DOI: 10.30727/0235-1188-2022-65-1-72-90
Оригинальная исследовательская статья
Original research article

Субъектность объяснимого искусственного интеллекта*

А.Н. Райков
Институт проблем управления РАН, Москва, Россия,
МИРЭА – Российский технологический университет, Москва, Россия

Аннотация

Статья посвящена определению путей развития способности систем искусственного интеллекта (ИИ) давать объяснения своим выводам. Эта тема не нова, однако именно сейчас нарастание сложности этих систем заставляет ученых активизировать исследования в этом направлении. Современные нейронные сети содержат сотни слоев нейронов, число параметров этих сетей достигает триллионов, генетические алгоритмы порождают тысячи поколений решений, семантика моделей ИИ усложняется, достигая квантового и нелокального уровней. Ведущие компании мира вкладывают огромные средства в создание объяснимого ИИ (*Explainable AI, XAI*), однако результат пока остается неудовлетворительным – человек зачастую не может понять «объяснений» ИИ, потому что последний принимает решения иначе, чем человек, а возможно, потому что получить хорошее объяснение невозможно в рамках классической парадигмы ИИ. С похожей проблемой ИИ столкнулся лет 40 назад, когда экспертные системы содержали всего несколько сотен логических правил-продукций. Проблема тогда была решена за счет усложнения логики и построения дополнительных баз знаний для объяснения выводов, даваемых ИИ. Сейчас, по-видимому, нужны иные подходы, прежде всего учитывающие внешнее окружение и субъектность систем ИИ. Настоящая работа акцентирует внимание на разрешении этой проблемы через погружение моделей ИИ в социально-экономическую среду, построение онтологий этой среды, а также учет профиля пользователя и формирование условий для целенаправленной сходимости решений и выводов ИИ к понятным пользователю целям.

Ключевые слова: философия искусственного интеллекта, онтологии, эпистемология, причинность, рефлексивно-активные системы, субъективная реальность.

* Работа поддержана Российским научным фондом (РНФ), грант № 21-18-00184 «Социогуманитарные основания критериев оценки инноваций, использующих цифровые технологии и искусственный интеллект».

Райков Александр Николаевич – доктор технических наук, ведущий научный сотрудник Института проблем управления РАН, профессор МИРЭА – Российского технологического университета.
ANRaikov@mail.ru
https://orcid.org/0000-0002-6726-9616

Для цитирования: *Райков А.Н.* Субъектность объяснимого искусственного интеллекта // Философские науки. 2022. Т. 65. № 1. С. 72–90. DOI: 10.30727/0235-1188-2022-65-1-72-90

Subjectivity of Explainable Artificial Intelligence*

A.N. Raikov
Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia,
MIREA – Russian Technological University, Moscow, Russia

Abstract

The article addresses the problem of identifying methods to develop the ability of artificial intelligence (AI) systems to provide explanations for their findings. This issue is not new, but, nowadays, the increasing complexity of AI systems is forcing scientists to intensify research in this direction. Modern neural networks contain hundreds of layers of neurons. The number of parameters of these networks reaches trillions, genetic algorithms generate thousands of generations of solutions, and the semantics of AI models become more complicated, going to the quantum and non-local levels. The world's leading companies are investing heavily in creating explainable AI (XAI). However, the result is still unsatisfactory: a person often cannot understand the "explanations" of AI because the latter makes decisions differently than a person, and perhaps because a good explanation is impossible within the framework of the classical AI paradigm. AI faced a similar problem 40 years ago when expert systems contained only a few hundred logical production rules. The problem was then solved by complicating the logic and building added knowledge bases to explain the conclusions given by AI. At present, other approaches are needed, primarily those that consider the external environment and the subjectivity of AI systems. This work focuses on solving this problem by immersing AI models in the social and economic environment, building ontologies of this environment, taking into

* The work was supported by the Russian Science Foundation, grant no. 21-18-00184 "Social and humanitarian foundations for evaluation criteria for innovations based on digital technologies and artificial intelligence."

account a user profile and creating conditions for purposeful convergence of AI solutions and conclusions to user-friendly goals.

Keywords: philosophy of artificial intelligence, ontologies, epistemology, causality, reflexive-active systems, subjective reality.

Alexander N. Raikov – D.Sc. in Technology, Leading Research Fellow, Institute of Control Sciences, Russian Academy of Sciences; Professor, MIREA – Russian Technological University.
ANRaikov@mail.ru
https://orcid.org/0000-0002-6726-9616

For citation: Raikov A.N. (2022) Subjectivity of Explainable Artificial Intelligence. *Russian Journal of Philosophical Sciences = Filosofskