микросетью и коммунальной сетью с учетом их показателей надежности. Исследования показали, что микросети и коммунальные сети *выигрывают*, когда они *сотрудничают* в обмене энергией независимо от их индивидуальных *вкладов* в *коалицию* по обмену энергией.

В статье [178] значения Шепли использовались для проектирования электросетей. Выводом исследования стала доказанность возможности применять значения Шепли для моделирования и проектирования децентрализованных энергетических систем.

В статье [179] значения Шепли использовались для составления расписания зарядки электроавтомобилей в случае создания зарядных станций на территории компаний-работодателей. Для этого была сформирована коалиция компаний и владельцев электроавтомобилей. В статье доказано, что создание коалиции позволяет компании и владельцам электромобилей снизить годовую плату за потребляемую электроэнергию.

Значения Шепли могут быть рассчитаны в пакетах iml [150], fastshap [180], для R. В Julia можно использовать Shapley.jl [181].

Кроме аддитивного объяснения Шепли, описание которого будет представлено в разделе 2.4.2, можно использовать подход breakdown, который демонстрирует вклад каждой функции в прогноз с последовательным добавлением значений признаков. Достоинство метода – высокая скорость расчета значений, недостаток – зависимость от значений функций, которые уже находятся в коалиции [143].

2.4.2. АДДИТИВНОЕ ОБЪЯСНЕНИЕ ШЕПЛИ (SHAP)

Метод аддитивного объяснения Шепли (SHAP) был впервые предложен Ландбергом и Ли в [182]. SHAP основан на оптимальных значениях Шепли. SHAP является самостоятельным алгоритмом, так как:

- авторы предложили альтернативные подходы к оценке значений Шепли:
  - KernelSHAP, который оценивает значения Шепли с использованием *ядерного сглаживания* (kernel density estimation);
  - TreeSHAP, который оценивает значения Шепли на основе деревьев решений;
- SHAP может быть использован с глобальными методами объяснения, основанными на агрегации значений Шепли.

При использовании SHAP игроки могут быть не только признаками, но и группой признаков (например, пиксели могут быть объединены в группу пикселей) [143]. В статье [183] представлена модификация алгоритма SHAP, которая заключается в объяснении ряда признаков, имеющих близкое значение для пользователя при визуальной интерпретации в задаче прогнозирования солнечной радиации, для прогнозирования генерации солнечных электростанций.

Одним из новшеств, которое привносит SHAP, является возможность использования подхода LIME и значений Шепли:

$$g(z') = \varphi_0 + \sum_{j=1}^{M} \varphi_j z'_j , \tag{2.16}$$

где g – модель объяснения; $z = \{0,1\}^M$ – вектор коалиции, показывающий присутствие признака в игре; M – максимальный размер коалиции.

Для вычисления интересующего экземпляра x вектор коалиции состоит только из единиц, таким образом:

$$g(x') = \varphi_0 + \sum_{j=1}^{M} \varphi_j . \tag{2.17}$$

SHAP удовлетворяет следующим свойствам:

- локальной точности, которая при $\varphi_0 = E_X(f(x))$ и при всех $x'_j = 1$ совпадает со свойством эффективности значений Шепли;
- отсутствию признака – отсутствующий признак получает атрибут нуль:

$$x'_j = 0 \Rightarrow \varphi_j = 0 ; \tag{2.18}$$

- согласованности, которая говорит, что если модель изменяется так, что предельный вклад значения признака увеличивается или остается прежним (независимо от других признаков), то число Шепли также увеличивается или остается прежним.

Рассмотрим модификации SHAP.

KernelSHAP оценивает вклад каждого признака в прогноз для экземпляра x и состоит из пяти шагов:

1) выборки коалиции $z'_k \in \{0,1\}^M$, $k \in \{1, ..., K\}$;

2) получения прогноза для каждого z'_k с преобразованием z'_k в исходное пространство признаков и применением модели $f : f\left(h_x\left(z'_k\right)\right)$, где $h_x(z') = z,\ h_x : \{0,1\}^M \to R^p$;

3) вычисления весов для каждого z'_k с использованием ядра SHAP, которое может быть сформулировано как

$$\pi_x(z') = \frac{(M-1)}{\binom{M}{|z'|}|z'|(M-|z'|)}, \tag{2.19}$$

где M – максимальный размер коалиции; $|z'|$ – количество имеющихся признаков в экземпляре z';

4) подбора взвешенной линейной модели;

5) возвращения значений Шепли φ_k, коэффициентов линейной модели.

Так как выборка коалиции игнорирует структуру между имеющимися и отсутствующими признаками, KernelSHAP ошибочно присваевает большие веса маловероятным случаям. Для того чтобы избежать такой ситуации, можно сделать выборки из условного распределения, а не из предельного. В результате значения Шепли будут иметь другую интерпретацию: например, признак, который мог бы вообще не использоваться моделью, может иметь ненулевое значение Шепли при использовании условной выборки. Для предельной игры этот признак всегда будет получать значение Шепли, равное 0, поскольку в противном случае это нарушит аксиому фиктивности.

Различие между KernelSHAP и LIME заключается во взвешивании экземпляров регрессионной модели. LIME оценивает экземпляры на основании расстояния до исходного экземпляра. KernelSHAP взвешивает экземпляры в соответствии с весом, который коалиция получает при оценке значения Шепли.

TreeSHAP является быстрой модификацией KernelSHAP. TreeSHAP определяет функцию значения, рассчитывая условное ожидание $E_{X_j|X_{-j}(f(x)|x_j)}$ вместо предельного. Проблемой nfrjuj подхода является то, что алгоритм может давать признакам, которые не влияют

на функцию *f*, ненулевые значения, если этот признак коррелирует с другим признаком.

Рассмотрим примеры применения SHAP для различных задач.

В статье [184] SHAP применяется для объяснения полученного значения диэлектрической проницаемости материала (рис. 2.5).

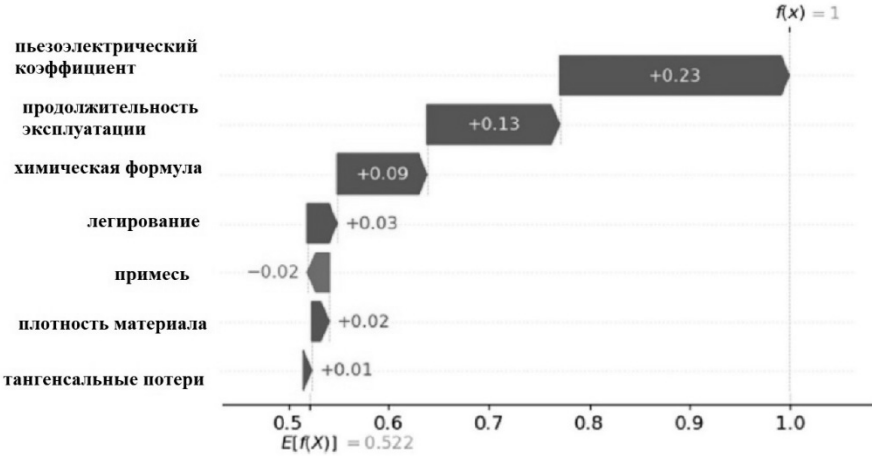


Рис. 2.5. Важность признаков SHAP для прогноза диэлектрической проницаемости материала

Визуализация на рис. 2.5 демонстрирует значимость отдельных признаков, начиная с ожидаемого значения ($E[f(X)] = 0{,}522$) и заканчивая полученным конкретным прогнозом ($f(X) = 1$). Анализ показывает, что переменные «пьезоэлектрический коэффициент» и «продолжительность эксплуатации» оказывают значительное положительное влияние на прогноз, в то время как переменная «примесь» оказывает незначительное отрицательное влияние.

При глобальной интерпретации можно рассчитать важность каждого признака SHAP по всем данным:

$$I_j = \frac{1}{N}\sum_{i=1}^{n}\left|\varphi_j^{(i)}\right|, \tag{2.20}$$

После расчета признаков происходит сортировка признаков по убыванию важности для построения графика.

В статье [185] значения важности признаков SHAP были рассчитаны для выделения наиболее влияющих признаков на коэффициент самообеспечения электроэнергией (ESSR) хозяйств (рис. 2.6).

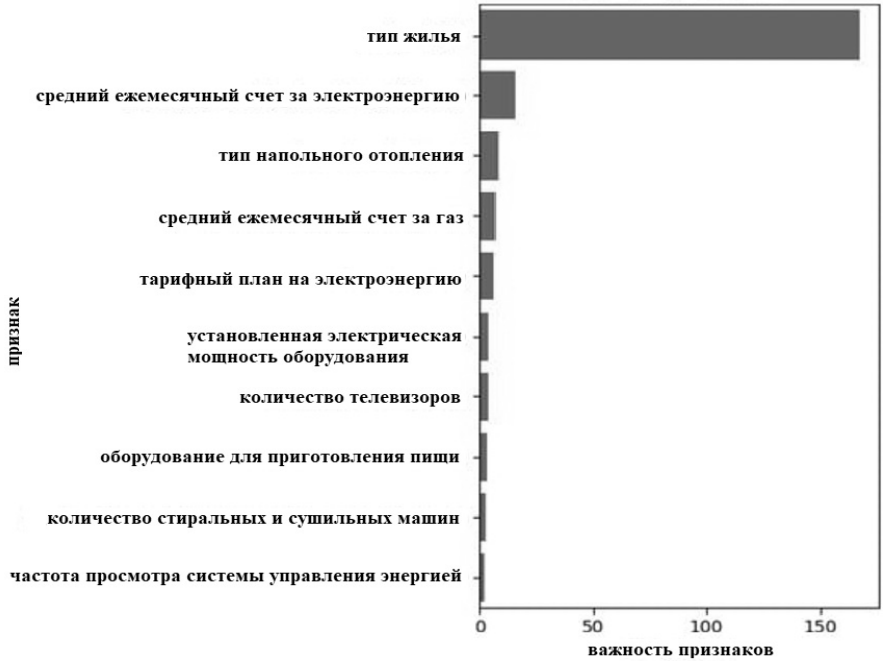


Рис. 2.6. Важность признаков SHAP, повлиявших на значение ESSR

Для того чтобы визуализировать общую силу признаков, которая влияет на прогноз, строят «силовое» влияние SHAP значений. Прогноз начинается с базовой линии (среднего значения всех прогнозов). На графике каждое значение Шепли представляется стрелкой, которая либо увеличивает (положительное значение), либо уменьшает (отрицательное значение) прогноз.

В статье [186] исследуется применение SHAP в задаче по автоматизации операций в центре обработки данных. На рис. 2.7 представлена локальная интерпретация результатов прогноза потребленной электрической энергии, полученных с помощью экстремального градиентного бустинга, с использованием силового графика SHAP.

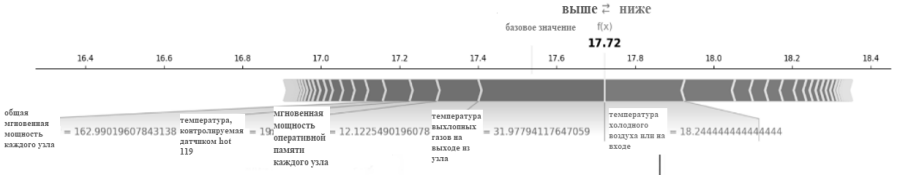


Рис. 2.7. Пример силового графика SHAP

Для того чтобы проанализировать связь между значением признаков и их влиянием на прогноз, можно построить сводный график. Каждая точка на сводном графике будет значением Шепли для признака и экземпляра. Положение на оси *y* определяется признаком, а на оси *x* – значением Шепли. Перекрывающиеся точки смещаются в направлении оси *y*, поэтому получается представление о распределении значений Шепли для каждого признака. Признаки упорядочены в соответствии с их важностью.

На рис. 2.8 показано влияние признаков на результаты прогнозирования потребления электрической энергии, полученные с помощью экстремального градиентного бустинга [186].

Для того чтобы увидеть точную форму связи, можно построить графики зависимости $\left\{\left(x_j^{(i)}, \varphi_j^{(i)}\right)\right\}_{i=1}^{n}$. График зависимости показывает дисперсию значений Шепли по оси *y* относительно значений признака по оси *x*. Пример графика зависимости между значениями Шепли и признаками для значений «*огонь, пламя или горячие вещества*», полученный в исследовании по оценке риска возникновения пожара на стадионах [187], представлен на рис. 2.8.

Для того чтобы раскрасить график зависимости признаков SHAP с самыми сильными взаимодействиями, можно рассчитать эффект взаимодействия между признаками, который в соответствии с теорией игр определяется как

$$\varphi_{i,j} = \sum_{S \subseteq \{i,\, j\}} \frac{|S|!\left(M - |S| - 2\right)!}{2(M-1)!}\, \delta_{ij}(S), \qquad (2.21)$$

где $\delta_{ij}(S)$ – взаимодействие признаков без учета индивидуальных эффектов, которое рассчитывается в соответствии с

$$\delta_{ij}(S) = f_x(S \cup \{i, j\}) - f_x(S \cup \{i\}) - f_x(S \cup \{j\}) + f_x(S), \quad (2.22)$$

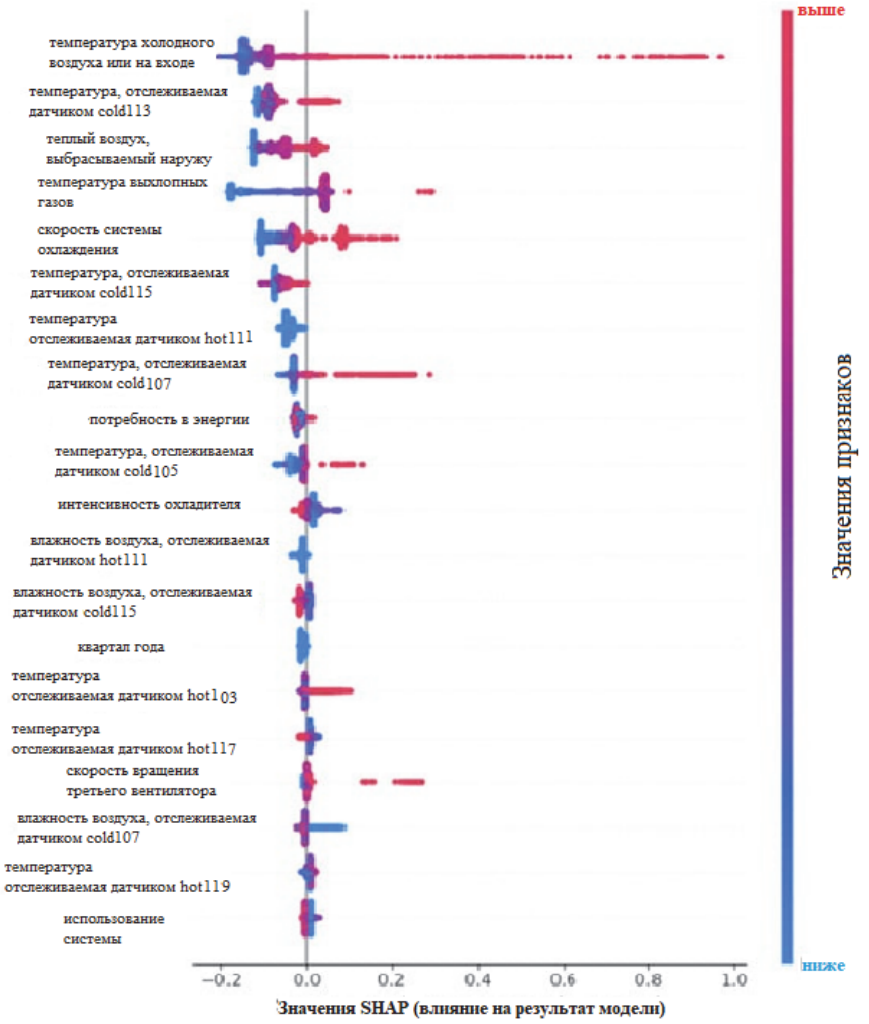


Рис. 2.8. Пример сводного графика SHAP

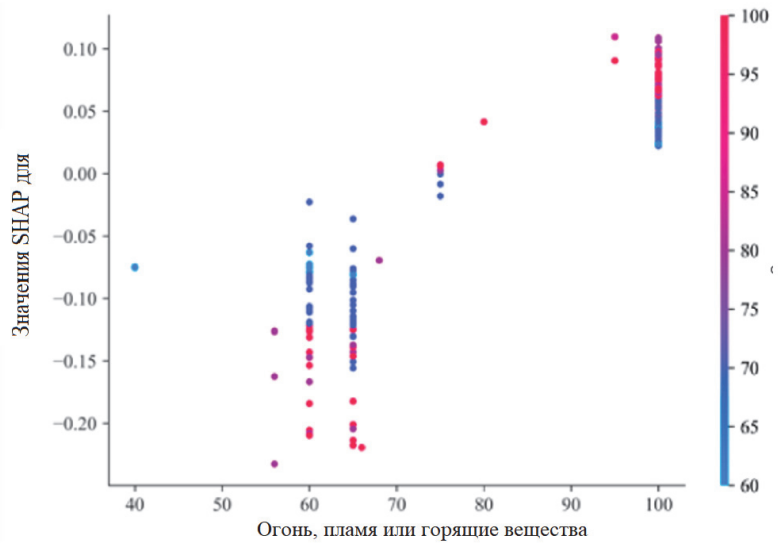


Рис. 2.9. Пример графика зависимости между значениями Шепли и признаками для значений «*огонь, пламя или горячие вещества*»

Для нахождения похожих экземпляров можно провести кластеризацию значений Шепли каждого экземпляра: другими словами, кластеризация экземпляров базируется на сходстве объяснений (рис. 2.10) [188].

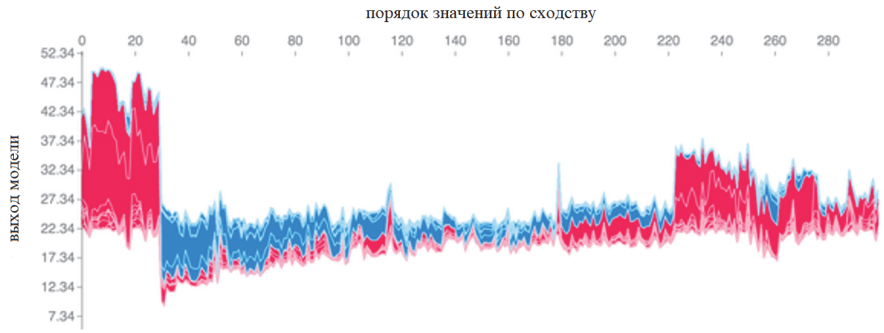


Рис. 2.10. Пример кластеризации значений Шепли

Объяснения SHAP группируются по сходству объяснений. Каждая позиция на оси *x* – это экземпляр данных. Красные значения SHAP увеличивают прогноз, синие значения уменьшают его.

Преимущества SHAP:

- наличие теоретической основы из теории игр;
- возможность объединения подходов LIME и значений Шепли;
- быстрая реализация для древовидных моделей;
- возможность согласования глобальной интерпретации с локальными объяснениями.

Недостатки метода:

- необходимость расчета значений Шепли для многих экземпляров;
- игнорирование зависимостей между признаками в KernelSHAP, которое может привести к приданию большого веса маловероятным точкам данных;
- получение неинтуитивных атрибуций признаков в TreeSHAP;
- необходимость доступа к данным.

SHAP был реализован в пакете Shap на Python [189]. Эта реализация работает для моделей на основе деревьев в библиотеке машинного обучения Scikit-learn для Python.

SHAP реализован на R в пакетах Shapper [190] и Fastshap [191].

2.5. МЕТОДЫ ОБЪЯСНЕНИЯ РЕЗУЛЬТАТОВ ГЛУБОКИХ НЕЙРОСЕТЕВЫХ МОДЕЛЕЙ

Существуют методы объяснения результатов глубоких нейросетевых моделей:

- визуальные (метод создают изображения или графики для объяснения);
- текстовые (данные в текстовой форме);
- математические и числовые.

В настоящем разделе рассматриваются визуальные методы объяснения.

2.5.1. МЕТОД CAM

Одним из популярных визуальных методов является метод карты активации класса (CAM – Class Activation Mapping). CAM способен локализовать на изображении признаки сверточной нейронной сети, имевшие наибольшее влияние на принятие решений о классификации. CAM использует слой усреднения после сверточных слоев и перед последним полносвязным слоем [192].

Пусть $f_k(x, y)$ обозначает активационную единицу, ω_c^k – вес, соответствующий классу c для единицы k. Тогда вход для слоя *sorftmax*, соответствующий классу c, для k определяется как

$$S_c = \sum_{x,y} \sum_k \omega_c^k f_k(x, y). \quad (2.23)$$

Тогда карта активации класса M_c, показывающая важность активации в пространственной точке (x, y), для классификации класса c будет рассчитана так:

$$M_c = \sum_k \omega_c^k f_k(x, y). \quad (2.24)$$

Метод CAM нашел широкое распространение в медицине. В статье [193] приводится обзор 45 исследований, в которых показывается эффективность применения CAM для обнаружения патологических областей на снимках. В статье [194] приводится исследование по эффективности применения CAM для распознавания патологий на снимках легких, в статье [195] – исследование по применению CAM для визуализации очагов заболевания на рентгеновских снимках грудной клетки.

В статье [196] CAM используется для распознавания действий людей. Эксперимент был проведен с использованием тепловых изображений Dongguk.

В статье [197] предлагается применение CAM для обработки изображений дистанционного зондирования земли: CAM используется для извлечения признаков частей объекта и фильтрации фоновых помех, что позволяет экспертам анализировать параметры объектов без учета помех. Эксперименты были проведены в задаче распознавания воздушных судов на изображениях дистанционного зондирования.

CAM реализован в PyTorch на языке Python [198, 199]. Intel реализовал набор инструментов CAM-Visualizer [200]. CAM может быть применен в MATLAB [201].

2.5.2. МЕТОД GRAD-CAM

На основе метода CAM был создан метод градиентно-взвешенной карты активации классов (Grad-CAM – Gradient-weighted Class Activation Mapping). Grad-CAM использует градиенты относительного

целевого класса c, который поступает в конечный сверточный слой. Grad-CAM генерирует грубую карту локализации $L^c_{\text{Grad-CAM}} \in R^{v \times u}$ шириной v и высотой u, выделяющей важные пиксели для классификации изображения.

Сначала вычисляются градиенты $\frac{\partial y^c}{\partial A_k}$ классовой оценки y^c по отношению к картам активации A_k последнего сверточного слоя. Градиенты возвращаются после усреднения по размеру карты активации Z, а затем вычисляются веса важности нейронов a_k^c в соответствии с формулой [202]

$$a_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_k(i,j)}. \tag{2.25}$$

Весовой коэффициент a_k^c показывает важность признака k для класса c. Наконец, тепловые карты Grad-CAM создаются с использованием активаций прямого распространения следующим образом:

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k a_k^c A_k\right). \tag{2.26}$$

Grad-CAM нашел широкое применение в инженерных отраслях.

В статье [203] приводится исследование по применению Grad-CAM для объяснения оценки ситуации в неструктурированной дорожной среде, которое показало, что объяснение результатов может способствовать внедрению автономных транспортных средств в Индии. В статье [204] Grad-CAM используется для объяснения типов объектов в облаке точек LiDAR на дорогах.

В статье [205] Grad-CAM используется для определения ключевых характеристик, существенно влияющих на оценку переходной устойчивости энергосистемы. Применение Grad-CAM позволяет выявить слабые места сети.

В статье [206] Grad-CAM используется для объяснения результатов детектирования частичных разрядов в трансформаторах. Эксперимен-

тальные данные включали искусственные дефекты, созданные в лабораторных условиях, и шум, зафиксированный в трансформаторах с литой изоляцией при помощи акустических датчиков. Исследование показало, что применение Grand-CAM для объяснения результатов глубокой нейронной сети позволило повысить надежность системы определения частичных разрядов. На рис. 2.11 представлены изображения активации.

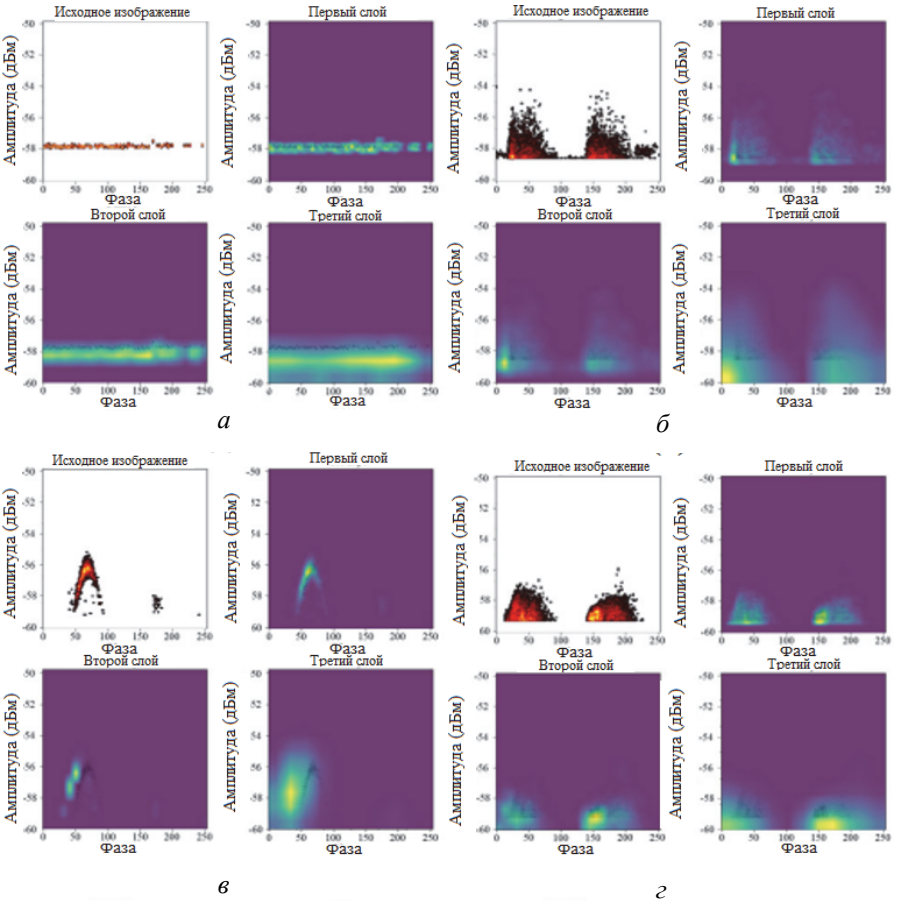


Рис. 2.11. Карты активации Grad-CAM для объяснения вида частичного разряда:

а – шум; *б* – поверхностный разряд; *в* – коронный разряд; *г* – разряд пустот, происходящий во внутренних пустотах литой изоляции

В статье [207] Grad-CAM используется для объяснения решений по оценке износа буровых долот. В [208] рассматривается задача детектирования посторонних предметов в шинах. В [209] этот метод помогает оценивать неисправность подшипников. На рис. 2.12 представлены карты активации Grad-CAM для различных типов неисправностей подшипников.

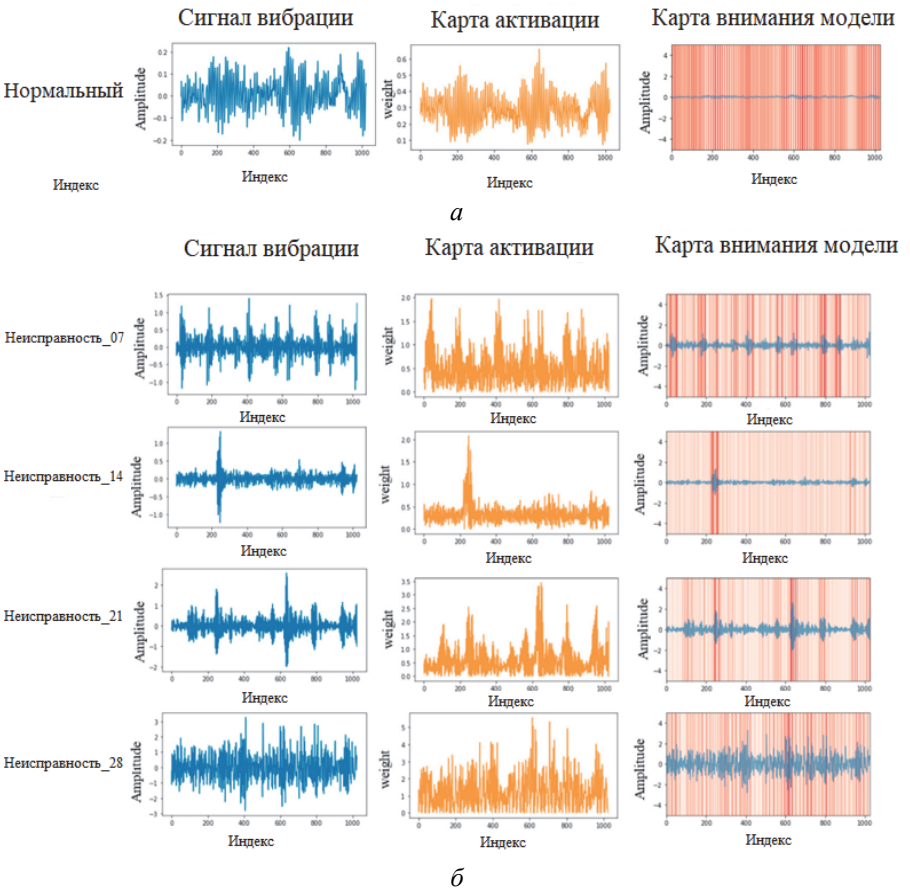
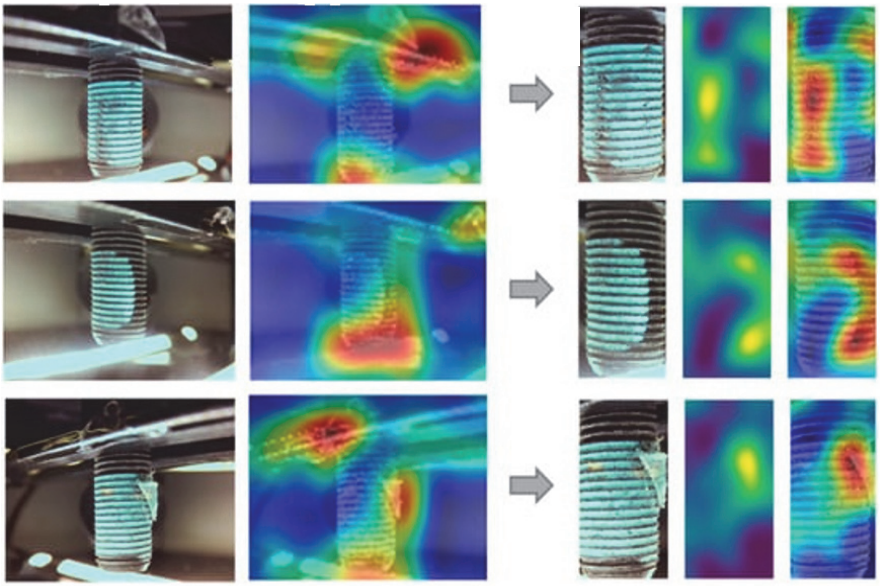


Рис. 2.12. Карты активации Grad-CAM для различных типов неисправностей подшипника:

а – нормальный сигнал; *б* – сигналы неисправностей внутренней дорожки подшипника

В [210] Grad-CAM используется для детектирования наличия покрытия на болтовых соединениях, которые препятствуют их ослаблению. На рис 2.13, демонстрирующем применение Grad-CAM, видно, что на результаты определения больше влияют фоновые факторы (освещение, зажимы), а не площадь нанесения покрытия. Для минимизации этих факторов рекомендуется предварительно обрабатывать изображения таким образом, чтобы оставлять только анализируемую область (ROI – region of interest).



а *б*

Рис. 2.13. Карты активации Grad-CAM для детектирования наличия покрытия: *а* – анализ всего изображения; *б* – анализ предобработанного изображения

Grand-CAM реализован в библиотеке PyTorch на языке Python [211] и в MATLAB [212].

2.5.3. МЕТОД LRP

Метод распространения релевантности по слоям (LRP – Layer-Wise Relevance Propagation) основан на декомпозиции решения: проводит оценку релевантности между активациями $x^{(i)}$ нейрона i и его входом

на слое l $\left(R_l^{(i)}\right)$. Более конкретно оценка релевантности $R_l^{(i)}$ слоя l рассчитывается относительно слоя $l + 1$ как [213]

$$R^l(i) = \sum_j \frac{x(i)\omega(i,j)}{\sum_i x(i)\omega(i,j)} R^{l+1}(j), \tag{2.27}$$

где $\omega(i,j)$ – вес между нейронами i и j.

Оценка объяснительной силы LRP представлена в исследовании [214]. Алгоритм содержал несколько этапов. Изначально были созданы состязательные примеры на основе исходных изображений внесением искажений в оригинальные изображения. Назначением состязательных примеров было моделирование возмущений. Затем для исходных и состязательных примеров рассчитывалась оценка релевантности. На последнем этапе происходила оценка объяснительной силы статистическими методами (анализом частотного распределения).

Проведенный анализ показал отсутствие существенной разницы между картами исходных и состязательных изображений. Следовательно, в текущей версии LRP не соответствует требованиям к надежности объяснения решений глубоких нейронных сетей.

В статье [216] LRP используется для объяснения решений системы, направленной на стабилизацию работы микросети с учетом экономического фактора. В статье продемонстрировано, что наиболее важным входом в системе является время реакции участников, за которым следует коэффициент эластичности цен, а потребление или генерация электроэнергии практически не влияет на стабильность микросети.

В статьях [217, 218] LRP применяется для объяснения решений по распознаванию объектов на изображениях, полученных с помощью радиолокатора с синтезированной апертурой (SAR – Synthetic Aperture Radar). Сложность интерпретирования SAR-изображений заключается во влиянии на них методов предварительной обработки. В статьях LRP применяется для определения целевых областей SAR.

LRP может быть реализован с помощью библиотек в PyTorch на языке Python [219–222] и на MATLAB [108].

2.5.4. МЕТОД PRM

Метод карт пикового отклика (PRM – Peak Response Maps) используется для повышения способности локализации экземпляров сверточных нейронных сетей. PRM находит максимальные активации классов, которые определяют их баллы в каждом местоположении изображения, после чего рассчитанные активации распространяются обратно на входное изображение для генерации карт пикового отклика. При этом места расположения пиков рассчитываются в соответствии с формулой [222]

$$P_c = \left\{(i_1, j_1),\ ...,\ \left(i_{N^c}, j_{N^c}\right)\right\}, \tag{2.28}$$

где N^c – номер пика c-й карты отклика M_c, которая извлекается из локальных максимумов внутри окна размером 3×3.

Ядро выборки рассчитывается в точке (x, y) как

$$G^c(x, y) = \sum_{k=1}^{N^c} f(x - i_k, y - j_k), \tag{2.29}$$

где $x \in [0, H]$, $y \in [0, W]$; f – функция выборки, которая извлекается из признаков пиков.

Оценка доверия к классу s^c рассчитывается из свертки карты отклика и ядра выборки как

$$s^c = M^c G^c. \tag{2.30}$$

Градиенты, которые будут передаваться обратно, рассчитываются как

$$\delta^c = \frac{1}{N^c} \frac{\partial L}{\partial s^c} G^c, \tag{2.31}$$

где L – функция потерь классификации.

В настоящее время PRM не находит такого широкого применения, как ранее рассмотренные методы.

В статье [223] в задаче сегментации экземпляров при отсутствии достаточного количества маркированных данных для обучения модели PRM применяется для того, чтобы показать успешность применения метода, разработанного авторами. Результаты демонстрируют низкую точность определения объектов на изображениях при использовании PRM (рис. 2.14).

а *б* *в*

Рис. 2.14. Репрезентативные результаты определения объектов:

а – входное изображение; *б* – истинные результаты; *в* – результаты, полученные PRM

Несмотря на то что с помощью PRM можно довольно точно локализовать экземпляры каждого класса, авторы [224] отмечают также проблему PRM, связанную с недостаточной надежностью информации для сегментации. Это связано с тем, что при применении PRM нельзя понять, какие области можно считать полным экземпляром.

В статье [225] PRM используется для объяснения обнаружения ветряных турбин на спутниковых снимках. На рис. 2.15 показана последовательность шагов алгоритма нахождения объектов:

- на вход слабо контролируемой модели локализации подают спутниковое изображение;

- карту прогнозируемого отклика, локализующую турбины на изображении, из контролируемой модели локализации передают в модуль обнаружения пиков;
- модуль обнаружения пиков определяет локальные максимумы на карте отклика и объединяет их на основе близости;
- прогнозируемые местоположения на изображении преобразуются в географические местоположения (координаты широты, долготы) турбин.

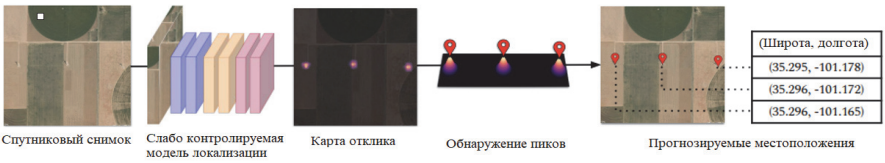


Рис. 2.15. Алгоритм локализации ветряных турбин

По результатам исследования [225] метод PRM показал лучшие результаты по локализации ветряных турбин в сравнении с CAM и Grad-CAM.

Решения с использованием PRM могут быть реализованы в PyTorch на языке Python [226] и Matlab [227].

2.5.5. МЕТОД CLEAR

Метод класса повышенной внимательности (CLEAR – CLass-Enhanced Attentive Response) представляет собой подход для визуализации решений глубокой нейронной сети с использованием значений активации сети. Он использует алгоритм обратной свертки (*деконволюции*) для получения индивидуальных карт внимания каждого класса. После прямого прохода применяется обратная свертка для получения деконволюционного выходного слоя l с K ядрами [228]:

$$h(l) = \sum_{k=1}^{K} z(k,\ l)\omega(k,\ l), \tag{2.32}$$

где $z(k,\ l)$ – карта признаков слоя l; $\omega(k,\ l)$ – вес ядра.

Итоговый ответ слоя l рассчитывается так:

$$R(l) = h(1)h(2)...h(l). \tag{2.33}$$

Индивидуальные карты внимания $R(x', c)$ класса c и обратно проецированного входа x' вычисляются для всех L слоев:

$$R(x', c) = h(1)h(2)...h(L), \tag{2.34}$$

После нахождения данных значений строится карта внимания доминирующего класса $C(x')$, которая показывает наиболее влиятельные классы:

$$C(x') = \arg\max R(x', c). \tag{2.35}$$

При этом доминирующая карта реакций (карта ответов) $D_c(x')$ определяется на основе идентифицированного класса:

$$D_c(x') = R(x', c). \tag{2.36}$$

Карта класса повышенной реакции строится как

$$M = C(x') + D_c(x'). \tag{2.37}$$

В статье [229] авторы метода применили его для объяснения и визуализации прогнозов фондового рынка, модифицировав метод в CLEAR-Trade. Результаты показывают, что CLEAR-Trade может обеспечить объяснение процесса принятия решений для регулирования рынка, тем самым повышая вероятность их потенциального внедрения в финансовой отрасли.

В 2019 году был предложен метод CLEAR-DR [230], который основывался на CLEAR и решал задачу классификации диабетической ретинопатии (DR (diabetic retinopathy)). Авторы применили метод в наборе из более чем 50 000 изображений сетчатки глаза. Они создали тепловые карты для каждого из пяти классов: легкая, умеренная, тяжелая, пролиферативная, отрицательная диабетическая ретинопатия. Результаты исследования показали, что на правильно классифицированных изображениях карты CLEAR-DR соответствовали правильным частям анатомии глаза, а в случаях неправильной классификации CLEAR-DR не смог сосредоточиться на соответствующей аномалии. Таким образом, разработанный метод не отвечал требованиям надежности.

2.5.6. МЕТОД DEEPRESOLVE

Метод глубокой решимости (DeepResolve) использует карты признаков из промежуточных слоев и анализирует, как сеть комбинирует эти признаки для классификации входного изображения [192]. DeepResolve

вычисляет изображение, специфичное для каждого класса, которое называется картой важности признаков (FIM – Feature Importance Map) [192]:

$$H^c = \arg\max_H \left(S_c(H) - \lambda \|H\|_2^2 \right), \tag{2.38}$$

где c – целевой класс; S_c – классовый балл, полученный из последнего слоя; $H \in R^{K \times W}$ содержит карты признаков размером W для всех K нейронов из определенного слоя.

После получения FIM рассчитывается оценка важности признака:

$$\Phi_c = \left(\varphi_c^1, \varphi_c^2, ..., \varphi_c^k \right), \tag{2.39}$$

где φ_c^k рассчитывается из FIM как

$$\varphi_c^k = \frac{1}{W} \sum_{i=1} (H^i(i))_c , \tag{2.40}$$

где i – индекс нейрона, k – индекс канала в слое.

Процесс расчета инициализируется случайным образом и повторяется T раз с различными начальными параметрами для получения нескольких оценок H_c^t, Φ_c^t. После этого рассчитывается взвешенная дисперсия IL_c^k для получения общих оценок важности нейронов $\overline{\Phi_c}$, которая определяет сходства и различия между классами и используется для построения матрицы сходства. Взвешенная дисперсия рассчитывается по формуле

$$IL_c^k = \mathrm{var}\left(\varphi_c^t \right). \tag{2.41}$$

Матрица различий между классами для пары классов C_i, C_j рассчитывается как

$$D_{C_i C_j} = \Phi_{C_i} - \Phi_{C_j} . \tag{2.42}$$

В статье [231] этот метод применяется для анализа связи дезоксирибонуклеиновой кислоты (ДНК) с фактором транскрипции и гистоновыми метками для 40 000 синтетических последовательностей ДНК.

2.6. ПРОЧИЕ МЕТОДЫ ОБЪЯСНИМОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

2.6.1. МЕТОДЫ ВИЗУАЛИЗАЦИИ ГРАФИКОВ

В настоящем разделе описываются методы, которые могут быть применены для объяснения графиков и диаграмм рассеяния.

Стохастическое вложение соседей с *t*-распределением (t-SNE – T-distributed stochastic neighbor embedding) определяется как метод визуализации, который проецирует высокоразмерные данные в двух- или трехмерные пространства с использованием условных вероятностей для представления расстояний между точками данных, поиска сходств между ними.

В статье [232] метод tSNE применяется для анализа и сокращения размерности датасета погодных данных, используемого для прогноза выработки ветряных электростанций, и визуализации параметров функционирования станций. Результаты моделирования показали положительное влияние уменьшения размерности датасета на прогнозирование мощности выработки ветряной электростанции.

В статье [233] tSNE применяется для улучшения качества интеграции справочных данных с данными дистанционного зондирования земли. После удаления 15 % некорректных справочных данных глобальная точность классификации увеличилась примерно на 6 % относительно базовой классификации, полученной с помощью случайного удаления того же количества справочных данных. Эффективность была проверена путем классификации трех неоднородных областей. На рис. 2.16 представлен результат tSNE визуализации в начале и после кластеризации данных.

tSNE может быть реализована в Sklearn [234] на Python.

Метод главных компонент (PCA – Principal Component Analysis) является методом линейного снижения размерности, который применяется при первичном анализе данных, визуализации и предварительной обработке данных. Впервые метод для объяснения глубоких нейронных сетей был применен в [235].

Для входного изображения $r_\theta \in \Omega$ с индексом $\theta = [1, \Theta]$ можно получить высокоразмерное представление изображения, выдаваемое глубокой нейронной сетью. После его центрирования и получе-

ния $F^L(r_\theta)$ вычисляют собственные значения ковариационной матрицы:

$$\frac{1}{\Theta}\sum_{\theta=1}^{\Theta}(F^L(\theta))(F^L)(\theta)^{\mathrm{T}}. \tag{2.43}$$

а *б*

Рис. 2.16. tSNE визуализация:

а – до кластеризации; *б* – после кластеризации

При заданных параметрах $\Theta = \Theta_1, \ldots, \Theta_N$ и образцовом положении изображения t с параметрами $\theta \in \Theta$ для конкретного фактора k признаки рассчитываются как

$$F_k^L(t) = \frac{|\Theta_k|}{|\Theta|}\sum_{\theta\in\Theta|\theta_k=t} F^L(\Theta). \tag{2.44}$$

Таким образом, получаются проекционные вложения относительно различных факторов изображения (рис. 2.17).

В статье [232] было представлено, что для анализа и размерности погодного датасета PCA показал хуже результат, чем tSNE.

PCA можно реализовать на Python [236] и MATLAB [237].

TreeView является методом, разделяющим пространство признаков на более мелкие подпространства, где каждое подпространство представляет собой отдельный фактор. Сначала входные данные X преобразуются в признаки Y: $T_1 : X \to Y$. Затем Y классифицируются и преобразуются в пространство меток Z: $T_2 : Y \to Z$ [192].

Пространство признаков Y разделяется на K подпространств, которые строятся путем кластеризации похожих нейронов в соответствии с их активациями. Каждый кластер i описывает определенный фактор S_i. Затем из меток кластеров строится новый K-мерный признак, для которого TreeView создает визуализацию.

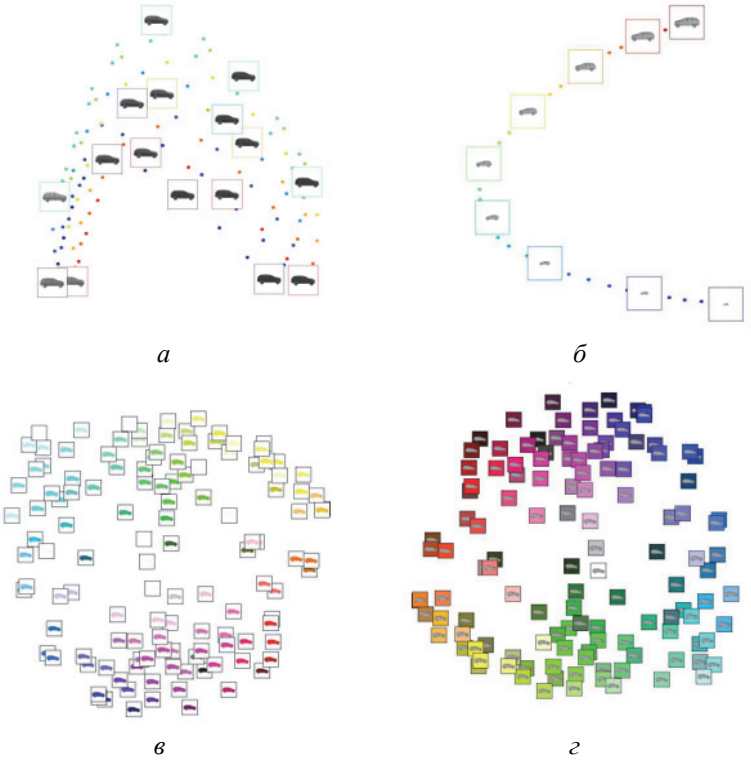


Рис. 2.17. PCA-вложения для различных факторов:
а – освещение; *б* – размер; *в* – цвет объекта; *г* – цвет фона

Для слоя l ответы нейронов обозначаются как $Y_l \in R^{N_l \times T}$, где N_l – количество фильтров; T – количество данных. Y_l кластеризуется на K кластеров с активациями $F_l \in R^{K \times N_l \times T}$ в соответствии со сходством скрытых активаций. Затем из меток кластеров строятся новые интерпретируемые признаки $M \in R^{K \times T}$.

Предсказание метки кластера базируется на обученном классификаторе $P_l^i\ P_l^i$, который обучается с использованием признаков M.

2.6.2. МЕТОДЫ ОБЪЯСНЕНИЯ ТЕКСТОВЫХ ДАННЫХ

Методы объяснения текстовых данных создают текст на естественном языке для интерпретации решения.

Метод оценки значений активации ячеек является методом объяснимости для рекуррентных сетей с долговременной краткосрочной памятью (LSTM – Long Short-Term Memory). Метод использует язык на уровне символов для объяснения долгосрочных зависимостей LSTM. Векторы значений подаются в LSTM на каждом временном шаге и проецируются на последовательности слов с полностью связанными слоями. Значения активации на каждом временном шаге моделируют следующий символ в последовательности и используются для интерпретации модели.

Зрительный вопросно-ответный алгоритм (VQA – Visual Question Answering), предложенный в [239], анализирует области изображения и слова вопроса для генерации ответа. Эти слова $Q = (q_1, q_2, ..., q_T)$ проецируются с помощью слоя вложений в пространство с меньшей размерностью для получения векторных представлений слов $Q_\varpi = (q_{1\varpi}, q_{2\varpi}, ..., q_{T\varpi})$. Для моделирования отношений между соседними словами к векторным представлениям слов применяется одномерный сверточный слой.

После этого модуль LSTM моделирует скрытое представление вопроса q_t^h для каждого временного шага t:

$$q_t^h = \max\left(\text{LSTM}\left(\text{conv}\left(q_{t:t+s}^h\right)\right)\right), \tag{2.45}$$

где s – поле приема одномерной свертки.

Механизм совместного внимания принимает на вход изображение $V = (v_1, v_2, ..., v_N)$ и представления слов на каждом иерархическом уровне $r \in \{\varpi, p, s\}$, где ϖ – слово; p – фраза; s – предложение. На основании V и r происходит генерация обращенных к вниманию изображений v_{att}^r и вопросов q_{att}^r. Окончательное предсказание ответа

основывается на всех этих изображениях и вопросах, которые моделируются с помощью полносвязных слоев.

$$h^{\varpi} = \mathrm{th}\left(W_{\varpi}\left(q_{\mathrm{att}}^{\varpi} + v_{\mathrm{att}}^{\varpi}\right)\right); \tag{2.46}$$

$$h^{p} = \mathrm{th}\left(W_{p}\left(q_{\mathrm{att}}^{p} + v_{\mathrm{att}}^{p}\right),\ h^{\varpi}\right); \tag{2.47}$$

$$h^{s} = \mathrm{th}\left(W_{s}\left(q_{\mathrm{att}}^{s} + v_{\mathrm{att}}^{s}\right),\ h^{p}\right); \tag{2.48}$$

$$p = \mathrm{softmax}\left(W_{h}h^{s}\right), \tag{2.49}$$

где W_{ϖ}, W_{p}, W_{s}, W_{h} – веса полносвязных слоев.

ИТОГИ

В настоящем разделе проведен анализ возможности применения методов объяснимого искусственного интеллекта. Можно сделать вывод, что только при применении методов объяснимого искусственного интеллекта возможно добиться доверия пользователей к интеллектуальным системам и обеспечить надежность их работы.

Одна из самых распространенных классификаций методов объяснимого искусственного интеллекта – по стратегии интерпретации. По этому типу методы подразделяются на интерпретируемые и апостериорные. Апостериорные, в свою очередь, разделяются на локальные и глобальные. В данном обзоре в основном были рассмотрены методы локальной интерпретации, которые позволяют дать объяснение только для конкретного экземпляра входных данных. Рассмотрен алгоритм аддитивного объяснения Шепли, который может также давать глобальные объяснения на основании анализа локальных интерпретаций, и специализированные методы объяснения результатов глубоких нейросетевых моделей.

Необходимо отметить, что в настоящее время ни один из методов не удовлетворяет принципам XAI: объяснимости, значимости, точности и работе в пределах знаний.

3. РАЗРАБОТКА СИСТЕМ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ

3.1. ОСНОВНЫЕ ПОНЯТИЯ В ОБЛАСТИ ПРИНЯТИЯ РЕШЕНИЙ

Актуальность исследования методов поддержки принятия решений обусловлена ролью, которую решения играют в организации и регулировании человеческой деятельности, преодолении проблем и достижении целей. Решение определяет направление движения и позволяет организации или человеку перейти от текущего состояния к желаемому. Оно может быть результатом анализа данных, оценки рисков и прогнозирования последствий. Решение влияет на эффективность и результативность деятельности, а также может определять успех или неудачу в достижении поставленных целей. Правильно принятое решение может привести к значительным выгодам и улучшению ситуации, в то время как неправильное решение может привести к потерям и негативным последствиям.

В широком понимании решение представляет собой выбор определенного действия или курса действий из множества возможных альтернатив с целью достижения желаемого результата или решения проблемы, включая анализ ситуации, оценку альтернатив, выбор наиболее предпочтительного варианта и его последующую реализацию. Этот выбор осуществляется по определенным критериям, позволяющим оценивать альтернативы с точки зрения одной или нескольких целей достаточно многозначно, и его смысл определяется в соответствии с конкретным направлением исследования. В современной научной литературе решение трактуется как процесс, выбор и как результат выбора. Решение как процесс протекает во времени и осуществляется в несколько этапов: этап подготовки, принятия и реализации решения. Этап принятия

решения представляет акт выбора, осуществляемый ЛПР, которое действует в соответствии с определенными правилами. Решение как результат выбора является предписанием к действию (планом работы, вариантом, проектом).

В [240] автор рассматривает различия в психологическом и экономическом подходе к определению понятия «решение». Экономика исследует преимущественно внешние условия формирования и реализации решений, а психология – внутренние. В психологии [241] решение – это «формирование мыслительных операций, снижающих исходную неопределенность проблемной ситуации», либо [242] «один из необходимых моментов волевого действия и способов его выполнения, предполагающего предварительное осознание целей и средств действия, мысленное совершение действия, предшествующее фактическому действию, мысленное обсуждение оснований, говорящих за или против его выполнения и т. п. Данный процесс заканчивается принятием решения». С этой точки зрения понятие «решение» синонимично понятию «принятие решения». В [243] принятие решения определятся как «рациональное предпочтение одной альтернативы из некоторого набора возможных направлений действий».

В экономике решение рассматривается как результат некоторого процесса: решение – «результат труда, получаемый в процессе переработки информации»; «формально зафиксированный проект какого-либо изменения» [244]. В менеджменте решение – это «выбор альтернативы из множества вариантов действий для достижения поставленной цели» [245].

В иностранной литературе в области принятия решения встречается два термина [246]: «decision making» (принятие решения, решаться, совершать выбор) и «problem solving» (решение проблемы (задачи)).

Понятие «принятие решения» рассматривают в узком и широком смысле. В узком смысле – это заключительный акт деятельности по выявлению и анализу различных вариантов, направленный на выбор и утверждение лучшего варианта решения. В широком смысле принятие решения – это процесс, протекающий во времени, осуществляемый в несколько этапов.

Принятие решения – это выбор одного курса действия, одной альтернативы из ряда имеющихся. Если нет альтернатив, то нет выбора, следовательно, нет и решения.

В результате анализа рассмотренных определений можно сделать вывод, что феномен принятия решения обладает следующими характеристиками:

1) наличием цели у субъекта любого решения (лица, принимающего решение). Это означает, что у ЛПР есть ясное представление о том, чего необходимо достичь или какую проблему решить. Цель является конечным результатом или желаемым состоянием, которое лицо стремится достичь через свои действия и решения;

2) наличием минимум двух вариантов решения. Необходимо рассмотреть и оценить несколько альтернативных путей или подходов для достижения поставленной цели или решения проблемы. Это позволяет ЛПР быть осознанным, информированным и гибким в своих действиях, что в конечном итоге приводит к более успешному и эффективному результату;

3) наличием критериев, по которым сопоставляются варианты решения. Критерии представляют собой параметры или факторы, которые помогают определить, какой вариант решения является наиболее эффективным. Наличие критериев позволяет систематизировать и структурировать процесс принятия решения, а также обеспечивает объективность и обоснованность выбора.

Задачи принятия решений имеют следующие характерные особенности:

1) многоцелевой характер. Часто при выборе определенного курса действий необходимо учитывать и сбалансировать несколько различных, иногда противоречивых целей. В таких ситуациях решение должно удовлетворять не только одной главной цели, но и ряду других, которые могут включать экономические, социальные, экологические и другие факторы. Многоцелевое принятие решений требует комплексного подхода и анализа, поскольку каждая цель может иметь свои приоритеты, ограничения и взаимосвязи с другими целями. Важно определить вес каждой цели и найти оптимальное решение, которое обеспечит наилучшее достижение всех целей в совокупности;

2) воздействие фактора времени. Не всегда можно сразу наблюдать последствия принятого решения. Часто трудно бывает указать конкретный промежуток времени, в течение которого можно наблюдать то или иное последствие;

3) неформализуемые понятия. Такие понятия, как интуиция, эмоции, ценности, субъективные оценки, политические действия, являются

примерами очень важных неформализуемых понятий, которые существенно усложняют задачу;

4) неопределенность. В момент принятия решения неизвестны последствия каждой из альтернатив;

5) возможности получения информации. Чем больше информации доступно, тем более полное представление о ситуации можно получить. Полнота информации позволяет более точно оценить последствия принятия решений и выбрать наиболее оптимальный вариант. Часто удается получить некоторую информацию, помогающую решить, какую из альтернатив следует выбрать. Однако получение такой информации может потребовать больших затрат времени и денег, и к тому же она может быть не вполне достоверной. Решения должны приниматься на основе актуальной информации из надежных и достоверных источников;

6) динамические аспекты процесса принятия решений. После того как решение выработано, может оказаться, что задача не исчерпана до конца, и потребуется принять очередное решение через некоторый промежуток времени. Важно распознать заранее такие динамические аспекты проблемы и увидеть, какие возможности могут открыться в будущем благодаря принятому решению.

Многие важные задачи не обладают всеми перечисленными особенностями, но часто одной оказывается вполне достаточно, чтобы сделать задачу трудноразрешимой.

Для достижения поставленных целей любое решение должно соответствовать следующим требованиям:

1) быть реальным. Это означает, что оно должно быть основано на достижении поставленных целей с учетом реально доступных ресурсов и времени;

2) содержать механизм реализации. Механизм реализации представляет собой набор шагов, процедур и действий, необходимых для достижения цели. Механизм реализации должен быть четко сформулирован и структурирован, включать в себя определение ролей и ответственностей участников процесса, распределение ресурсов, установление сроков и этапов выполнения работ;

3) быть обоснованным, т. е. основываться на тщательном анализе данных, информации и фактов, а также на применении соответствующих научных методов и моделей. Обоснованность решения подразуме-

вает, что оно базируется на систематическом сборе и обработке информации, использовании проверенных подходов для оценки ситуации и выработки оптимального варианта действий. Такое решение гарантирует, что выбранный подход наиболее эффективен и адекватен и соответствует имеющимся данным и требованиям, что повышает вероятность достижения желаемого результата и снижает риски ошибок;

4) быть эффективным. т. е. обеспечивать достижение поставленных целей с минимальными затратами ресурсов и времени. Эффективность решения оценивается по его способности реализовать желаемый результат с наилучшим соотношением затрат и полученной выгоды. Для достижения эффективности важно учитывать все аспекты задачи, включая ограничения и риски, и выбирать такой подход, который обеспечивает максимальную отдачу от вложенных усилий. Эффективное решение позволяет не только достичь поставленных целей, но и повысить общую эффективность деятельности;

5) быть устойчивым по эффективности к возможным ошибкам в определении исходных данных, т. е. даже при наличии неточностей или неточных данных решение должно продолжать работать эффективно и достигать поставленных целей. Важно, чтобы решение было способно адаптироваться к изменяющимся условиям и корректироваться в случае обнаружения ошибок или неточностей в исходных данных. Такая устойчивость решения позволяет минимизировать риски и обеспечить его эффективность даже при наличии непредвиденных ситуаций или ошибок в исходных данных;

6) быть реализуемым, в том числе не содержать положений, которые нарушат исполнение в результате вызываемых им конфликтов. Решение должно быть разработано таким образом, чтобы учитывать интересы и потребности всех заинтересованных сторон, а также возможные противоречия и конфликты, которые могут возникнуть в процессе его реализации;

7) быть гибким, т. е. обладать способностью реагировать на перемены внешних или внутренних факторов, что позволит своевременно оценить изменения в ситуации и принять решение о необходимости разработки нового решения, которое будет соответствовать новым условиям;

8) предусматривать возможность верификации и контроля исполнения. Решение должно быть проверяемым на соответствие поставленным целям, требованиям и ожиданиям, а также подлежащим мониторингу

и отслеживанию процесса его реализации. Верификация позволяет убедиться в правильности алгоритмов, корректности данных и достижении ожидаемых результатов, в то время как контроль исполнения позволяет своевременно выявлять отклонения от плана, принимать корректирующие меры и обеспечивать эффективное выполнение решения. Важно, чтобы решение предусматривало механизмы верификации и контроля исполнения, гарантирующие его правильность, эффективность и достижение поставленных целей. Это позволяет обеспечить качество и надежность решения, а также возможность его корректировки и оптимизации в случае необходимости.

Существует множество видов решений, которые могут быть приняты в зависимости от ситуаций. Рассмотрим некоторые из наиболее распространенных видов решений:

1) оперативные решения принимаются в реальном времени для решения повседневных задач или проблем. Они обычно краткосрочные и направлены на немедленное решение конкретной ситуации;

2) тактические решения принимаются для достижения конкретных целей или задач в рамках более широкой стратегии. Они обычно краткосрочные и направлены на решение проблемы в рамках определенного временного периода;

3) стратегические решения принимаются для определения долгосрочных целей и направления развития организации. Они обычно связаны с изменениями в стратегии, структуре или процессах и могут иметь значительное влияние на будущее;

4) кризисные решения принимаются в условиях неожиданных или экстремальных ситуаций, таких как кризисы, катастрофы или чрезвычайные обстоятельства. Они требуют быстрого и эффективного реагирования для минимизации потенциального ущерба или риска;

5) компромиссные решения принимаются в условиях конфликта или разногласий между различными сторонами или интересами. Они направлены на достижение компромисса или согласия между различными вариантами или требованиями;

6) инновационные решения включают в себя новые идеи, подходы или технологии для решения проблемы или достижения цели.

В зависимости от ситуации и контекста решения могут быть комбинацией этих различных видов или иметь свои уникальные характеристики.

При всем различии подходов к определению принятия решения – это всегда процесс выбора одной из возможных альтернатив развития событий, протекающий во времени и состоящий из нескольких этапов.

1. Идентификация проблемы и цели. Цель – это идеальное представление желаемого результата. Если фактическое состояние не соответствует желаемому, то существует проблема. Проблема всегда связана с определенными условиями, которые обобщенно называют ситуацией. Совокупность проблемы и ситуации образует проблемную ситуацию. Выявление и описание проблемной ситуации дает исходную информацию для постановки задачи принятия решений. На этом этапе определяется проблема, которую необходимо решить, и цель, которую нужно достичь. Этот этап включает анализ ситуации, выявление причин и последствий проблемы и определение желаемого результата.

2. Сбор информации. На этом этапе собирается необходимая информация для принятия решения, включая анализ данных, проведение исследований, консультации с экспертами или получение мнений от заинтересованных сторон.

3. Формирование множества решений. На этом этапе генерируются различные альтернативные (взаимоисключающие) варианты решения проблемы и достижения цели и оценивается их предпочтительность. Предпочтение – это интегральная оценка качества решений, основанная на объективном анализе (знании, опыте, проведении расчетов и экспериментов) и субъективном понимании ценности и эффективности решений.

4. Оценка и сравнение вариантов. На этом этапе проводится оценка и сравнение различных вариантов решения на основе определенных критериев. Анализируются преимущества и недостатки каждого варианта, оцениваются риски и ожидаемые результаты.

5. Принятие решения. На этом этапе выбирается наиболее подходящий вариант решения на основе проведенной оценки и сравнения. Принятое решение должно быть обоснованным и соответствовать решению проблемы и поставленной цели. Обобщенной характеристикой решения служит его эффективность. Эта характеристика включает эффект решения, определяющий степень достижения целей, отнесенный к затратам на их достижение. Решение тем эффективнее, чем больше степень достижения целей и меньше затраты на их реализацию.

6. Реализация решения. Этап реализации принятого решения, т. е. выполнение необходимых действия для его осуществления. Включает

разработку плана действий, выделение ресурсов и координацию действия сотрудников или заинтересованных сторон.

7. Оценка результатов. На этом этапе проводится оценка результатов принятого решения. Включает анализ достигнутых результатов, сравнение с ожиданиями и оценку эффективности решения. Результаты оценки могут быть использованы для корректировки или улучшения решения в будущем.

Стоит отметить, что этапы реализации решения и оценки результатов не относятся к процедуре принятия решения как таковой, однако их включение в общую схему важно с практической точки зрения, поскольку эти этапы замыкают жизненный цикл процесса возникновения, разрешения и исчезновения проблемной ситуации. Все описанные выше этапы помогают обеспечить систематический и обоснованный подход к процессу принятия решений.

Процесс принятия решения условно представлен в виде схемы на рис. 3.1.

Рис. 3.1. Этапы процесса принятия решения

Совершенствование процесса принятия решений можно реализовать с помощью новых технологий и инструментов, обеспечивающих объективный анализ данных с построением модели предполагаемого развития событий и разрешения проблемных ситуаций. Таким инструментом является информационная система в виде поддержки принятия решений с целью сбора, оптимизации, анализа данных, выявления ошибок в настоящем и прогнозирования дальнейшего хода развития событий с учетом гипотетических вариантов. Автоматизированные системы принятия решений включают в себя разнообразные базы данных и знаний, обработать которые самостоятельно человек не способен. В настоящее время такие системы, как и другие достижения искусственного интеллекта, выступают некоторой формой поддержки в работе человека.

3.2. ПОНЯТИЕ И ОСОБЕННОСТИ СИСТЕМ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ

История создания систем поддержки принятия решений начинается в 1960-х годах. В этот период были разработаны первые компьютерные системы, которые использовались для анализа данных и поддержки принятия решений. Тогда же компанией IBM была создана первая система под названием «DSS-1» (Decision Support System 1). Эта система использовала математические модели и статистические методы для анализа данных и принятия решений. В 1970-х годах появились более сложные СППР, которые нашли свое применение в различных отраслях, таких как финансы, производство и управление. В 1980-х годах СППР стали более распространенными и доступными благодаря развитию компьютерных технологий. Были разработаны различные типы СППР, такие как экспертные системы, системы прогнозирования и системы управления знаниями. В 1990-х годах СППР стали интегрироваться с другими информационными системами: системами управления базами данных и системами управления предприятием. Это позволило более эффективно использовать данные и принимать решения на основе целостной информации. С развитием Интернета и технологий веб-приложений в 2000-х годах СППР стали доступными через веб-интерфейсы, у пользователей появилась возможность получать доступ к системам принятия решений из любого места и в любое время. К основным предпосылкам появления СППП относят:

1) увеличение сложности принимаемых решений. Современные организации сталкиваются со всё более сложными и многогранными

проблемами, требующими анализа большого объема данных и учета множества факторов. Традиционные методы принятия решений становятся недостаточными для эффективного управления в подобных ситуациях;

2) рост объема данных. С развитием информационных технологий и с доступностью больших объемов данных организации имеют возможность собирать и анализировать значительное количество информации для принятия решений;

3) улучшение аналитических возможностей. С развитием алгоритмов машинного обучения, искусственного интеллекта и аналитических методов стало возможным более точно и эффективно анализировать данные и делать прогнозы. СППР используют эти аналитические возможности для поддержки принятия решений;

4) развитие конкуренции. Это привело к уменьшению времени на принятие решения, выросла ответственность за его качество. В таких условиях средства, затраченные на анализ информации и принятие более качественного решения, многократно себя оправдывают. Информационные системы – один из важнейших факторов выживания социально-экономического объекта в жесткой конкурентной борьбе;

5) улучшение коммуникации и сотрудничества. СППР обеспечивают возможность совместной работы и обмена информацией между различными участниками процесса принятия решений. Это способствует более эффективному сотрудничеству и принятию обоснованных решений.

СППР продолжают развиваться и становятся все более интеллектуальными, используя методы искусственного интеллекта, машинного обучения и анализа данных для поддержки принятия решений. Сегодня СППР применяются в различных областях, включая бизнес, финансы, здравоохранение, производство и другие. Объем мирового рынка интеллектуальных решений, в том числе в области систем поддержки принятия решений, в 2022 году составил 10,55 млрд долларов США, и ожидается, что к 2032 году он достигнет примерно 45,15 млрд долларов США, увеличившись в среднем на 15,7 % в течение прогнозируемого периода с 2023 по 2032 год. Росту этого рынка способствует необходимость принятия решений, ориентированных на данные, в современной сложной и конкурентной бизнес-среде.

Одной из важных тенденций, преобладающих на рынке интеллектуальных решений, является активное внедрение методологий принятия

решений, ориентированных на данные. Совершенствование алгоритмов машинного обучения и обработки естественного языка расширяет возможности систем интеллектуального анализа решений, делая их более эффективными и точными. Развитию отрасли интеллектуального анализа решений способствует несколько факторов. Во-первых, необходимость для предприятий оставаться гибкими и адаптируемыми в условиях быстро меняющейся динамики рынка содействует внедрению СППР. Во-вторых, резкий рост объема больших данных и усложнение процессов принятия решений стимулируют высокий спрос на инструменты анализа данных. Наконец, пандемия COVID-19 подчеркнула важное значение принятия решений на основе данных для управления кризисами и обеспечения готовности к ним, что еще больше ускорило рост рынка.

Несмотря на многообещающие перспективы, рынок интеллектуальных решений сталкивается с определенными проблемами. Основное препятствие связано со сложностью интеграции этих технологий в существующую инфраструктуру, что требует значительных затрат времени и ресурсов. Опасения по поводу конфиденциальности и безопасности данных остаются актуальными, и организациям приходится тщательно изучать нормативно-правовую базу. Кроме того, препятствием служит нехватка квалифицированных специалистов, владеющих технологиями анализа решений.

Системы поддержки принятия решений – это класс информационных систем, которые обеспечивают руководителей различных уровней знаниями и информацией, позволяющими принимать обоснованные управленческие решения в различных сферах деятельности.

3.3. АРХИТЕКТУРА ИНФОРМАЦИОННОЙ СИСТЕМЫ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ ПРИ ПРОГНОЗИРОВАНИИ ГЕНЕРАЦИИ ФОТОЭЛЕКТРИЧЕСКОЙ СТАНЦИИ

Разрабатываемая информационная система предназначена для прогнозирования генерации электроэнергии фотоэлектрической станции на основе анализа метеорологических данных, данных о солнечной радиации и других релевантных факторов. Система использует методы объяснимого искусственного интеллекта для обеспечения прозрачности и понятности процессов принятия решений, что существенно повышает уровень коммуникации с пользователями (рис. 3.2).

Важной частью системы является база данных системы (рис. 3.3).

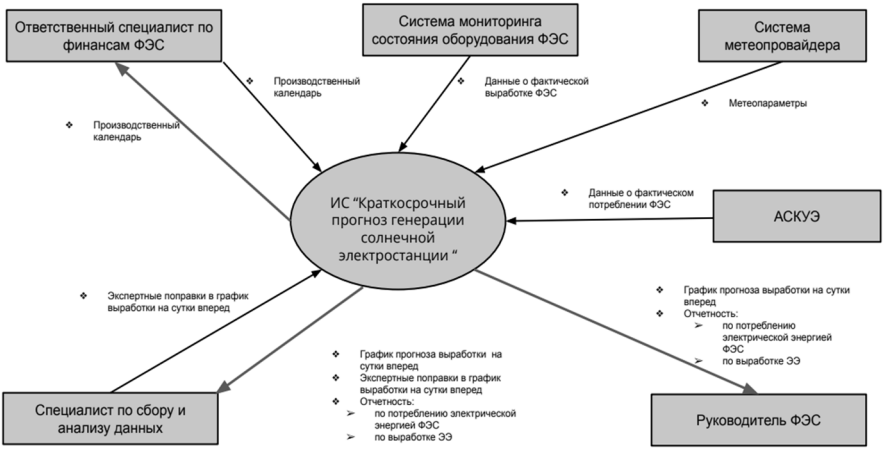


Рис. 3.2. Контекстная диаграмма информационной системы

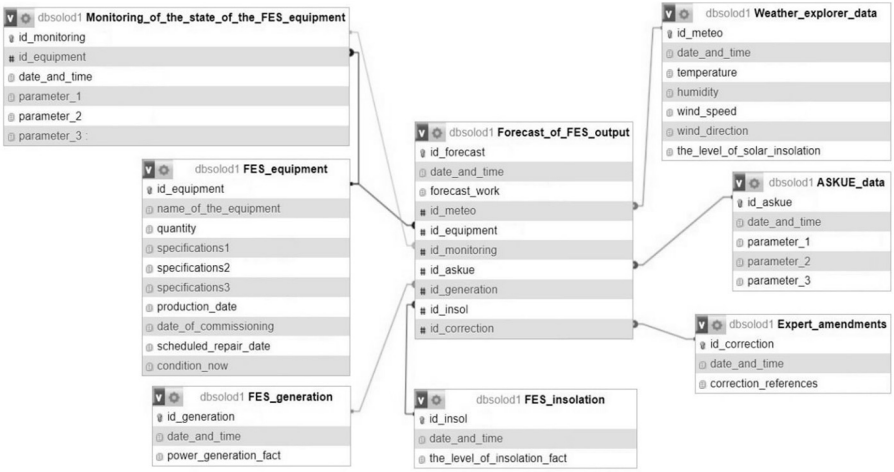


Рис. 3.3. Модель базы данных системы

Таблица Monitoring_of_the_state_of_the_FES_equipment содержит данные о состоянии осинового оборудовании станции:

- id_monitoring – идентификатор записи с результатами мониторинга (первичный ключ);

- id_equipment – идентификатор оборудования (внешний ключ);
- data_and_time – момент снятия контролируемых параметров;
- parameter_1, parameter _2, … – значения параметров (их описание выходит за рамки данной работы).

Таблица FES_equipment содержит данные об основном электрооборудовании станции:

- id_equipment – идентификатор оборудования (первичный ключ);
- name_of_the_equipment – диспетчерское имя оборудования;
- quantity – количество аналогичных единиц оборудования;
- specifications_1, specifications_2, … – паспортные данные (их описание выходит за рамки данной работы);
- production_data – дата изготовления;
- date_of_commissioning – дата ввода в эксплуатацию;
- scheduling_repair_data – дата ближайшего планового ремонта;
- condition_now – индекс фактического технического состояния.

Таблица FES_generation содержит ретроспективные данные о генерации электроэнергии:

- id_generation – идентификатор отдельной генерирующей энергоустановки (первичный ключ);
- date_and_time – дата и время записи о выработанной электроэнергии (с шагом 1 час);
- power_generation_fact – выработка электроэнергии за час (кВт).

Таблица Forecast_of_FES_output содержит данные о прогнозах выработки:

- id_forecast – идентификатор записи (первичный ключ);
- date_and_time – дата и время записи;
- id_meteo – идентификатор набора метеоданных, которые использовались для построения прогноза (внешний ключ);
- id_generation – идентификатор отдельной генерирующей энергоустановки (внешний ключ);
- id_insol – идентификатор записи об инсоляции, которая использовалась для построения прогноза (внешний ключ).

Таблица FES_insolation содержит ретроспективные данные о инсоляции:

- id_insol – идентификатор записи (первичный ключ);
- date_and_time – дата и время записи (с шагом 1 час);
- the_level_of_insolation _fact – значение инсоляции.

Таблица FESWeather_explorer_data содержит ретроспективные метеоданные:

- id_meteo – идентификатор записи (первичный ключ);
- date_and_time – дата и время записи (с шагом 1 час);
- далее идут метеорологические параметры.

Информационная система включает в себя следующие основные компоненты.

Модуль сбора данных

Задачи:

- сбор данных из различных источников: метеостанций, метеопровайдеров, SCADA-система станции;
- интеграция с API для автоматического обновления данных.

Модуль предварительной обработки данных

Задачи:

- очистка данных от шумов и аномалий;
- нормализация и стандартизация данных для подготовки к анализу;
- формирование признаков, включая временные ряды и статистические характеристики.

Модуль моделирования

Задачи:

- реализация алгоритмов машинного обучения для прогнозирования генерации электроэнергии;
- использование методов XAI для объяснения предсказаний модели на основе аддитивного объяснение Шепли;
- обучение и тестирование моделей на исторических данных.

Модуль принятия решений

Задачи:

- интеграция прогнозов с текущими операционными данными фотоэлектрической станции для поддержки принятия решений по управлению и оптимизации работы станции;
- предоставление рекомендаций операторам станции на основе анализа и прогнозов.

Модуль визуализации и интерфейса пользователя

Задачи:

- предоставление графического интерфейса для визуализации данных, прогнозов и объяснений модели;
- предоставление инструментов для интерактивного анализа данных и прогнозов;
- предоставление отчетов и дашбордов для мониторинга производительности станции.

База данных

Задачи:

- хранение исходных данных, промежуточных результатов обработки и моделей машинного обучения;
- управление доступом и запросами к базе данных.

Каждый модуль взаимодействует с другими через определенные API и протоколы. Например, модуль сбора данных передает данные в модуль предварительной обработки, который в свою очередь подготовляет их для модуля моделирования. Модуль моделирования использует обработанные данные для обучения и тестирования моделей, результаты которых передаются в модуль принятия решений. Этот модуль анализирует прогнозы и текущие операционные данные для формирования рекомендаций, которые отображаются в модуле визуализации и интерфейса пользователя.

Каждый модуль реализуется как отдельный микросервис, что позволяет легко масштабировать и обновлять компоненты независимо друг от друга.

3.4. ВЫБОР ТЕХНОЛОГИЙ РЕАЛИЗАЦИИ ИНФОРМАЦИОННОЙ СИСТЕМЫ

1. Язык программирования: Python.
2. Библиотеки машинного обучения: TensorFlow, Keras, Scikit-learn.
3. Система управления базой данных: PostgreSQL (для хранения исторических данных и результатов моделирования).
4. Веб-фреймворки для разработки пользовательского интерфейса: Flask.
5. Инструменты визуализации: Matplotlib, Seaborn, Bokeh.

Python является оптимальным выбором для разработки информационной системы поддержки принятия решений при прогнозировании генерации фотоэлектрической станции по совокупности факторов.

Python имеет богатую экосистему библиотек и фреймворков для работы с данными, машинного обучения, веб-разработки и других задач, связанных с разработкой ИС. Примеры таких библиотек включают TensorFlow, PyTorch, Scikit-learn, Keras для машинного обучения; Django и Flask для веб-разработки; Pandas и NumPy для работы с данными.

Python имеет простой и понятный синтаксис, что делает его легким для изучения и использования. Это особенно важно для команды разработчиков с разным уровнем опыта.

Хотя Python – не самый быстрый язык программирования, его производительность достаточно высока для большинства задач. Более того, существуют инструменты, такие как Cython и Numba, которые позволяют компилировать Python-код в более быстрый машинный код.

Существует развитое сообщество разработчиков, которые активно поддерживают и развивают язык и его библиотеки, что обеспечивает доступ к большому количеству ресурсов, документации и примеров кода.

Кроме того, программное обеспечение, написанное на языке Python, может легко интегрироваться с библиотеками на таких языках программирования, как C, C++ и Java.

3.5. ОПИСАНИЕ ВЫБОРКИ ДАННЫХ МЕТОДА КРАТКОСРОЧНОГО ПРОГНОЗИРОВАНИЯ ГЕНЕРАЦИИ ФОТОЭЛЕКТРИЧЕСКОЙ СТАНЦИИ И ИНТЕРПРЕТАЦИИ ПРОГНОЗОВ

В качестве объекта исследования выбрана одна из ФЭС, расположенных в Республике Алтай. Республика Алтай благодаря своему природному и климатическому потенциалу обладает значительными возможностями для развития возобновляемых источников энергии. Территория республики получает большое количество солнечного излучения, особенно в летний период. В Республике Алтай 300 из 365 дней – солнечные.

Исходная выборка данных содержит результаты измерений выработки ФЭС и солнечной инсоляции в период с 06.05.2020 по 23.02.2022.

К ней необходимо было добавить метеорологические факторы. Архив доступных метеорологических наблюдений на ближайшей метеорологической станции имеет дискретность по времени три часа, поэтому рассматривается задача прогнозирования на три часа вперед.

Из выборки данных были исключены значения, когда выработка равна нулю, поскольку такие записи не информативны для прогнозирования.

В качестве целевой переменной выбрана суммарная инсоляция, приходящаяся на единицу площади солнечной панели, поскольку по ее значению можно вычислить выработку в зависимости от оборудования ФЭС.

Полный набор метеорологических факторов приведен в табл. 3.1 (использованы обозначения из источника данных).

Корреляционный анализ позволил исключить часть малоинформативных метеорологических признаков: P_0, P, P_a, U, D, V, V_{max}, L, T_d, R, T_r, E, T_g, E', S.

Таблица 3.1

Метеорологические параметры исходной выборки

Параметр	Описание параметра	Единица измерения / формат
Местное время	Местное время	дд.ММ.гг чч:мм
T	Температура воздуха на высоте 2 м над поверхностью земли (и далее все высоты от поверхности земли)	°C
P_0	Атмосферное давление на уровне станции	мм. рт. ст.
P	Атмосферное давление, приведенное к среднему уровню моря	мм. рт. ст.
P_a	Барическая тенденция: изменение атмосферного давления за последние три часа	мм. рт. ст.
U	Относительная влажность на высоте 2 м	%
D	Направление ветра на высоте 10–12 м	Текстовое описание (текст)

Окончание табл. 3.1

Параметр	Описание параметра	Единица измерения / формат
V	Скорость ветра на высоте 10–12 м	м/с
$V_{\max}$	Максимальное значение порыва ветра на высоте 10–12 м	м/с
ClCover	Общая облачность	% + текст
LowLevelCl	Слоисто-кучевые, слоистые, кучевые и кучево-дождевые облака (облака нижнего яруса – до 2 км в средних широтах)	Текст
AmountLowLevCl	Количество наблюдаемых облаков низкого уровня, при их отсутствии – среднего	% + текст
ClCeil	Высота основания самых низких облаков	м + текст
MidLevCl	Высококучевые, высокослоистые, слоисто-дождевые облака (облака среднего яруса – от 2 до 6 км в средних широтах)	Текст
HighLevCl	Перистые, перисто-кучевые и перисто-слоистые облака (верхний ярус облаков – от 6 до 13 км в средних широтах)	Текст
L	Горизонтальная дальность видимости	км
T_d	Температура точки росы на высоте 2 м	°C
R	Количество выпавших осадков	мм
T_r	Период, за который накоплено указанное количество осадков	ч
E	Состояние поверхности почвы без снега или измеримого ледяного покрова	Текст
T_g	Минимальная температура поверхности почвы за ночь	°C
E'	Состояние поверхности почвы со снегом или измеримым ледяным покровом	Текст
S	Высота снежного покрова	см

Данные по облачности в формате естественного языка требуют разработки специализированных алгоритмов обработки [38], поэтому

в настоящем исследовании они не используются. В результате для анализа сформирован следующий набор признаков:

- порядковый номер дня в году (*D*);
- час суток (*H*);
- температура воздуха, °C (*T*);
- общая облачность, % (Cov);
- высота основания самых низких облаков, м (*Cels*);
- количество наблюдаемых облаков низкого уровня, при их отсутствии – среднего, % (*ALow*);
- расчетная инсоляция на границе атмосферы, Вт/м^2 (*CI*).

3.6. АНАЛИЗ ДАННЫХ

Был проведен анализ корреляции признаков (рис. 3.4 и 3.5).

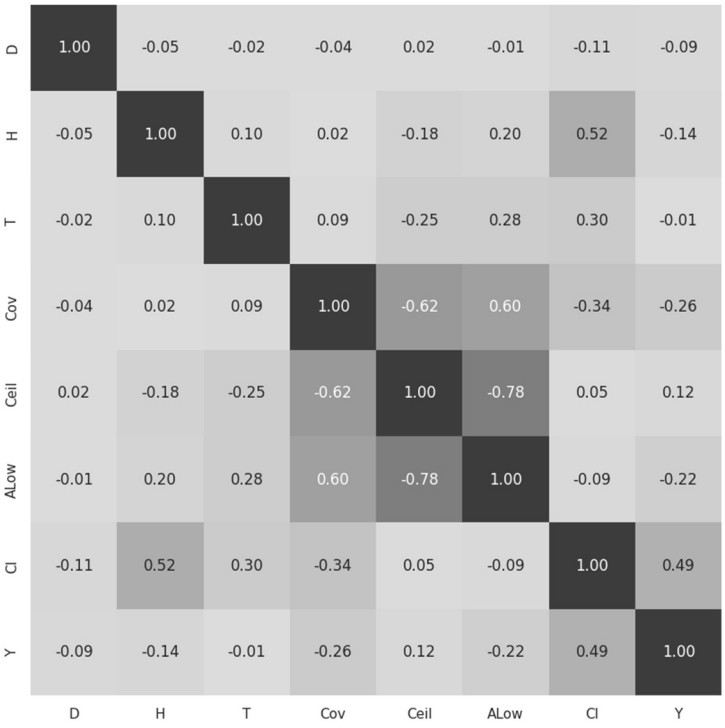


Рис. 3.4. Матрица коэффициентов в корреляции Пирсона

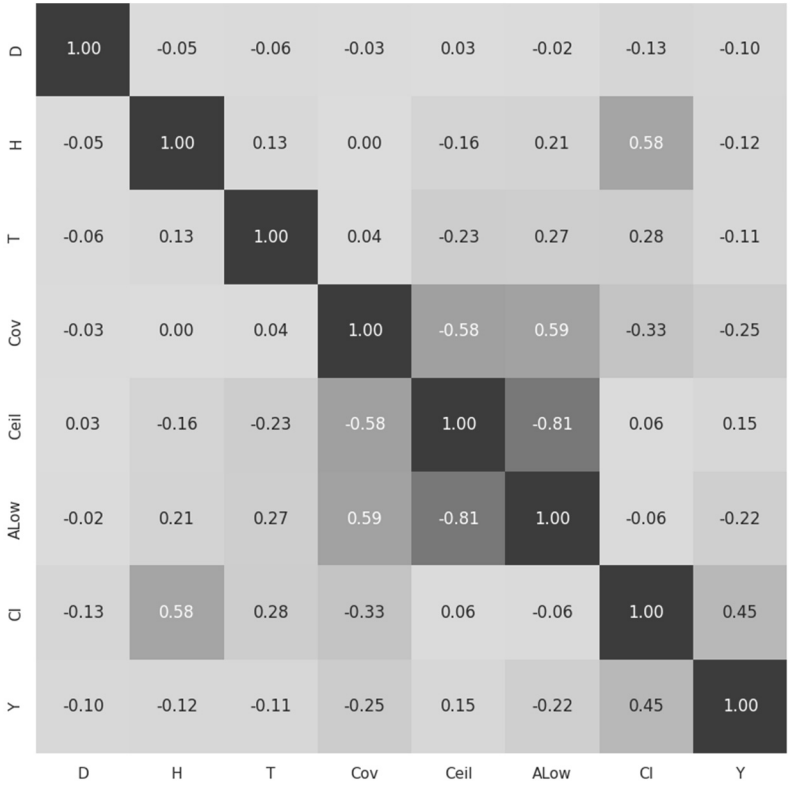


Рис. 3.5. Матрица коэффициентов в корреляции Спирмена

Видно, что целевая переменная, инсоляция (*Y*), наиболее коррелирует с расчетной инсоляцией на границе атмосферы и параметрами облачности.

Среди параметров нет ни одной пары с коэффициентом корреляции по модулю выше 0,85, значит, не следует убирать ни один из признаков.

На рис. 3.6–3.25 показаны результаты визуализации признаков и их зависимостей. На рис. 3.26–3.29 приведены фрагменты целевой переменной (фрагменты временного ряда).

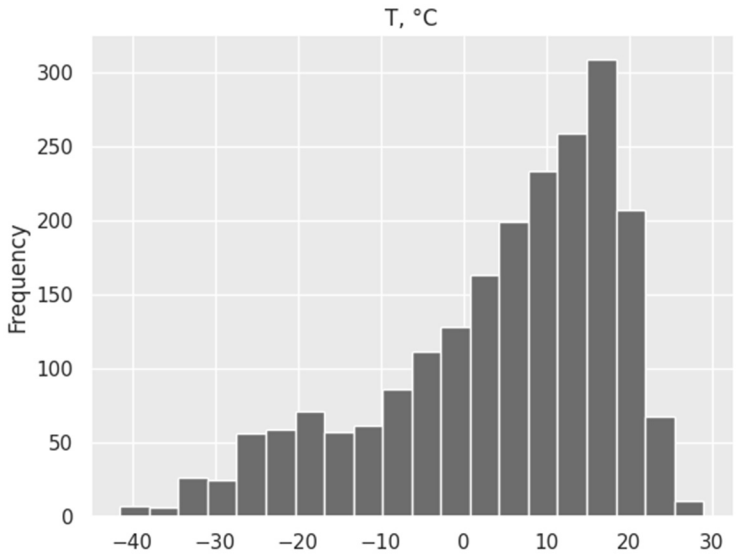


Рис. 3.6. Распределение температуры воздуха

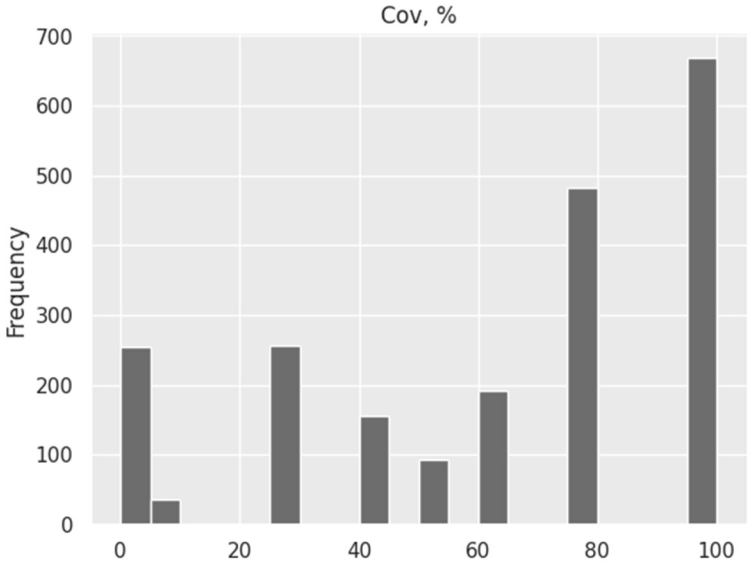


Рис. 3.7. Распределение облачности

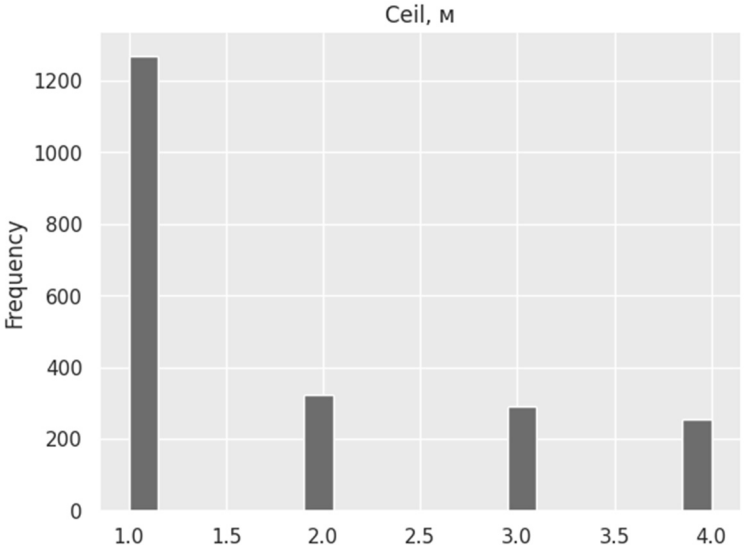


Рис. 3.8. Распределение высоты облаков

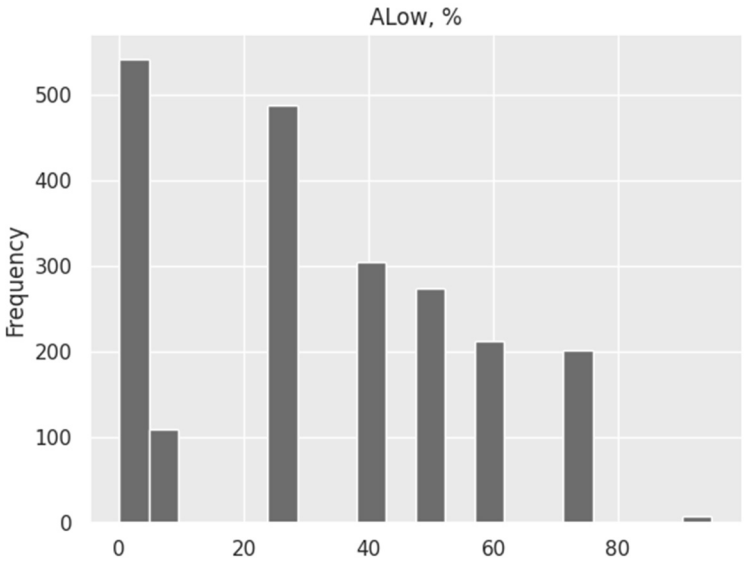


Рис. 3.9. Распределение количества облаков нижнего слоя

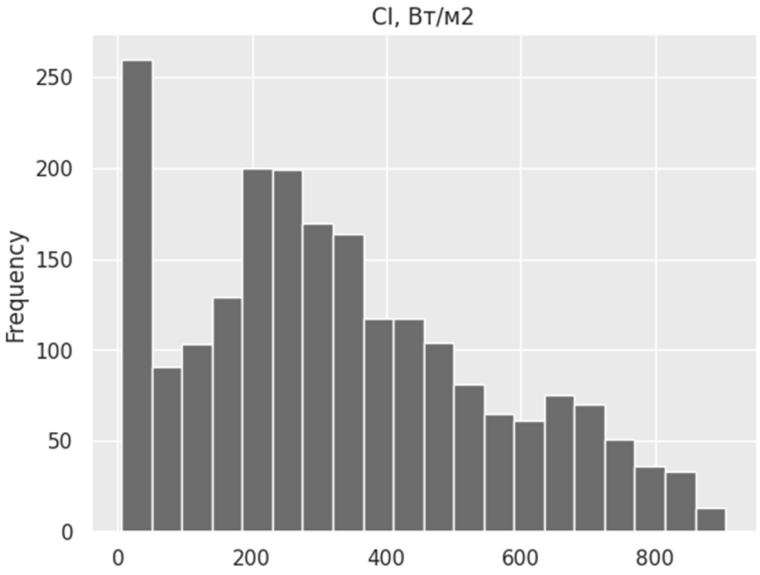


Рис. 3.10. Распределение расчетной инсоляции

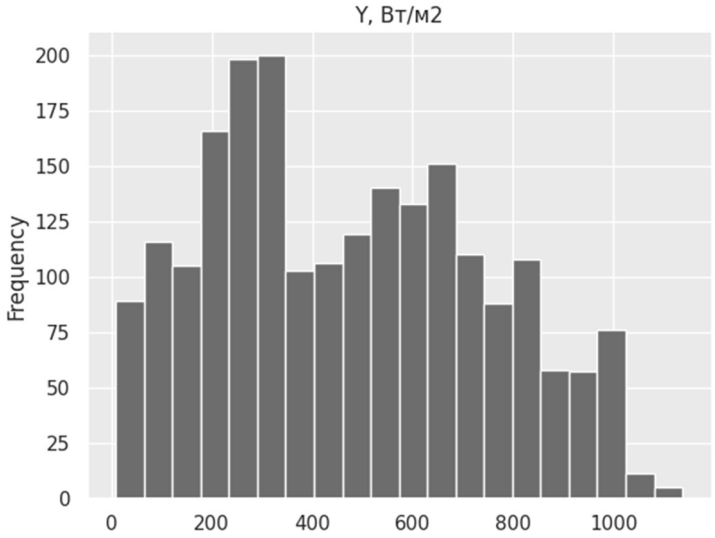


Рис. 3.11. Распределение инсоляции на поверхности Земли

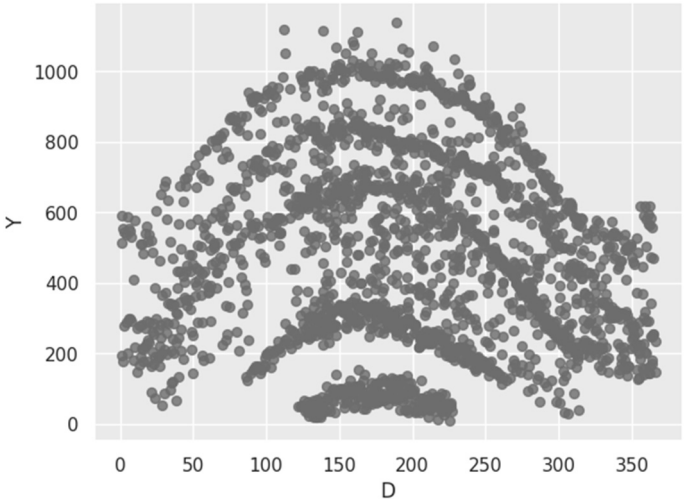


Рис. 3.12. Зависимость инсоляции на поверхности земли от номера дня в году

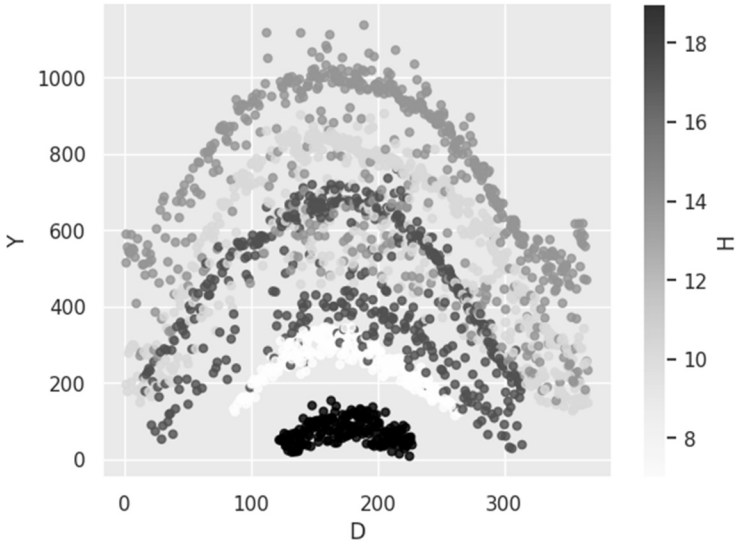


Рис. 3.13. Зависимость инсоляции на поверхности земли от номера дня в году при изменении часа суток

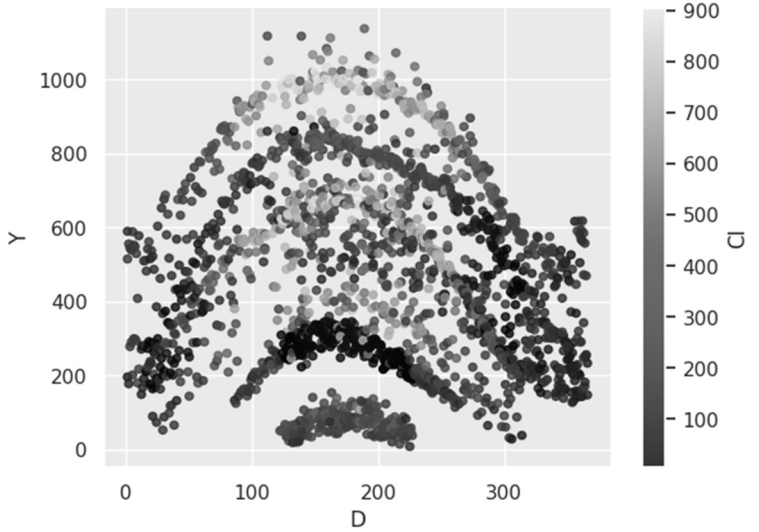


Рис. 3.14. Зависимость инсоляции на поверхности земли от номера дня в году при изменении расчетной инсоляции на границе атмосферы

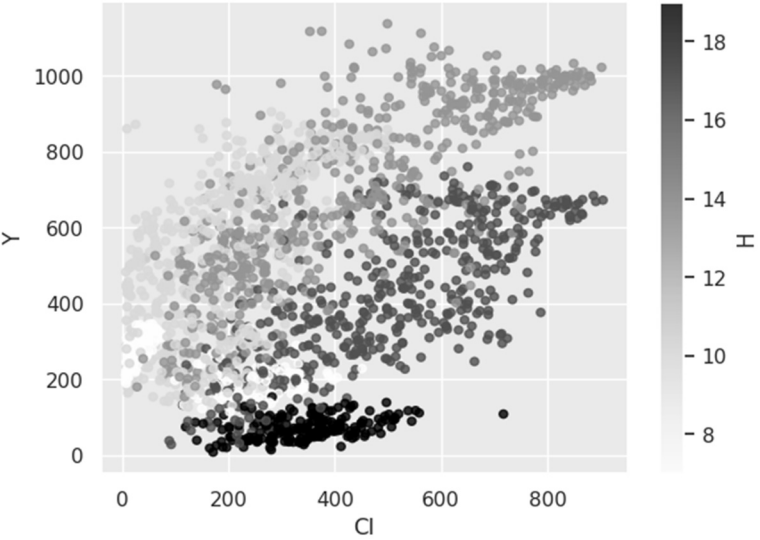


Рис. 3.15. Зависимость инсоляции на поверхности земли от расчетной инсоляции на границе атмосферы при изменении часа суток

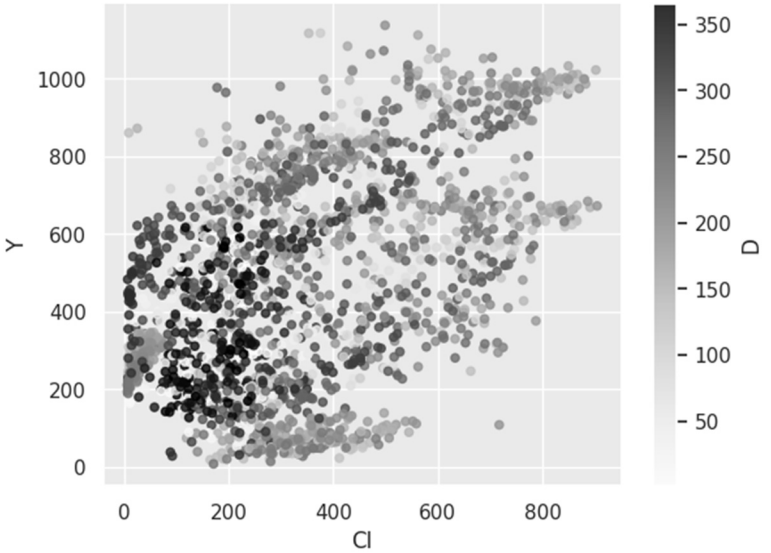


Рис. 3.16. Зависимость инсоляции на поверхности земли от номера дня в году при изменении дня года

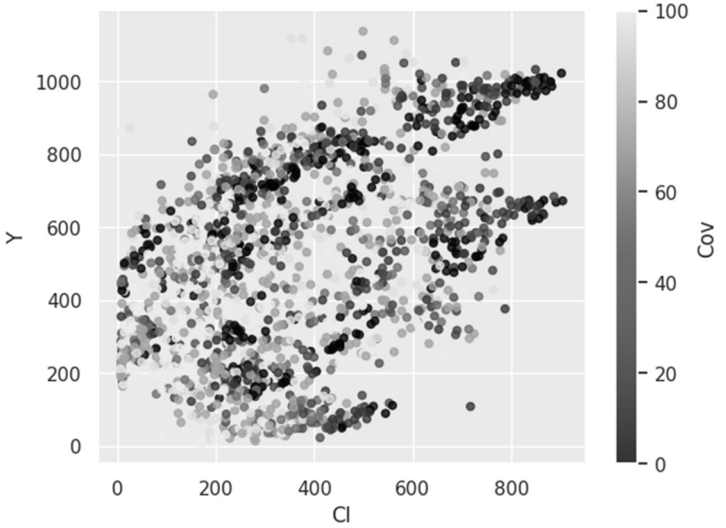


Рис. 3.17. Зависимость инсоляции на поверхности земли от номера дня в году при изменении облачности

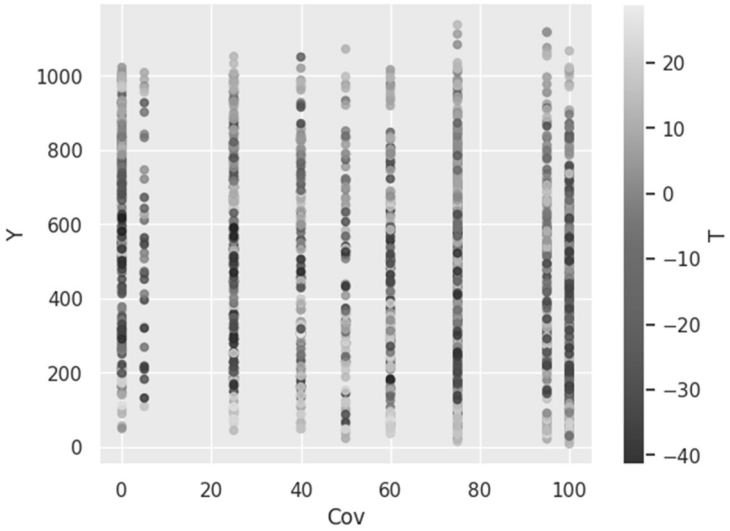


Рис. 3.18. Зависимость инсоляции на поверхности земли от облачности при изменении температуры

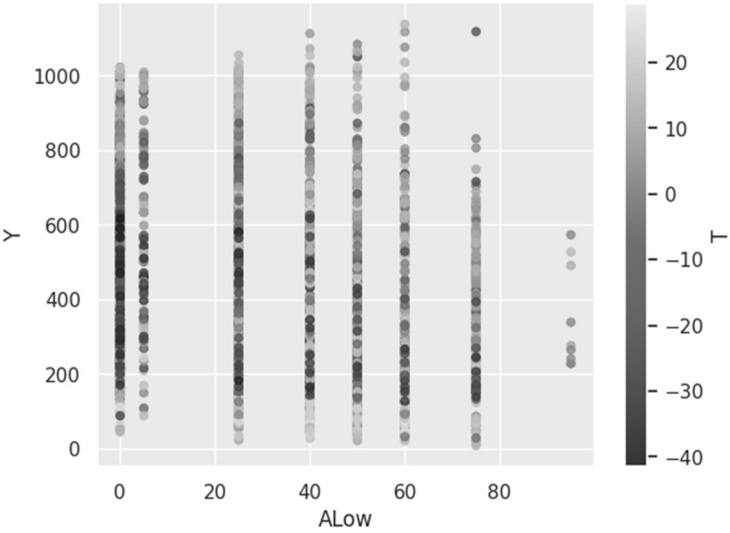


Рис. 3.19. Зависимость инсоляции на поверхности земли от количества облаков нижнего слоя при изменении температуры

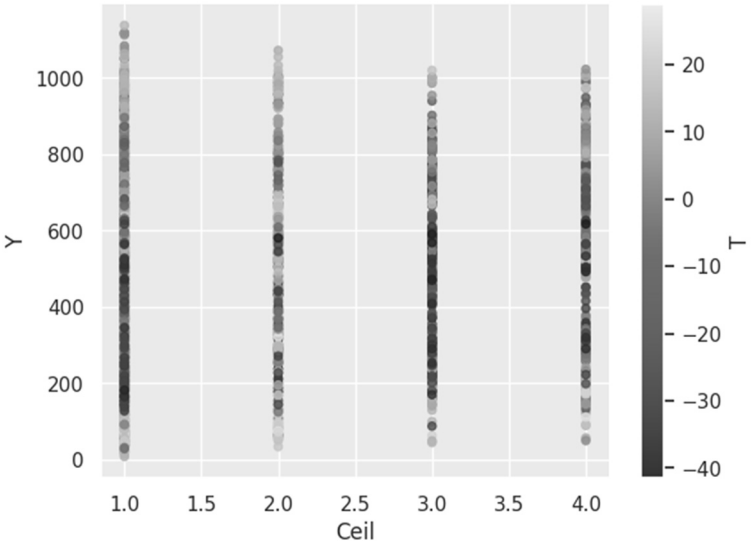


Рис. 3.20. Зависимость инсоляции на поверхности земли от высоты облаков нижнего слоя при изменении температуры

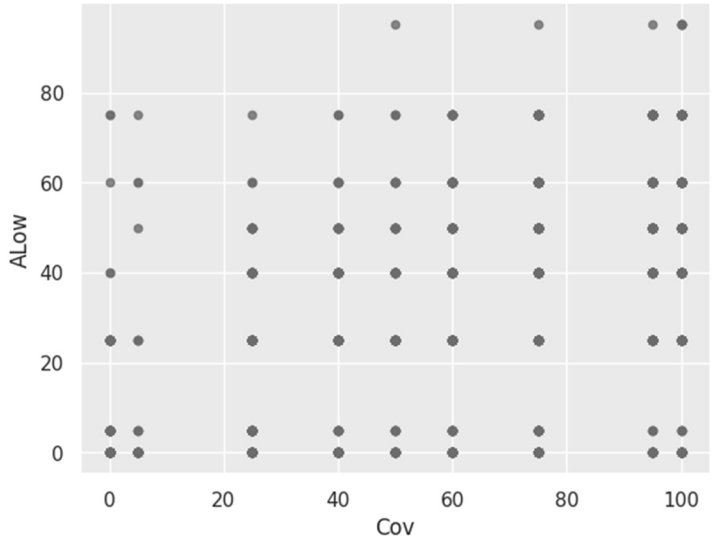


Рис. 3.21. Зависимость количества облаков нижнего слоя от их высоты

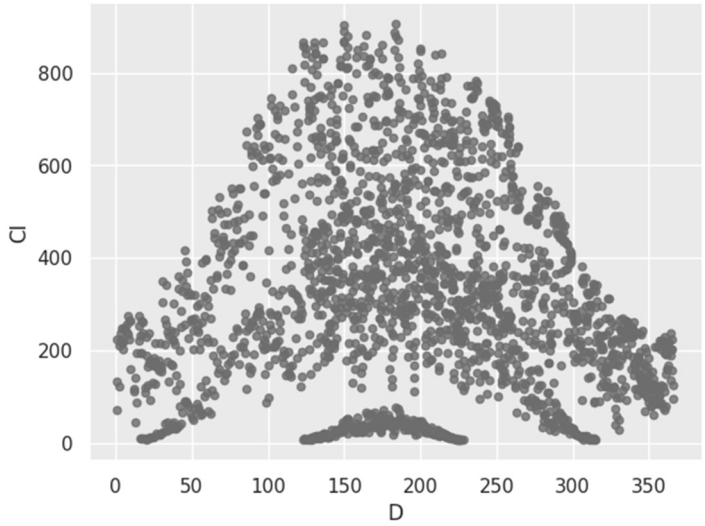


Рис. 3.22. Зависимость инсоляции на поверхности атмосферы от дня года

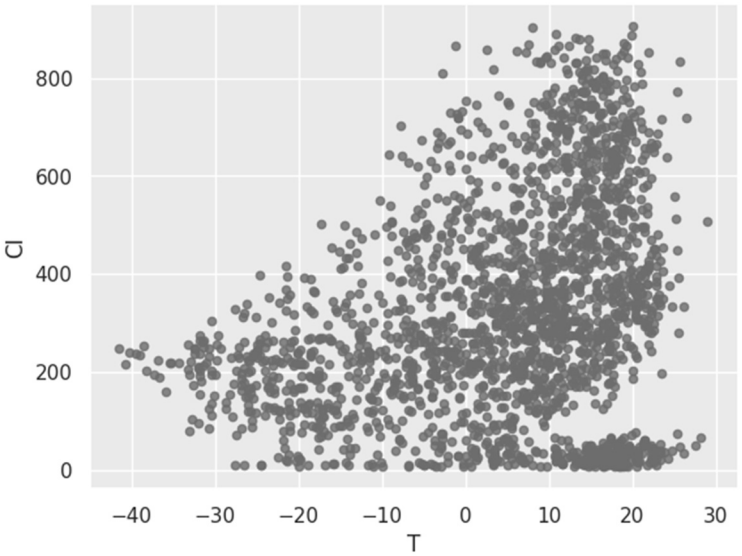


Рис. 3.23. Зависимость температуры от инсоляции на поверхности атмосферы

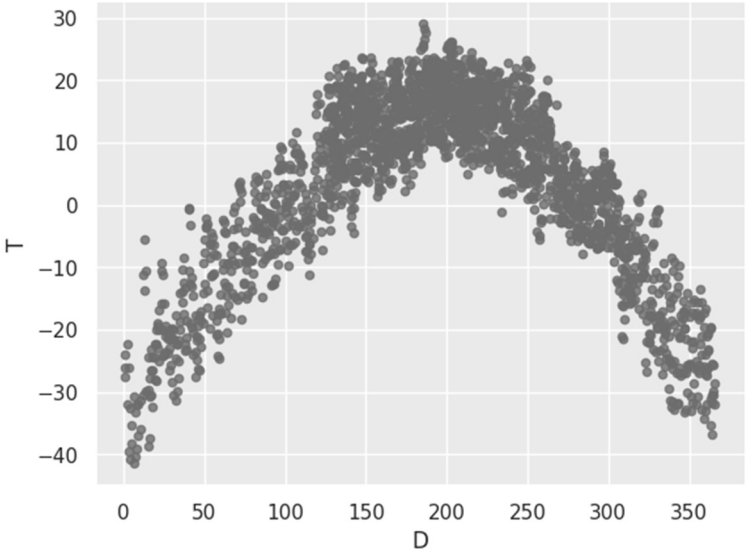


Рис. 3.24. Зависимость температуры от дня года

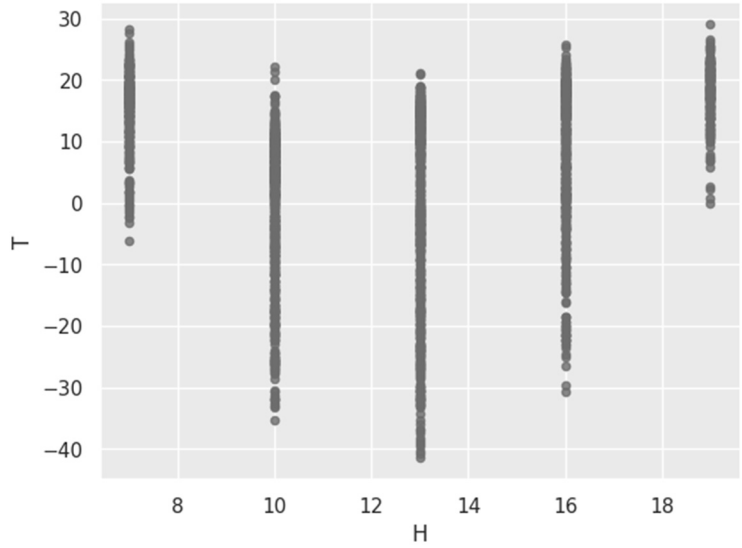


Рис. 3.25. Зависимость температуры от часа суток

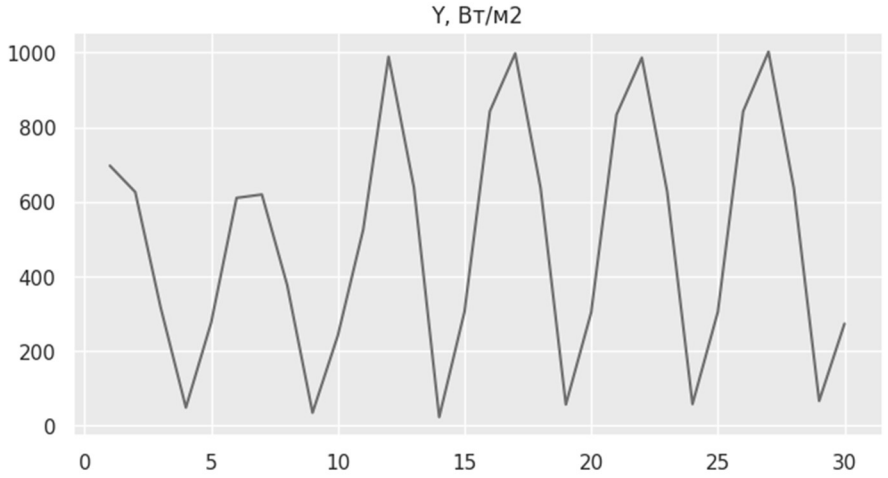


Рис. 3.26. Фрагмент инсоляции на поверхности земли (только дневные часы)

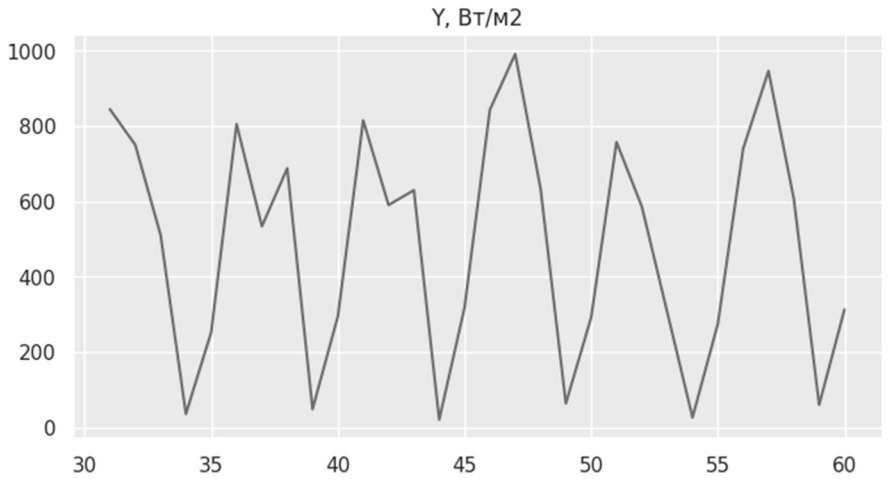


Рис. 3.27. Фрагмент инсоляции на поверхности земли (только дневные часы)

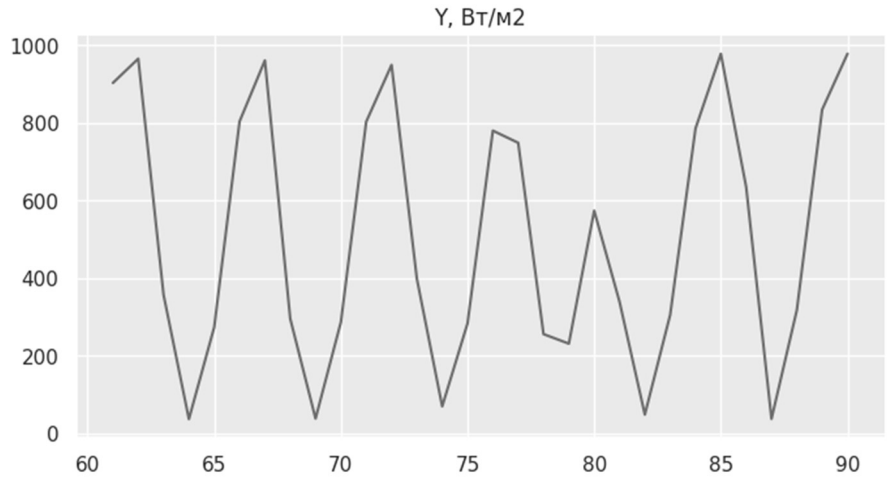


Рис. 3.28. Фрагмент инсоляции на поверхности земли (только дневные часы)

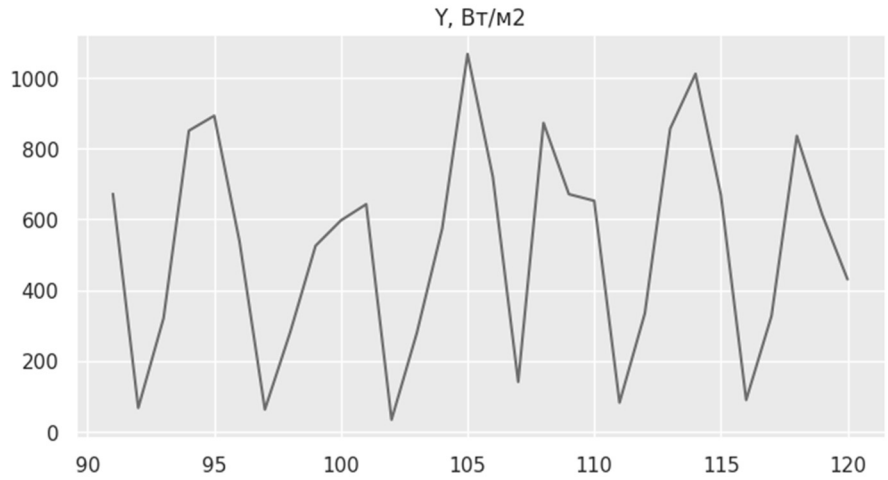


Рис. 3.29. Фрагмент инсоляции на поверхности земли (только дневные часы)

3.7. ПРИМЕНЕНИЕ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ

Проведено исследование моделей машинного обучения:

- линейная регрессия с регуляризацией Тихонова (LR);
- *k*-ближайших соседей (kNN);
- адаптивный бустинг, базовая модель – дерево решений (AB);
- экстремальный градиентный бустинг (XGB);
- случайный лес (RF);
- быстрый бустинг (LB);
- категориальный бустинг (CB).

Выборка была разделена на обучающую и валидационную в отношении 5 к 1.

В табл. 3.2 приведены результаты моделей при использовании различных гиперпараметров.

Таблица 3.2

Результаты моделей

Модель	Гиперпараметры	R^2 на обучающей выборке	R^2 на валидационной выборке
LR	–	0,5	0,492
kNN	Число соседей: 2	0,854	0,810
kNN	Число соседей: 3	0,866	0,812
kNN	Число соседей: 5	0,875	0,809
kNN	Число соседей: 7	0,880	0,793
AB	Максимальная глубина деревьев: 6 Число деревьев: 150	0,900	0,856
AB	Максимальная глубина деревьев: 6 Число деревьев: 50	0,897	0,852
AB	Максимальная глубина деревьев: 8 Число деревьев: 50	0,955	0,874
AB	Максимальная глубина деревьев: 8 Число деревьев: 100	0,955	0,873
AB	Максимальная глубина деревьев: 10 Число деревьев: 50	0,988	0,860

Продолжение табл. 3.2

Модель	Гиперпараметры	R^2 на обучающей выборке	R^2 на валидационной выборке
AB	Максимальная глубина деревьев: 9 Число деревьев: 50	0,977	0,862
XGB	Максимальная глубина деревьев: 3 Число деревьев: 150	0,943	0,878
XGB	Максимальная глубина деревьев: 3 Число деревьев: 50	0,909	0,876
XGB	Максимальная глубина деревьев: 5 Число деревьев: 50	0,956	0,880
XGB	Максимальная глубина деревьев: 7 Число деревьев: 50	0,989	0,871
XGB	Максимальная глубина деревьев: 5 Число деревьев: 100	0,998	0,866
RF	Максимальная глубина деревьев: 11 Число деревьев: 150	0,961	0,874
RF	Максимальная глубина деревьев: 11 Число деревьев: 50	0,959	0,871
RF	Максимальная глубина деревьев: 9 Число деревьев: 50	0,936	0,873
RF	Максимальная глубина деревьев: 7 Число деревьев: 50	0,896	0,857
LB	Максимальная глубина деревьев: 9 Число деревьев: 150	0,922	0,883
LB	Максимальная глубина деревьев: 9 Число деревьев: 50	0,894	0,872
LB	Максимальная глубина деревьев: 7 Число деревьев: 150	0,921	0,882
LB	Максимальная глубина деревьев: 5 Число деревьев: 150	0,919	0,878
LB	Максимальная глубина деревьев: 11 Число деревьев: 150	0,922	0,883

Окончание табл. 3.2

Модель	Гиперпараметры	R^2 на обучающей выборке	R^2 на валидационной выборке
CB	Максимальная глубина деревьев: 7 Число деревьев: 100	0,687	0,673
CB	Максимальная глубина деревьев: 7 Число деревьев: 500	0,890	0,870
CB	Максимальная глубина деревьев: 9 Число деревьев: 500	0,904	0,871
CB	Максимальная глубина деревьев: 7 Число деревьев: 1000	0,914	0,880
CB	Максимальная глубина деревьев: 6 Число деревьев: 1000	0,897	0,876

На рис. 3.30–3.33 показаны результаты моделей с оптимизированными гиперпараметрами.

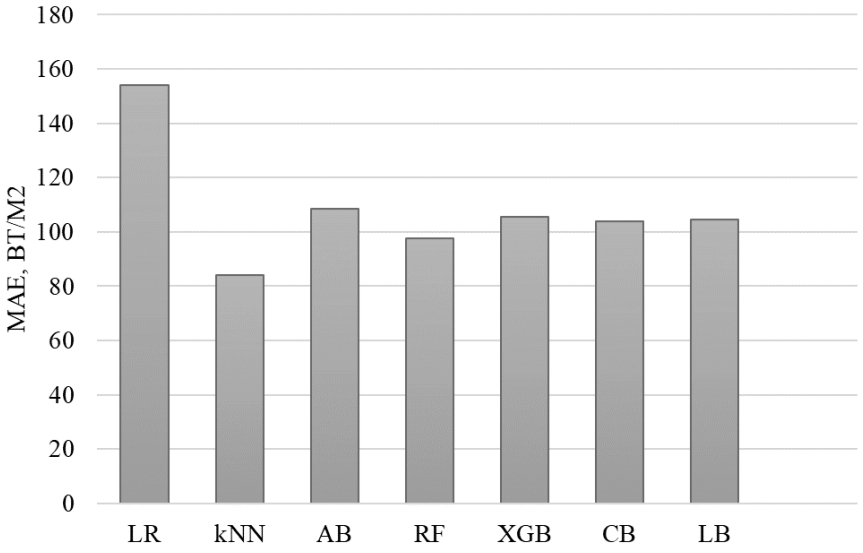


Рис. 3.30. Сравнение алгоритмов по средней абсолютной ошибке

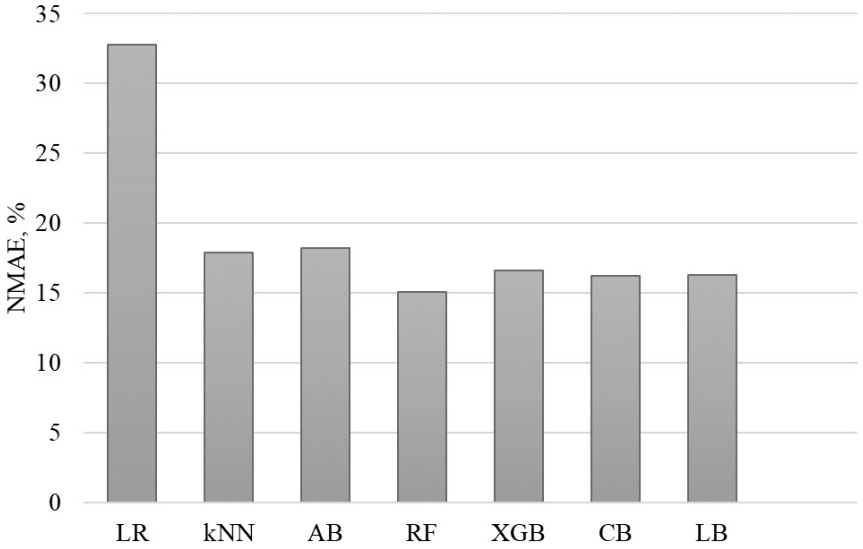


Рис. 3.31. Сравнение алгоритмов по нормализованной средней абсолютной ошибке

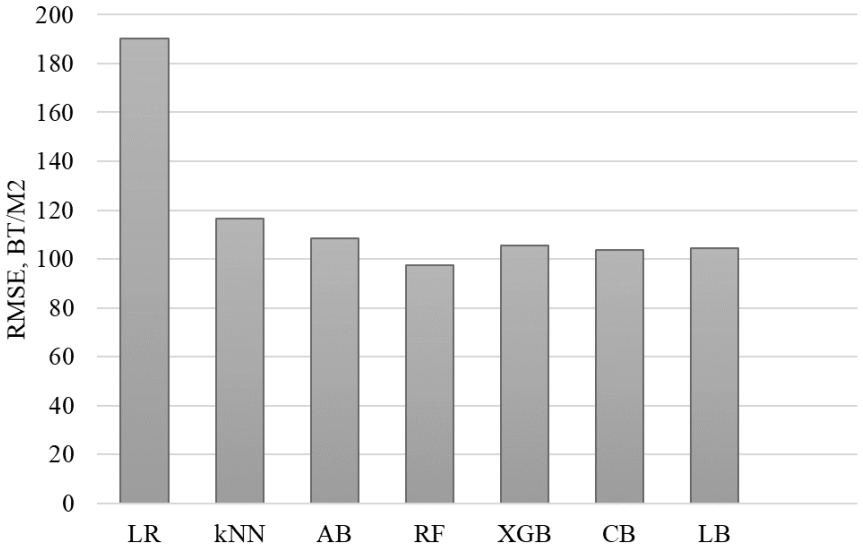


Рис. 3.32. Сравнение алгоритмов по среднеквадратической ошибке

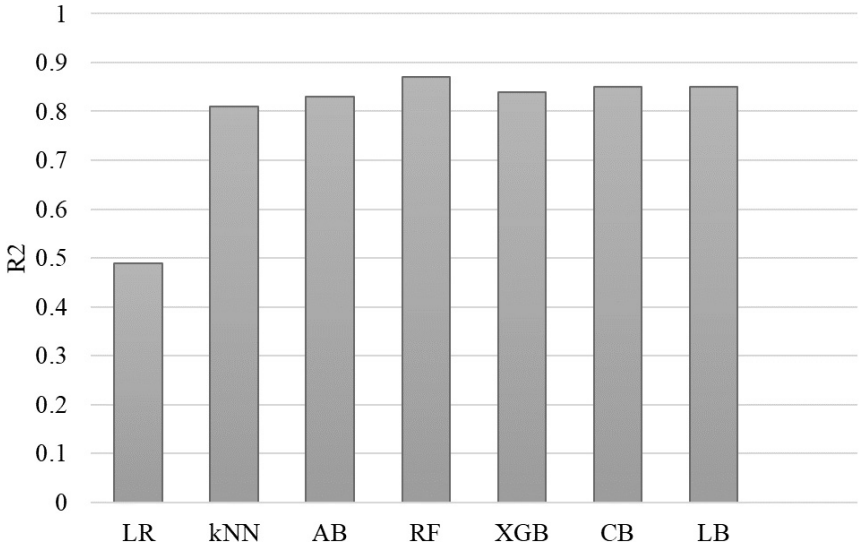


Рис. 3.33. Сравнение алгоритмов по коэффициенту детерминации

Полученные результаты позволяют сделать следующие выводы:

1) искомая зависимость инсоляции от метеорологических факторов является нелинейной (об этом говорит низкая точность линейной регрессии);

2) искомая зависимость требует построения сложных кусочно-непрерывных моделей, поскольку результаты алгоритма kNN оказались существенно хуже результатов ансамблей деревьев решений;

3) наилучшая точность получена при использовании алгоритма случайного леса;

4) для ансамблевых моделей изменение максимальной глубины деревьев оказывает намного большее влияние, чем изменение количества деревьев;

5) полученный уровень точности соответствует наилучшим результатам, описанным в научной литературе по теме прогнозирования инсоляции и выработки ФЭС.

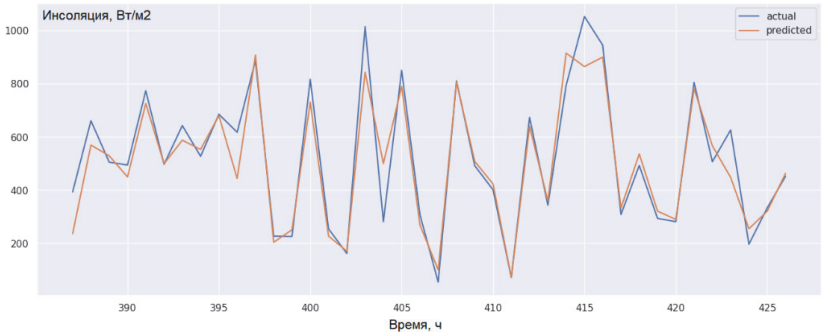


Рис. 3.34. Фрагмент тестовой выборки с сопоставлением фактических и прогнозных значений

На рис. 3.34 показан фрагмент тестовой выборки с наложением истинного графика и прогнозного (ночные часы исключены из выборки).

3.8. АЛГОРИТМ ОБЪЯСНЕНИЯ ПРОГНОЗОВ

Алгоритм объяснения прогнозов основан на методе аддитивного объяснения Шепли, описанного в разделе 2. Для каждого отдельного прогноза и каждого признака используется расчетное выражение (2.3), дающее для каждого прогнозного часа свой вектор значимости признаков. Затем выполняется его визуализация, как показано на рис. 3.35–3.38, где дана интерпретация прогнозов для нескольких часов одного дня.

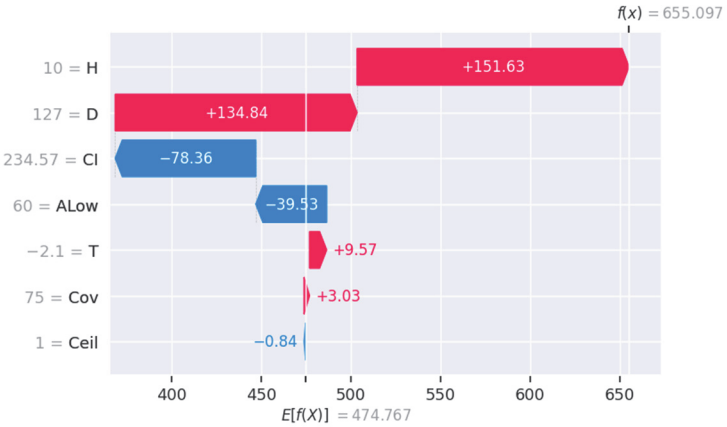


Рис. 3.35. Визуализация объяснения прогноза 10:00 7 мая

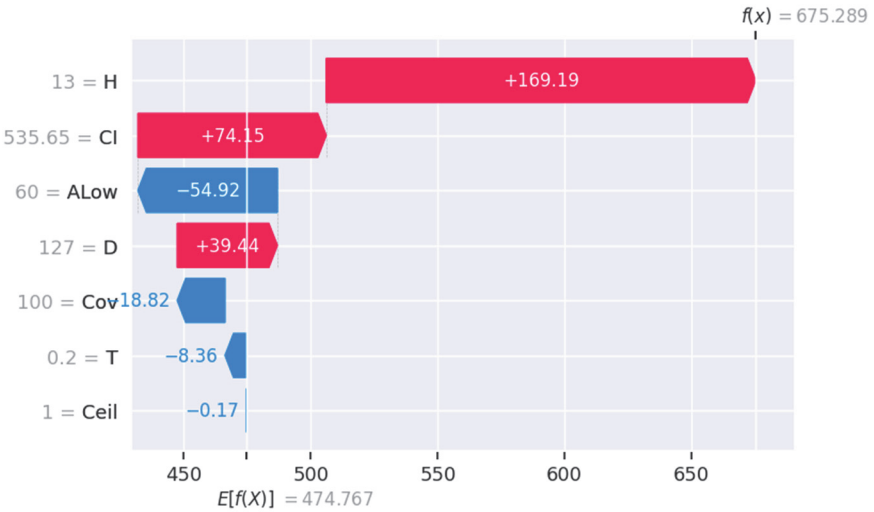


Рис. 3.36. Визуализация объяснения прогноза 13:00 7 мая

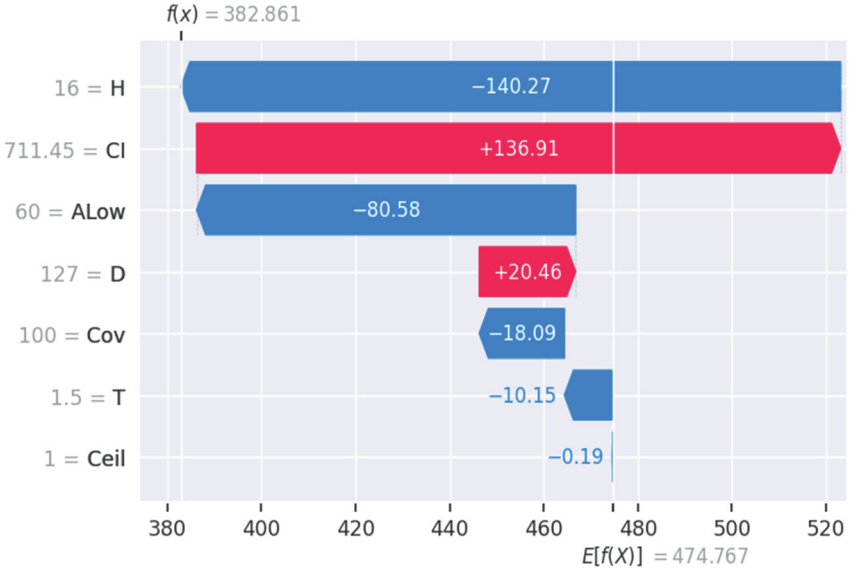


Рис. 3.37. Визуализация объяснения прогноза 16:00 7 мая

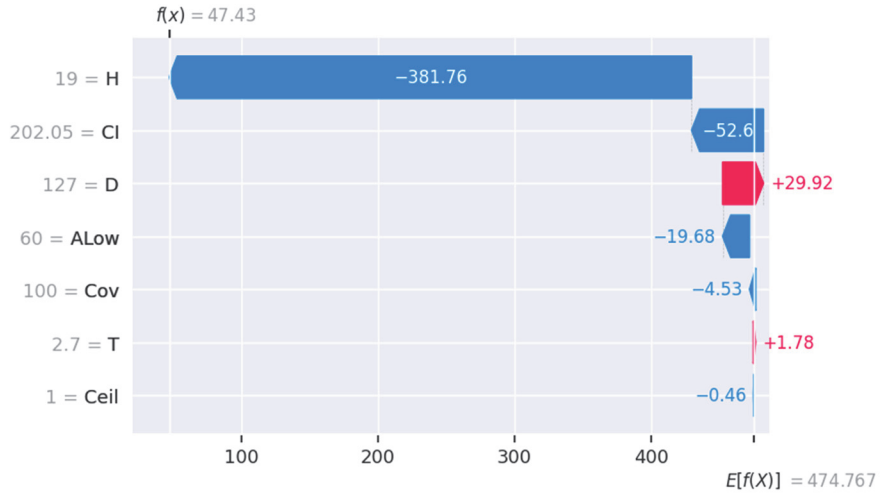


Рис. 3.38. Визуализация объяснения прогноза 19:00 7 мая

В 10 утра по местному времени (не совпадает с солнечным) в мае солнце уже высоко, поэтому признаки дня года и часа суток привели к повышению прогнозного значения относительно среднего по выборке. Невысокий уровень расчетной инсоляции и низкая облачность – к снижению прогноза.

В 13 часов дня, так как солнце достигло наивысшей точки среди рассматриваемых, час суток и расчетное значение инсоляции привели к повышению прогноза, однако облачность вызвала необходимость корректировки прогноза.

В 16 часов дня из-за снижения положения солнца над горизонтом и облачности прогноз стал существенно ниже, чем на 10 и 13 часов.

В 19 часов, поскольку это вечерняя время, номер часа суток оказал наибольшее влияние на близкое к нулю значение инсоляции.

Таким образом, специалист, оценивающий прогнозы, может понимать, как они строятся, и определять их корректность.

ЗАКЛЮЧЕНИЕ

Технологии искусственного интеллекта развиваются с высокой скоростью, что создает пробелы в существующем законодательстве. Адекватное правовое регулирование становится критически важным для обеспечения защиты прав человека, предотвращения дискриминации и этичного использования искусственного интеллекта. Поэтому важно создавать базовые принципы правового регулирования искусственного интеллекта, которые могли бы быть универсальными и независимыми от технологических особенностей. Тем не менее необходим постоянный мониторинг и обновление законодательства в случае выхода новых технологий из правового поля.

Поскольку наиболее успешные информационные технологии, в том числе с применением искусственного интеллекта, как правило, являются глобальными, важно согласовывать законодательства различных стран. Необходимо отметить, что из-за конкуренции за доминирующую роль в области искусственного интеллекта не стоит ожидать от каких-либо стран введения внутренних существенных ограничений на развитие технологий. Более вероятно продолжение и усиление государственной поддержки в этой области.

Регулирование искусственного интеллекта требует комплексного подхода, объединяющего юриспруденцию, технологии, этику и социальные науки. Это предполагает взаимодействие специалистов из разных областей для разработки эффективных норм и стандартов.

Авторы считают, что методы объяснимого искусственного интеллекта способны сыграть важную роль в повышении доверия к интеллектуальным информационным системам. Кроме того, они могут обеспечить прозрачность алгоритмов и возможность их интерпретации, чтобы пользователи и регулирующие органы могли понимать, как принимаются решения, что привело к тому или иному результату, на каком этапе

и по чьей вине случился тот или иной инцидент. Это позволит минимизировать риски ошибочных решений, избежать введения пользователей в заблуждение, дискриминации и нарушений этики при использовании интеллектуальных информационных систем.

Таким образом, развитие правовой базы и методов объяснимого искусственного интеллекта является ключевым фактором для интеграции интеллектуальных информационных систем в экономику и социальную сферу.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Marcucci S.* AI Localism in Practice: Examining How Cities Govern AI / S. Marcucci, U. Kalkar, S. Verhulst. – URL: https://files.thegovlab.org/ailocalism-in-practice.pdf (дата обращения: 12.09.2024). – Текст: электронный.

2. *Филипова И. А.* Правовое регулирование искусственного интеллекта : учеб. пособие / И. А. Филипова. – Нижний Новгород : Нижегородский госуниверситет, 2020. – С. 90.

3. *McCarthy J.* What is artificial intelligence / J. McCarthy. – 2007 – URL: http://wwwformal.stanford.edu/jmc/whatisai/whatisai.html (дата обращения: 12.09.2024). – Текст: электронный.

4. Роботизированная история с древних времён до наших дней [сайт]. – 2023. – URL: https://habr.com/ru/companies/inferit/articles/761622/ (дата обращения: 12.09.2024).

5. Принципы работы с ИИ, разработанные на асиломарской конференции [сайт]. – 2017. – URL: https://futureoflife.org/open-letter/ai-principles-russian/ (дата обращения: 12.09.2024).

6. В американских ИТ кадровая катастрофа: на все США за год открыто лишь 700 рабочих мест дней [сайт]. – 2024. – URL: https://www.cnews.ru/news/top/2024-01-09_v_it-sfere_ssha_kadrovaya_katastrofa (дата обращения: 12.09.2024).

7. First International Symposium on Roboethics – 30th – 31rd January 2004, Villa Nobel, Sanremo, Italy. – URL: http://www.roboethics.org/sanremo2004/ (дата обращения: 12.09.2024). – Текст: электронный.

8. World Robot Declaration [сайт]. – 2004. – URL: http://prw.kyodonews.jp/prwfile/prdata/0370/release/200402259634/index. (дата обращения: 12.09.2024).

9. Нормы гражданского права о робототехнике и Хартия робототехники [сайт]. – 2017. – URL: https://robotrends.ru/pub/1725/normy-grazhdanskogo-prava-o-robototehnike-i-hartiya-robototehniki (дата обращения: 12.09.2024).

10. *Asaro P.* What Should We Want from a Robot Ethic? / P. Asaro // International Review of Ethics. – 2006. – № 12. – P. 9–16.

11. Microsoft CEO Satya Nadella lays out 10 Laws of AI (and Human Behavior) [сайт]. – 2016. – URL: https://www.geekwire.com/2016/microsoft-ceo-satya-nadella-10-laws-ai/ (дата обращения: 12.09.2024).

12. Кодекс этики в сфере ИИ // – Комиссия по реализации Кодекса этики в сфере искусственного интеллекта// – Альянс в сфере искусственного интеллекта // – URL: https://ethics.a-ai.ru/ – (дата обращения: 12.09.2024) Текст: электронный

13. К Кодексу этики ИИ присоединились 11 российских и международных компаний [сайт]. – 2023. – URL: https://tass.ru/ekonomika/19362997 (дата обращения: 12.09.2024).

14. Реализация потенциала локализма ИИ [сайт]. – 2020. – URL: https://www.project-syndicate.org/commentary/local-regulation-of-artificial-intelligence-uses-by-stefaan-g-verhulst-1-and-mona-sloane-2020-02/russian?barrier=accesspaylog (дата обращения: 12.09.2024).

15. The Responsible Use and Design of Artificial Intelligence at the Local Level [сайт]. – 2024. – URL: https://ailocalism.org/# (дата обращения: 12.09.2024)

16. Тест: как хорошо вы знаете законы робототехники [сайт]. – 2024. – URL: https://trends.rbc.ru/trends/industry/65d720ad9a79476b3c51f792?from=copy (дата обращения: 12.09.2024).

17. Технологические тренды в 2023 году, Deloitte [сайт]. – 2023. – URL: https://ai.gov.ru/knowledgebase/infrastrukturaii/2023_tehnotrendy_v_2023_godu_tech_trends_2023_deloitte/ (дата обращения: 12.09.2024).

18. Иммерсивные технологии – будущее реального и виртуального опыта [сайт]. – 2022. – URL: https://trends.rbc.ru/trends/industry/62d15e099a794704c379cf3b?from=copyou (дата обращения: 12.09.2024).

19. 2023 На пути к более интеллектуальному завтрашнему дню: влияние ИИ в эпоху после COVID-19/Towards a smarter tomorrow: Impact of AI in the post-COVID [сайт]. – 2023. – URL: https://ai.gov.ru/knowledgebase/investitsionnaya-aktivnost/2023_na_puti_k_bolee_intellektualynomu_zavtrashnemu_dnyu_vliyanie_ii_v_epohu_posle_covid-19_towards_a_smarter_tomorrow_impact_of_ai_in_the_post-covid_era_pwc_india/ (дата обращения: 12.09.2024).

20. 2023 DeepTech 2022–2023. Инвестиционная активность: направления и тренды, Агентство инноваций города Москвы, Интеррос, Восход [сайт]. – 2023. – URL: https://ai.gov.ru/knowledgebase/obrazovanie-i-kadry/2023_deeptech_2022-2023_investicionnaya_aktivnosty_napravleniya_i_trendy_agentstvo_innovaciy_goroda_moskvy_interros_voshod/ (дата обращения: 12.09.2024).

21. 2023 Промежуточный отчет: Управление ИИ в интересах человечества/ Interim Report: Governing AI for Humanity, ООН, // – AI Advisory /BodyGoverning AI for Humanity// – 2023. – URL: https://ai.gov.ru/knowledgebase/mezhdunarodnye-dokumenty-po-razvitiyu-ii/2023_promeghutochnyy_otchet_upravlenie_ii_v_interesah_chelovechestva_interim_report_governing_ai_for_humanity_oon_ai_advisory_body (дата обращения: 12.09.2024) – Текст: электронный

22. *Anderson J.* The Future of Human Agency / J. Anderson, L. Rainie // Pew Research Center. – 2023 – P. 16–18.

23. *Wittenberg C.* Labeling AI-Generated Content: Promises, Perils, and Future Directions / C. Wittenberg, Z. Epstein, A. J. Berinsky, D. G. Randп // MIT Schwarzman College of Computing. – 2023.

24. Report on the Impact of Artificial Intelligence and Associated Technologies on Arms Control, Nonproliferation, and Verification // United States Departement of State Washington DC 20520. – October 31, 2023.

25. Этика для роботов: кто и как регулирует разработки ИИ [сайт]. – 2023. – URL https://bcs-express.ru/novosti-i-analitika/etika-dlia-robotov-kto-i-kak-reguliruet-razrabotki-ii: (дата обращения: 12.09.2024).

26. *Hoffmann M.* Adding Structure to AI Harm An Introduction to CSET's AI Harm Framework / M. Hoffmann, H. Heather Frase // Center for Security and Emerging Technology. – 2023.

27. Ответственность людей и решения искусственного интеллекта [сайт]. – 2021. – URL: https://habr.com/ru/articles/589913/ (дата обращения: 12.09.2024).

28. *Апресян Р. Г.* Ответственность – Институт Философии Российской Академии Наук – 2018 / Р. Г. Апресян. – URL: https://iphlib.ru/library/collection/newphilenc/document/HASH01beef80ca72acb7a42e9dbe – (дата обращения: 12.09.2024). – Текст: электронный.

29. Цитаты известных личностей [сайт]. – 2024. – URL: https://ru.citaty.net/tsitaty/627539-mikhail-aleksandrovich-bakunin-svoboda-odnogo-cheloveka-zakan-chivaetsia-tam-gde-nach/ (дата обращения: 12.09.2024)

30. *Козельская Н. Л.* Формы вины на примере ряда зарубежных государств / Н. Л. Козельская. – Государственный университет – учеб.-науч. производственный комплекс г. Орёл – URL: https://cyberleninka.ru/article/n/formy-viny-na-primere-ryada-zarubezhnyh-gosudarstv/viewer (дата обращения: 12.09.2024). – Текст: электронный.

31. Формы вины. [сайт]. – 2024 – URL: https://www.yaklass.ru/p/obshchestvoznanie/11-klass/pravo-7270216/iuridicheskaia-otvetstvennost-7276184/re-389f303e-52cd-4a06-976b-14b5a1874518?ysclid=lzz2fq8w46863190054 (дата обращения: 12.09.2024).

32. Вымогатели начали использовать ИИ для подделки голосовых в Telegram. Чем новая схема опасна для пользователей [сайт]. – 2024. – URL: https://www.rbc.ru/technology_and_media/10/01/2024/659d37899a79473f8a99e35f?from=copy (дата обращения: 12.09.2024).

33. «Оказалось, что на меня оформлен заём»: как личные данные россиян сливают мошенникам [сайт]. – 2022 – URL: https://russian.rt.com/russia/article/950984-razvod-dannye-moshenniki (дата обращения: 12.09.2024).

34. Юридическая ответственность [сайт]. – 2024 – URL: https://www.yaklass.ru/p/obshchestvoznanie/11-klass/pravo-7270216/iuridicheskaia-otvetstvennost-7276184 (дата обращения: 12.09.2024).

35. *Shepardson D.* U.S. opens special probe into fatal Tesla pedestrian crash in California – 2022 / D. Shepardson. – URL: https://www.reuters.com/business/autos-

transportation/us-opens-new-probe-into-fatal-tesla-pedestrian-crash-california-2022-07-07/ (дата обращения: 12.09.2024) – Текст: электронный.

36. *Chen A*. IBM's Watson gave unsafe recommendations for treating cancer – 2018 / A. Chen. – URL: https://www.theverge.com/2018/7/26/17619382/ibms-watson-cancer-ai-healthcare-science (дата обращения: 12.09.2024). – Текст: электронный.

37. *Гурова М. Е.* Право интеллектуальной собственности: авторское право на труды искусственного интеллекта / М. Е. Гурова // Вопросы студенческой науки. – 2021. – № 6 (58). – С. 231–234.

38. *Материк М.* Ответственность искусственного интеллекта в правовом поле / М. Материк. – 2024. – URL: https://dgtlaw.ru/analytic/otvetstvennost-iskusstvennogo-intellekta-v-pravovom-pole (дата обращения: 12.09.2024). – Текст: электронный.

39. Власти разработали механизм защиты от причиненного технологиями ИИ вреда [сайт]. – 2023. – URL: https://www.rbc.ru/economics/07/12/2023/657063269a79470fdb112ac3?from=copy (дата обращения: 12.09.2024).

40. Искусственный интеллект победил в категории «Креатив» на Sony World Photography Awards 2023 [сайт]. – 2023. – URL: https://adindex.ru/news/creative/2023/04/18/312028.phtml (дата обращения: 12.09.2024).

41. Ограниченный и субъективный, безразличный и прожорливый: четыре главных проблемы искусственного интеллекта [сайт]. – 2023. – URL https://habr.com/ru/articles/586942/ (дата обращения: 12.09.2024).

42. *Могиленко А.* Применение алгоритмов искусственного интеллекта в мировой энергетике / А. Могиленко. – URL: https://www.eprussia.ru/epr/345-346/4513899.htm?ysclid=m00v8sb4x7682255314 (дата обращения: 12.09.2024). – Текст: электронный.

43. Xcel Energy оптимизирует в США работу умных энергосчетчиков с помощью IoT [сайт]. – 2019. – URL https://tass.ru/ekonomika/6899094 (дата обращения: 12.09.2024).

44. IBM повысит точность прогнозов солнечной энергии на 30% [сайт]. – 2015. – URL: http://www.abercade.ru/research/industrynews/13495.html (дата обращения: 12.09.2024).

45. DeepMind предложила сократить расходы энергии с помощью нейронных сетей [сайт]. – 2019. – URL https://22century.ru/energetics/45567?ysclid=m00vn1foa2674129398 (дата обращения: 12.09.2024).

46. Что представляет собой искусственный интеллект (ИИ) [сайт]. – 2023. – URL: https://habr.com/ru/articles/710350/ (дата обращения: 12.09.2024).

47. Искусственный интеллект в электроэнергетике: зачем и на что он способен. Пример ИИ-системы) [сайт]. – 2022. – URL: https://habr.com/ru/articles/674110/(дата обращения: 12.09.2024).

48. *Хальясмаа А. И.* Синтез моделей и методов автоматизированной диагностики высоковольтного оборудования электрических станций и подстанций : автореф. дис. … д-ра техн. наук / А. И. Хальясмаа. – Новосибирск, 14 мая 2024.

49. С BIM 360 на Pilot-ICE Enterprise за 20 дней [сайт]. – 2024. – URL: https://isicad.ru/ru/articles.php?article_num=19915 (дата обращения: 12.09.2024).

50. Vizorlabs Platform [сайт]. – URL: https://vizorlabs.ru/(дата обращения: 12.09.2024).

51. Умная система контроля качества NordClan на предприятии «Росатома» [сайт]. – 2023. – URL: https://www.atomic-energy.ru/articles/2023/11/24/140762?ysclid=m00wbsqduf893257108 (дата обращения: 12.09.2024).

52. 8 кейсов использования компьютерного зрения на производстве [сайт]. – 2021. – URL: https://smartgopro.com/novosti2/computervision/?ysclid=m00wj6g5j2981239872 (дата обращения: 12.09.2024).

53. ИИ управляет китайской скоростной ЖД-сетью протяженностью 45 000 км [сайт]. – 2024. – URL: https://hightech.plus/2024/03/14/ii-upravlyaet-kitaiskoi-skorostnoi-zhd-setyu-protyazhennostyu-45-000-km (дата обращения: 12.09.2024).

54. Мосэнергосбыт улучшает клиентский опыт с помощью речевых технологий группы ЦРТ [сайт]. – 2023. – URL: https://www.mosenergosbyt.ru/individuals/news/mosenergosbyt-uluchshaet-klientskiy-opyt-s-pomoshchyu-rechevykh-tekhnologiy-gruppy-tsrt/ (дата обращения: 12.09.2024).

55. Natural Language Processing: Why the Wind Industry Needs It [сайт]. – 2017. – URL: https://www.sparkcognition.com/natural-language-processing-why-wind-industry-needs-it/ (дата обращения: 12.09.2024).

56. *Silberstein S.* U.S. Government Uses for Artificial Intelligence // – January 16, 2024. – URL: https://www.investopedia.com/artificial-intelligence-in-us-government-8406703 (дата обращения: 12.09.2024). – Текст: электронный.

57. AI Governance Alliance Briefing Paper Series – January 2024 – AI Governance Alliance Briefing Paper Series – World Economic Forum – 2024.

58. Interim Report: Governing AI for Humanity – December 2023. – URL: https://www.un.org/en/ai-advisory-body (дата обращения: 12.09.2024). – Текст: электронный.

59. iTeh STANDARD PREVIEW (standards.iteh.ai) – Information technology – Artificial intelligence (AI) – Bias in AI systems and AI aided decision making – Switzerland 2021 – ISO/IEC TR 24027:2021(E).

60. Информационно-аналитическая справка «Зарубежная практика установления требований к безопасности ИИ» – НЦРИИ – 2023.

61. AI Watch [сайт]. – URL: https://ai-watch.ec.europa.eu/index_en (дата обращения: 12.09.2024).

62. Мониторинг основных международных и иностранных документов и инициатив в сфере искусственного интеллекта, включая документы международных и зарубежных альянсов в сфере искусственного интеллекта, международных организаций и объединений по вопросам искусственного интеллекта, за четвертый квартал 2023 года – НЦРИИ –2023.

63. Artificial Intelligence in China Sino Fact Sheet giz-//German Industrie 4.0 Project 2017.

64. China's New Rules For Generative AI: An Emerging Regulatory Framework [сайт]. – 2015. – URL: https://www.fasken.com/en/knowledge/2023/08/chinas-new-rules-for-generative-ai (дата обращения: 12.09.2024).

65. *Бутл Р.* Искусственный интеллект и экономика: Работа, богатство и благополучие в эпоху мыслящих машин : пер. с англ. / Р. Бутл. – Москва : Альпина ПРО, 2023. – С. 13, 21–23.

66. *Баринова Н. В.* Цифровая экономика, искусственный интеллект, Индустрия 5.0: вызовы современности / Н. В. Баринова, В. Р. Баринов // Вестник Российского экономического университета имени Г. В. Плеханова. – 2022;(5):23-34. – URL: https://doi.org/10.21686/2413-2829-2022-5-23-34

67. Будущее сейчас. Как технологии искусственного интеллекта влияют на экономику и бизнес [сайт]. – 2023. – URL: https://sber.pro/digital/publication/budushhee-sejchas-kak-tehnologii-iskusstvennogo-intellekta-vliyayut-na-ekonomiku-i-biznes/?ysclid=lugth8spcl486509443 (дата обращения: 12.09.2024).

68. Мишустин: внедрение искусственного интеллекта в экономике РФ достигает 20 %, план к 2024 году – 50% [сайт]. – 2023. – URL: https://www.interfax-russia.ru/main/mishustin-vnedrenie-iskusstvennogo-intellekta-v-ekonomike-rf-dostigaet-20-plan-k-2024-godu-50 (дата обращения: 12.09.2024).

69. «Эти технологии уже стали частью нашей жизни». Искусственный интеллект меняет российскую экономику. Как это работает? [сайт]. – 2023. – URL: https://lenta.ru/articles/2023/06/14/airecr/

70. Альманах «Искусственный интеллект» // Аналитический сборник № 12, центр компетенций НТИ «Искусственный интеллект». – Москва, 2023.

71. Artificial Intelligence Index Report 2023 Stanford University Human-Cented Artificial Intellegence. – 2023.

72. *Asimov I.* (1988) Asimov's Book of Science and Nature Quotations / I. Asimov, J. A. Shulman. – New York : Grove Press, 1998.

73. *Krugman P.* (2017) New Zealand Parliament / P. Krugman. – Vol. 644, week 63. –2017.

74. Промышленные революции. Ключевые изменения и результаты [сайт]. – URL: https://skvot.2035.university/promyshlennye-revolyucii (дата обращения: 12.09.2024).

75. Четыре промышленные революции [сайт]. – 2021. – URL: https://postnauka.org/wtf/155993 (дата обращения: 12.09.2024).

76. *Ricardo D.* On the Principles of Political Economy and Taxation / D. Ricardo. – 1821.

77. FoldiMate (робот для укладки одежды) [сайт]. – 2019. – URL: https://www.tadviser.ru/index.php/Продукт:FoldiMate_(робот_для_укладки_одежды)?ysclid=lzz6r26jre261673597 (дата обращения: 12.09.2024).

78. Ученые в Сингапуре научили роботов собирать стул из IKEA [сайт]. – 2018. – URL: https://iz.ru/734076/2018-04-19/uchenye-v-singapure-nauchili-robotov-sobirat-stul-iz-ikea (дата обращения: 12.09.2024).

79. Робота-хирурга Da Vinci обвинили в убийстве женщины [сайт]. – 2024. – URL: https://naked-science.ru/community/929220 (дата обращения: 12.09.2024).

80. *Avent R.* The Wealth of Humans: Work, Power, and Status in the Twenty-First Century / R. Avent. – New York : St. Martin's Press, 2016. – P. 59.

81. *Bregman R.* Utopia for Realists / R. Bregman. – London: Bloomsbury Publishing, 2017. – P. 185.

82. *Schwab K*. Shaping the Future of the Fourth Industrial Revolution / K. Schwab. – London : Penguin Radom House, 2018. – P. 23.

83. *Crawford K.* Social Work and Human Development / K. Crawford, J. Janet Walker // British Library Cataloguing Data. – 2017.

84. *Иванова Л. Н.* Искусственный интеллект на службе управления инцидентами и конфликтами на предприятии / Л. Н. Иванова, Е. А. Налимова // Современные проблемы инновационной экономики. – 2020. – № 7. – С. 71–77.

85. *Шишацкий Н. Г.* Структурная модернизация как фактор повышения конкурентоспособности региона: (на примере Красноярского края) / Н. Г. Шишацкий, Е. А. Брюханова, Р. В. Гордеева. – Новосибирск : ИЭОПП СО РАН, 2020. – С. 509.

86. Что представляет собой искусственный интеллект (ИИ) – 2023. – URL: https://habr.com/ru/articles/710350/ (дата обращения: 12.09.2024).

87. *Walch K.* Common sense in AI remains elusive) [сайт]. – 2020. – URL: https://www.techtarget.com/searchenterpriseai/feature/Common-sense-in-AI-remains-elusive (дата обращения: 12.09.2024) Текст: электронный.

88. *Barney N*. Artificial superintelligence (ASI) – 2023 / N. Barney. – URL: https://www.techtarget.com/searchenterpriseai/definition/artificial-superintelligence-ASI (дата обращения: 12.09.2024) Текст: электронный.

89. Коротко об аппаратной части искусственного интеллекта [сайт]. – 2023. – URL: https://la-chatte.com/articles/web/korotko-ob-apparatnoy-chasti-iskusstvennogo-intellekta/?ysclid=lv3noxk61m698160035 (дата обращения: 12.09.2024).

90. Топ-10 производителей чипов для искусственного интеллекта 2023 года [сайт]. – 2023. – URL: https://globalcorporations.ru/top-10-proizvoditelej-chipov-dlya-iskusstvennogo-intellekta-2023-goda/?ysclid=lymwycdgat714625604 (дата обращения: 12.09.2024).

91. Обзор аппаратных решений для задач искусственного интеллекта: США, Китай, Россия [сайт]. – 2023. – URL: https://habr.com/ru/companies/baikalelectron/articles/750552/ (дата обращения: 12.09.2024).

92. Компании-производители чипов AI: рейтинг 2024 [сайт]. – 2024. – URL: https://acomsupply.com/news/novosti_elektronnykh_komponentov/kompanii_proiz

voditeli_chipov_ai_reyting_2024/?ysclid=m00linp4ry565222085 (дата обращения: 12.09.2024).

93. 7 лидеров в индустрии чипов, переживающей бум ИИ [сайт]. – 2023. – URL: https://lenta.profinansy.ru/selections/1928028?ysclid=m00liplzov768119213 (дата обращения: 12.09.2024).

94. Искусственный интеллект (мировой рынок) [сайт]. – 2024. – URL: https://www.tadviser.ru/index.php/Статья:Искусственный_интеллект_(мировой_рынок)?ysclid=m00lkfcxy3754360387 (дата обращения: 12.09.2024).

95. Top 20 AI Chip Makers of 2024: NVIDIA's Upcoming Competitors [сайт]. – 2024. – URL: https://research.aimultiple.com/ai-chip-makers/ (дата обращения: 12.09.2024).

96. Annual Report 2022 (Form 10-K) (англ.). Intel Corporation.

97. Конкурент Nvidia прекратит продажи в Китае из-за торговых ограничений США ИИ [сайт]. – 2023. – URL: https://kz.kursiv.media/2023-11-23/rntt-graphcore/?ysclid=lv4tujz252147103728 (дата обращения: 12.09.2024).

98. Cerebras launches new AI supercomputing processor with 2.6 trillion transistors [сайт]. – 2021. – URL: https://venturebeat.com/ai/cerebras-systems-launches-new-ai-supercomputing-processor-with-2-6-trillion-transistors/ (дата обращения: 12.09.2024).

99. AMD says U.S. told it to stop shipping top AI chip to China [сайт]. – 2022. – URL: https://www.reuters.com/technology/amd-says-us-told-it-stop-shipping-top-ai-chip-china-2022-08-31/ (дата обращения: 12.09.2024).

100. Полупроводники: мировой рынок [сайт]. – 2024. – URL: https://www.tadviser.ru/index.php/Статья:Полупроводники_%28мировой_рынок%29 (дата обращения: 12.09.2024).

101. Полупроводники: рынок США [сайт]. – 2023. – URL https://www.tadviser.ru/index.php/Статья:Полупроводники_%28рынок_США%29 (дата обращения: 12.09.2024).

102. Worldwide Silicon Wafer Shipments and Revenue Set New Records in 2022, SEMI Reports [сайт]. – 2022. – URL: https://www.semi.org/en/news-media-press-releases/semi-press-releases/worldwide-silicon-wafer-shipments-and-revenue-set-new-records-in-2022-semi-reports (дата обращения: 12.09.2024).

103. Полупроводники: рынок России [сайт]. – 2024. – URL https://www.tadviser.ru/index.php/Статья:Полупроводники_(рынок_России) (дата обращения: 12.09.2024).

104. США за год сократили экспорт микрочипов почти на 20 процентов [сайт]. – 2023. – URL https://ria.ru/20231002/chipy-1899823856.html (дата обращения: 12.09.2024).

105. Полупроводники: рынок Китая [сайт]. – 2024. – URL: https://www.tadviser.ru/index.php/Статья:Полупроводники_(рынок_Китая (дата обращения: 12.09.2024).

106. Полупроводники: рынок Индии [сайт]. – 2024. – URL: https://www.tadviser.ru/index.php/Статья:Полупроводники_(рынок_Индии (дата обращения: 12.09.2024).

107. Индекс интеллектуальной зрелости отраслей экономики, секторов социальной сферы и системы государственного управления Российской Федерации45// – Москва, 2023 // – Национальный центр развития искусственного интеллекта при Правительстве Российской Федерации – 2023.

108. The art of AI maturity [сайт]. – 2024. – URL: https://www.accenture.com/us-en/insights/artificial-intelligence/ai-maturity-and-transformation (дата обращения: 12.09.2024).

109. Аналитический отчет о развитии промышленных метавселенных [сайт]. – 2024. – URL: https://ict.moscow/research/analiticheskii-otchet-o-razvitii-promyshlennoi-metavselennoi/ (дата обращения: 12.09.2024).

110. Лидерство при помощи периферийных вычислений [сайт]. – 2023. – URL: https://ict.moscow/research/liderstvo-pri-pomoshchi-periferiinykh-vychislenii (дата обращения: 12.09.2024).

111. Правительство оценит внедрение искусственного интеллекта в госорганах [сайт]. – 2023. – URL: https://www.vedomosti.ru/technology/articles/2023/01/24/960150-pravitelstvo-otsenit-vnedrenie-iskusstvennogo-intellekta (дата обращения: 12.09.2024).

112. Индекс готовности приоритетных отраслей к внедрению искусственного интеллекта [сайт]. – 2023. – URL: https://ai.gov.ru/ai/implementation/?ysclid=lvdd2brgt7200410165 (дата обращения: 12.09.2024).

113. *Winograd T.* Understanding Computers and Cognition: A New Foundation for Design / T. Winograd, F. Flores. – Norword, NJ: Ablex Publishing Corporation, 1986. – 29 c.

114. Explainable artificial intelligence (XAI). Challenges of model interpretability. – Management Solution. – 2023. – 48 p.

115. For Principles of Explainable Artificial Intelligence. National Institute of Standards and Technology Interagency os Internal Report 8312, 2021. – 43 p. – DOI 10.6028/NIST.IR.8312

116. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence / A. Sakid, T. Abuhmed, E.-S. Shaker [et al.] // Information Fusion. – 2023. – Vol. 99. – P. 101805.

117. IEEE Approved Draft Standard for Transparency of Autonomous Systems. IEEE P7001/D4. 2021. 75 pp. ISBN:978-1-5044-8060-4.

118. Ethically aligned design. A vision for Prioritizing Human Well-being with Autonomous and Intellegent System. Version 2 For Public Discussion. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. – 266 p.

119. The Belmont Report. – URL: https://www.hhs.gov/ohrp/sites/default/files/the-belmont-report-508c_FINAL.pdf (дата обращения: 19.09.2024).

120. Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions / L. Longo, M. Brcic, F. Cabitza [et al.] // Information Fusion. – 2024. – Vol. 106. – P. 102301.

121. AI Explainability Whitepaper. Google. – URL: https://cerre.eu/wp-content/uploads/2020/07/ai_explainability_whitepaper_google.pdf (дата обращения: 19.09.2024).

122. Introduction to Vertex Explainable AI. – URL: https://cloud.google.com/vertex-ai/docs/explainable-ai/overview (дата обращения: 19.09.2024).

123. Whitepaper 2023. The rise of industrial explainable artificial intelligence (XAI) – Insights across the AI life cycle. – URL: https://assets.new.siemens.com/siemens/assets/api/uuid:3b4de373-57e2-4329-b025-2825db0172aa/WhitepaperXAI.pdf (дата обращения: 19.09.2024).

124. Atos XAI. Simplified Explainable AI. – URL: https://atos.net/wp-content/uploads/2022/08/atos-xai-simplified-explainable-white-paper.pdf (дата обращения: 19.09.2024).

125. Intel® Explainable AI Tools. – URL: https://www.intel.com/content/www/us/en/developer/articles/reference-implementation/explainable-ai-tools.html (дата обращения: 19.09.2024).

126. Explainable AI – how humans can trust AI. – URL: https://www.ericsson.com/49876b/assets/local/reports-papers/white-papers/explainable-ai-how-humans-can-trust-ai_whitepaper.pdf (дата обращения: 19.09.2024).

127. Quantum XAI. – URL: https://quantumxai.tech/whitepaper (дата обращения: 19.09.2024).

128. SideShift Token (xai). White paper. – URL: https://sideshift.ai/xai-white-paper (дата обращения: 19.09.2024).

129. Implementing Conversational and Generative AI in Financial Services Contact Centers. – URL: https://amelia.ai/white-paper/implementing-conversational-and-generative-ai-in-financial-services-contact-centers/ (дата обращения: 19.09.2024).

130. Demystifying AI: A Guide to Explainable AI (XAI) Techniques. – URL: https://www.coforge.com/hubfs/Demystifying%20AI%20-%20A%20Guide%20to%20Explainable%20AI%20(XAI)%20Techniques.pdf (дата обращения: 19.09.2024).

131. Nvidia: Explainable AI for credit risk management: applying accelerated computing to enable explainability at scale for AI-powered credit risk management using Shapley values and SHAP. – URL: https://www.gov.uk/ai-assurance-techniques/nvidia-explainable-ai-for-credit-risk-management-applying-accelerated-computing-to-enable-explainability-at-scale-for-ai-powered-credit-risk-management-using-shapley-values-and-shap (дата обращения: 19.09.2024).

132. Miscrodoft's vision for AI in the enterprise. – URL: https://info.microsoft.com/rs/157-GQE-382/images/EN-AU-CNTNT-Whitepaper-DigitalTransformation-MSFTvisionforAIintheenterprise.pdf (дата обращения: 19.09.2024).

133. What is AI ethics? – URL: https://www.ibm.com/topics/ai-ethics (дата обращения: 19.09.2024).

134. How Conversational XAI Makes AI More Responsible. – URL: https://www.deeploy.ml/wp-content/uploads/White-paper-conversational-XAI-6.pdf (дата обращения: 19.09.2024).

135. About xAI. – URL: https://x.ai/about (дата обращения: 19.09.2024).

136. Указ Президента Российской Федерации от 10.10.2019 № 490 «О развитии искусственного интеллекта в Российской Федерации». Дата опубликования: 11.10.2019. – 25 с.

137. ГОСТ Р 59276-2020. Системы искусственного интеллекта. Способы обеспечения доверия. Общие положения. Введ. с 23.12.2020 – Москва : Стандартинформ. – 12 с.

138. Приложение 1. Письмо от Минобрнауки России № МН-5/22720 от 21.2021. «О направлении доработанной модели компетенция». – URL: https://fgosvo.ru/uploadfiles/Information/Ps_MON_5_22720_21122022-1.pdf (дата обращения: 19.09.2024).

139. Образно-логический ИИ. – URL: https://explicable.ru/ (дата обращения: 19.09.2024).

140. Glagol. Телефонные роботы способные решать сложные задачи. – URL: https://glagol.ai/ (дата обращения: 19.09.2024).

141. Карточка проекта фундаментальных и поисковых научных исследований, поддержанного российским научным фондом. Номер 23-11-20024. – URL: https://rscf.ru/project/23-11-20024/ (дата обращения: 19.09.2024).

142. *Linardatos P*. Explainable AI: A Review of Machine Learning Interpretability Methods / P. Linardatos, S. Kotsiantis, V. Papastefanopolous // National Library of Medicine PubMed Central. – 2022. – 23(1). – 18.

143. *Christofer Molnar*. Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. – URL: https://christophm.github.io/interpretable-ml-book/ (дата обращения: 19.09.2024).

144. *Vilone G*. Classification of Explainable Artificial Intelligence Methods through Their Output Formats / G. Vilone, L. Longo // Machine Learning and Knowledge Extraction. – 2021. – 3(3). – P. 615–661.

145. AI Explainability 360 – Resources. – URL: https://aix360.res.ibm.com/resources#guidance (дата обращения: 19.09.2024).

146. *Goldstein A*. Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation / A. Goldstein, A. Kapelner, J. Bleich, E. Pitkin // Journal of Computational and Graphical Statistics. – 2014. – 24(1). – P. 1–22.

147. *Fernandez-Centeno M. A*. Identification of Trends in Dam Monitoring Data Series Based on Machine Learning and Individual Conditional Expectation Curves / M. A. Fernandez-Centeno, P. Alocen, M. A. Toledo // Water. – 2024. – 16. – 1239.

148. *Abdullah T. A. A*. A Review of Interpretable ML in Healthcare: Taxonomy, Applications, Challenges, and Future Directions / T. A. A. Abdullah, M. S. M. Zahid, W. A. Ali // Symmetry. – 2021. – 13. – 2439.

149. *Wadoux A. M. J.-C.* Beyond prediction: methods for interpreting complex models of soil variation / A. M. J.-C. Wadoux // Geouderma. – 2020. – 422. – 115953.

150. iml: Interpretable Machine Learning. – URL: https://cran.r-project.org/web/packages/iml/index.html (дата обращения: 20.09.2024).

151. 4.1. Partial Dependence and Individual Conditional Expectation plots. – URL: https://scikit-learn.org/stable/modules/partial_dependence.html (дата обращения: 20.09.2024).

152. *Ribeiro M. T*. "Why Should I Trust You?": Explaining the Predictions of Any Classifier / M. T. Ribeiro, S. Singh, C. Guestrin // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – 2016. – P. 1135–1144.

153. *Slack D*. Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods / D. Slack, S. Hilgard, E. Jia, S. Singh, H. Lakkaraju // Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. – 2020. – P. 180–186.

154. *Hulsmann J*. Local Interpretable Explanations of Energy System Designs / J. Hulsmann, J. Barbosa, F. Steinke // Energies. – 2023. – 16. – 2161.

155. *Khan A. M*. BIM Integration with XAI Using LIME and MOO for Automated Green Building Energy Performance Analysis / A. M. Khan, M. A. Tariq, S. K. U. Rehman, T. Saeed, F. K. Alqahtani, M. Sherif // Energies. – 2024. – 17. – 3295.

156. *Grzeszczyk T. A*. Justifying Short-Term Load Forecasts Obtained with the Use of Neural Models / T. A. Grzeszczyk, M. K. Grzeszczyk // Energies. – 2022. – 1. – 1852.

157. lime. – URL: https://github.com/marcotcr/lime (дата обращения: 20.09.2024).

158. lime: Local Interpretable Model-Agnostic Explanations. – URL: https://cran.r-project.org/web/packages/lime/index.html (дата обращения: 20.09.2024).

159. *Ribeiro M. T*. Anchors: High-Precision Model-Agnostic Explanations / M. T. Ribeiro, S. Singh, C. Guestrin // Thirty-Second AAAI Conference on Artificial Intelligence. – 2018. – Vol. 32. – № 1. – P. 1–9.

160. *Liu X*. Dynamic Anchor: A Feature-Guided Anchor Strategy for Object Detection / X. Liu, H.-X. Chen, B.-Y. Liu // Applied Sciences. – 2022. – 12. – 4897.

161. *Bashaiwth A*. An Explanation of the LSTM Model Used for DDoS Attacks Classification / A. Bashaiwth, H. Binsalleeh, B. AsSadhan // Applied Sciences. – 2023. – 13(15). – 8820.

162. *Sharma N. A*. Explainable AI Frameworks: Navigating the Present Challenges and Unveiling Innovative Applications / N. A. Sharma, R. R. Chand, Z. Buksh, A.B.M.S. Ali, A. Hanif, A. Beheshti // Algorithms. – 2024. – 17. – 227.

163. *Mirzaei S*. Explainable AI Evaluation: A Top-Down Approach for Selecting Optimal Explanations for Black Box Models / S. Mirzaei, H. Mao, R.R.O. Al-Nima, W.L. Woo, W.L. // Information 2024. – 15. – 4.

164. Anchor. – URL: https://github.com/marcotcr/anchor (дата обращения: 20.09.2024).

165. Anchorj. – URL: https://github.com/viadee/javaAnchorExplainer (дата обращения: 20.09.2024).

166. *Wachter S*. Counterfactual explanations without opening the black box: automated decisions and the GDPRc / S. Wachter, B. Mittelstadt, C. Russell // Harvard Journal of Law and Technology. – 2018. – 31(2). – Pp. 841–887.

167. *Dandl S*. Multi-Objective Counterfactual Explanations / S. Dandl, C. Molnar, M. Binder, B. Bischl // Parallel Problem Solving from Nature – PPSN XVI. – 2020. – P. 448–469.

168. *Tursunalieva A*. Making Sense of Machine Learning: A Review of Interpretation Techniques and Their Applications / A. Tursunalieva, D. L. J. Alexander, R. Dunne, J. Li, L. Riera, Y. Zhao // Applied Sciences. – 2024. – 14. – 496.

169. *Dindorf C*. Machine Learning and Explainable Artificial Intelligence Using Counterfactual Explanations for Evaluating Posture Parameters / C. Dindorf, O. Ludwig, S. Simon, S. Becker, M. Frohlich // Bioengineering. – 2023. – 10(5). – 511.

170. Counterfactuals: Counterfactual Explanations. – URL: https://github.com/dandls/moc/tree/master/counterfactuals (дата обращения: 20.09.2024).

171. ALIBI EXPLAIN. – URL: https://github.com/SeldonIO/alibi (дата обращения: 20.09.2024).

172. mace. General. – URL: https://github.com/charmlab/mace (дата обращения: 20.09.2024).

173. Diverse Counterfactual Explanations (DiCE) for ML. – URL: https://github.com/interpretml/DiCE (дата обращения: 20.09.2024).

174. *Shapley L*. A Value for n-Person Games. Contributions to the Theory of Games II. – Princeton University Press. – Pp. 307–317. DOI: 10.1515/9781400881970-018.

175. *Shao B*. VMD-WSLSTM Load Prediction Model Based on Shapley Values / B. Shao, Y. Yan, H. Zeng // Energies. – 2022. – 15. – 487.

176. *Pilling R*. Shapley Value-Based Payment Calculation for Energy Exchange between Micro- and Utility Grids / R. Pilling, S. C. Chang, P. B. Luh // Games. – 2017. – 8(4). – 45.

177. *Bandeiras F*. Application and Challenges of Coalitional Game Theory in Power Systems for Sustainable Energy Trading Communities / F. Bandeiras, A. Gomes, M. Gomes, P. Coelho // Energies. – 2023. – 16. – 8115.

178. *Yarar N. A*. Comprehensive Review Based on the Game Theory with Energy Management and Trading / N. Yarar, Y. Yoldas, S. Bahceci, A. Onen, J. Jung // Energies. – 2024. – 17. – 3749.

179. *Zima-Bockarjova M*. Charging and Discharging Scheduling for Electrical Vehicles Using a Shapley-Value Approach / M. Zima-Bockarjova, A. Sauhats, L. Petrichenko, R. Petrichenko // Energies. – 2020. – 13. – 1160.

180. Fastshap. – URL: https://github.com/bgreenwell/fastshap (дата обращения: 21.09.2024).

181. Shapley.jl. – URL: https://gitlab.com/ExpandingMan/Shapley.jl (дата обращения: 21.09.2024).

182. *Lumberg S. M*. A unified approach to interpreting model predictions / S. M. Lumberg, S.-I. Lee // Proceedings of the 31st International Conference on Neural Information Processing Systems. – 2017. – P. 4768–4777.

183. *Matrenin P. V*. Solar Irradiance Forecasting with Natural Language Processing of Cloud Observations and Interpretation of Results with Modified Shapley Additive Explanations / P. V. Matrenin, V. V. Gamaley, A. I. Khalyasmaa, A. I. Stepanova // Algorithms. – 2024. –17(4). – 150.

184. *Qayyum F*. Explainable AI for Material Property Prediction Based on Energy Cloud: A Shapley-Driven Approach / F. Qayyum, M. A. Khan, D.-H. Kim, H. Ko, G.-A. Ryu // Materials. – 2023. – 16(23). – 7322.

185. *Singh N. K*. LightGBM-, SHAP-, and Correlation-Matrix-Heatmap-Based Approaches for Analyzing Household Energy Data: Towards Electricity Self-Sufficient Houses / N. K. Singh, M. Nagahara // Energies. – 2024. – 17. – 4518.

186. *Gebreyesus Y*. AI for Automating Data Center Operations: Model Explainability in the Data Centre Context Using Shapley Additive Explanations (SHAP) / Y. Gebreyesus, D. Dalton, D. De Chiara, M. Chinnici, A. Chinnici // Electronics. – 2024. – 13. – 1628.

187. *Lu Y*. Machine Learning Models Using SHapley Additive exPlanation for Fire Risk Assessment Mode and Effects Analysis of Stadiums / Y. Lu, X. Fan, Y. Zhang, Y. Wang, X. Jiang // Sensors. – 2023. – 23(4). – 2151.

188. Local Model Interpretation: An Introduction. – URL: https://gilberttanner.com/blog/local-model-interpretation-an-introduction/ (дата обращения: 21.09.2024).

189. SHAP. – URL: https://github.com/shap/shap (дата обращения: 21.09.2024).

190. shapper. – URL: https://modeloriented.github.io/shapper/ (дата обращения: 21.09.2024).

191. fastshap. – URL: https://github.com/bgreenwell/fastshap (дата обращения: 21.09.2024).

192. Explainable AI (XAI): A survey of recents methods, applications and frameworks. – URL: https://theaisummer.com/xai/#layer-wise-relevance-propagation-lrp (дата обращения: 23.09.2024).

193. *Tang D*. Reviewing CAM-Based Deep Explainable Methods in Healthcare / D. Tang, J. Chen, L. Ren, X. Wang, D. Li, H. Zhang // Applied Sciences. – 2024. – 14(10). – 4124.

194. Four Transformer-Based Deep Learning Classifiers Embedded with an Attention U-Net-Based Lung Segmenter and Layer-Wise Relevance Propagation-Based Heatmaps for COVID-19 X-ray Scans / S. Gupta, A. K. Dubey, R. Singh [et al.] // Diagnostics. – 2024. – 14. – 1534.

195. Machine-Learning-Enabled Diagnostics with Improved Visualization of Disease Lesions in Chest X-ray Images / M.F. Rahman, T.-L. Tseng, M. Pokojovy [et. al.] // Diagnostics. – 2024. – 14. – 1699.

196. *Batchluun G*. CAM-CAN: Class activation map-based categorical adversarial network / G. Batchluun, J. Choi, K. R. Park // Expert Systems with Applications. – 2023. – 222. – 119809.

197. *Fu K*. MultiCAM: Multiple Class Activation Mapping for Aircraft Recognition in Remote Sensing Images / K. Fu, W. Dai, Y. Zhang, Z. Wang, M. Yan, M. X. Sun // Remote Sensing. – 2019. – 11. – 544.

198. TorchCAM: class activation explorer. – URL: https://github.com/frgfm/torch-cam (дата обращения: 23.09.2024).

199. torchcam 0.1.1. – URL: https://pypi.org/project/torchcam/0.1.1/ (дата обращения: 23.09.2024).

200. CAM-Visualizer: Class Activation Map Visualization Toolkit. – URL: https://www.intel.com/content/www/us/en/developer/articles/reference-implementation/class-activation-map-visualizer.html (дата обращения: 23.09.2024).

201. Investigate Network Predictions Using Class Activation Mapping. – URL: https://www.mathworks.com/help/deeplearning/ug/investigate-network-predictions-using-class-activation-mapping.html (дата обращения: 23.09.2024).

202. *Selvaraju R. R*. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization / R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra // 2017 IEEE International Conference on Computer Vision (ICCV. – 2017. – P. 618–626.

203. *Kolekar S*. Explainable AI in Scene Understanding for Autonomous Vehicles in Unstructured Traffic Environments on Indian Roads Using the Inception U-Net Model with Grad-CAM Visualization / S. Kolekar, S. Gite, B. Pradhan, A. Alamri // Sensors. – 2022. – 22. – 9677.

204. *Dworak D*. Adaptation of Grad-CAM Method to Neural Network Architecture for LiDAR Pointcloud Object Detection / D. Dworak, J. Baranowski // Energies. – 2022. – 15. – 4681.

205. *Qin C*. Identification of Transient Voltage Stability Weakness in Power Systems Based on Grad-CAM / C. Qin, Y. Jia, Z. Li, G. Li, L. Li, Y. Zhu // 2022 7th International Conference on Power and Renewable Energy (ICPRE). – 2022. – P. 53–57.

206. Classification of PRPD Pattern in Cast- Resin Transformers Using CNN and Implementation of Explainable AI (XAI) With Grad-CAM / H.-S. Kim, J. Jung, R. Hwang [et. al.] // IEEE Access. – 2024. – Vol. 12. – P. 53623–53632.

207. *Senjoba L*. Enhancing Interpretability in Drill Bit Wear Analysis through Explainable Artificial Intelligence: A Grad-CAM Approach / L. Senjoba, H. Ikeda, H. Toriya, T. Adach, Y. Kawamura // Applied Sciences. – 2024. – 14(9). – 3621.

208. *Saleh R.A.A*. AdvancingTire Safety: Explainable Artificial Intelligence-Powered Foreign Object Defect Detection with Xception Networks and Grad-CAM Interpretation / R.A.A. Saleh, F. Al-Areqi, M. Z. Konyar, K. Kaplan, S. Ongir, H. M. Ertunc // Applied Sciences. – 2024. – 14(10) – 4267.

209. *Lin C.-J*. Bearing Fault Diagnosis Using a Grad-CAM-Based Convolutional Neuro-Fuzzy Network / C.-J. Lin, J.-Y. Jhang // Mathematics. – 2021. – 9. – 1502.

210. *Noh E*. Automatic Screening of Bolts with Anti-Loosening Coating Using Grad-CAM and Transfer Learning with Deep Convolutional Neural Networks / E. Noh E, S. Hong // Applied Sciences. – 2022. – 12(4). – 2029.

211. Advanced AI explainability for PyTorch. – URL: https://github.com/jacobgil/pytorch-grad-cam (дата обращения: 23.09.2024).

212. gradCAM. – URL: https://www.mathworks.com/help/deeplearning/ref/gradcam.html (дата обращения: 23.09.2024).

213. *Samek W*. Interpreting the Predictions of Complex ML Models by Layer-wise Relevance Propagation / W. Samek, G. Montavon, A. Binder, S. Lapuschkin, K.-R. Muller // NIPS 2016 Workshop on Interpretable Machine Learning in Complex Systems. – 2016. – P. 1–5.

214. *Dieter T. R*. Evaluation of the Explanatory Power Of Layer-wise Relevance Propagation using Adversarial Examples / T. R. Dieter, H. Zisgen // Neural Process Letters. – 2023. – 55. – P. 8531–8550.

215. *Li G*. A spatial-temporal layer-wise relevance propagation method for improving interpretability and prediction accuracy of LSTM building energy prediction / G. Li, F. Li, C. Xu // Energy and Buildings. – 2022. – Vol. 217. – 112317.

216. *Erdem T*. Layer-Wise Relevance Propagation for Smart-Grid Stability Prediction / T. Erdem, S. Eken // Pattern Recognition and Artificial Intelligence. – 2021. – 2022. – P. 315–328.

217. CNN-LRP: Understanding Convolutional Neural Networks Performance for Target Recognition in SAR Images / B. Zang, L. Ding, Z. Feng [et. al.] // Sensors. – 2021. – 21(13). – 4536.

218. *Du M*. ULAN: A Universal Local Adversarial Network for SAR Target Recognition Based on Layer-Wise Relevance Propagation / M. Du, D. Bi, M. Du, X. Xu, Z. Wu // Remote Sensing. – 2023. – 15. – 21.

219. lrp-pf-auc 0.1.6. – URL: https://pypi.org/project/lrp-pf-auc/ (дата обращения: 23.09.2024).

220. The LRP Toolbox for Artificial Neural Networks (1.3.1). – URL: https://github.com/sebastian-lapuschkin/lrp_toolbox (дата обращения: 23.09.2024).

221. LRP. – URL: https://captum.ai/api/lrp.html (дата обращения: 23.09.2024).

222. *Zhou Y*. Weakly Supervised Instance Segmentation using Class Peak Response / Y. Zhou, Y. Zhu, Q. Ye, Q. Qiu, J. Jiao // Computer Vision and Patter Recognition. – 2018. – P. 3791–3800.

223. *Ou J.-R*. WS-RCNN: Learning to Score Proposals for Weakly Supervised Instance Segmentation / J.-R. Ou, S.-L. Deng, J.-G. Yu // Sensors. – 2021. – 21. – 3475.

224. *Huang X*. Weakly supervised segmentation via instance-aware propagation / X. Huang, Q. Zhu, Y. Liu, S. He // Neurocomputing. – 2021. – Vol. 447. – P. 1–9.

225. DeepWind: Weakly Supervised Localization of Wind Turbines in Satellite Imagery (Papers Track) / S. Zhou, J. Irvin, Z. Wang // Workshop on Tackling Climate Change with Machine Learning. – 2019. – 5.

226. Weakly Supervised Instance Segmentation using Class Peak Response. – URL: https://github.com/ZhouYanzhao/PRM (дата обращения: 23.09.2024).

227. Multi-parametric response map. – URL: https://www.mathworks.com/matlabcentral/fileexchange/52085-multi-parametric-response-map?s_tid=prof_contriblnk (дата обращения: 23.09.2024).

228. *Kumar D*. Explaining the Unexplained: A CLass-Enhanced Attentive Response (CLEAR) Approach to Understanding Deep Neural Networks / D. Kumar, A. Wong, D. W. Taylor // Computer Vision and Patter Recognition Workshop (CVPR-W) on Explainable Computer Vision. – 2017. – P. 36–44.

229. *Kumar D*. Opening the Black Box of Financial AI with CLEAR-Trade: A CLass-Enhanced Attentive Response Approach for Explaining and Visualizing Deep Learning-Driven Stock Market Prediction / D. Kumar, G. W. Taylor, A. Wong // Computer Vision and Pattern Recognition. – 2017. – P. 1–3.

230. *Kumar D*. Discovery Radiomics With CLEAR-DR: Interpretable Computer Aided Diagnosis of Diabetic Retinopathy / D. Kumar, G. W. Taylor, A. Wong // IEEE Access. – 2017. – P. 25891–25896.

231. *Liu G*. Visualizing Feature Maps in Deep Neural Networks using DeepResolve A Genomics Case Study / G. Liu, D. Gifford // Computer Science. – 2017. – P. 1–10.

232. *Gu J*. Wind Farm NWP Data Preprocessing Method Based on t-SNE / J. Gu, Y. Wang, D. Xie, Y. Zhang // Energies. – 2019. – 12. – 3622.

233. *Halladin-Dąbrowska A*. The t-SNE Algorithm as a Tool to Improve the Quality of Reference Data Used in Accurate Mapping of Heterogeneous Non-Forest Vegetation / A. Halladin-Dąbrowska, A. Kania, D. Kopec // Remote Sensing. – 2020. – 12(1). – 39.

234. TSNE. – URL: https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html (дата обращения: 23.09.2024).

235. *Aubry M*. Understanding deep features with computer-generated imagery / M. Aubry, B. Russell // Computer Vision and Pattern Recognition. – 2015. – P. 2875–2883.

236. NN-PCA. – URL: https://github.com/manasgaur/NN-PCA (дата обращения: 23.09.2024).

237. neuronPCA. – URL: https://www.mathworks.com/help/deeplearning/ref/neuronpca.html (дата обращения: 23.09.2024).

238. *Karpathy A.* Visualizing and Understanding Recurrent Networks / A. Karpathy, J. Johnson, L. Fei-Fei // ICLR 2016. – 2016. – P. 1–12.

239. *Lu J.* Hierarchical Question-Image Co-Attention for Visual Question Answering / J. Lu, J. Yang, D. Batra, D. Parikh // 30th Conference on Neural Information Processing Systems (NIPS 2016). – 2016. – P. 1–11.

240. *Шавкунова И. С.* «Решение» в психологии и экономике: интеграция научных подходов // Психология в экономике и управлении. – 2017. – Т. 9. – № 1. – С. 24–33.

241. *Карпенко Л. А*. Краткий психологический словарь / под ред. А. В. Петровского, М. Г. Ярошевского. – Москва : Феникс, 1998. – С. 512.

242. *Прохоров А. М.* Большая советская энциклопедия: в 30 т. / А. М. Прохоров // Советская энциклопедия. – 1976. – Т. 26. – 453 с.

243. *Мандрикова Е. Ю.* Виды личностного выбора и их индивидуально-психологические предпосылки : автореф. дис. … канд. психол. наук / Е. Ю. Мандрикова. – Москва, 2006. – 28 с.

244. *Стоева Н. И.* Управленческие решения и методы их принятия / Н. И. Стоева // Журнал социологических исследований. – 2020. – Т. 5. – № 3. – С. 34–36.

245. *Мескон М.* Основы менеджмента / М. Мескон, М. Альберт, Ф. Хедоури. – Москва : Дело, 2006. – С. 720.

246. *Brest P.* Problem solving, decision making, and professional judgment: a guide for lawyers and policymakers / P. Brest, L. H. Krieger. – Oxford : Oxford University Press, 2012. – P. 696.

НАУЧНОЕ ИЗДАНИЕ

Хальясмаа Кристина Ильмаровна
Степанова Алина Игоревна
Зиновьева Елена Леонидовна
Матренин Павел Викторович
Хальясмаа Александра Ильмаровна
Ерошенко Станислав Андреевич

ПРАВОВЫЕ АСПЕКТЫ ПРИМЕНЕНИЯ ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМ И МЕТОДЫ ИНТЕРПРЕТАЦИИ ИХ РАБОТЫ

Монография

Редактор *И.Л. Кескевич*
Выпускающий редактор *И.П. Брованова*
Корректор *И.Е. Семенова*
Художественный редактор *А.В. Ладыжская*
Компьютерная верстка *С.И. Ткачева*

Подписано в печать 19.12.2024
Формат 60 × 90 1/16. Бумага офсетная
Уч.-изд. л. 12,0. Печ. л. 12,0. Тираж 100 экз.
Изд. № 146. Заказ № 10

Налоговая льгота – Общероссийский классификатор продукции
Издание соответствует коду 95 3000 ОК 005-93 (ОКП)

Издательство Новосибирского государственного
технического университета
630073, г. Новосибирск, пр. К. Маркса, 20
Тел. (383) 346-31-87
E-mail: office@publish.nstu.ru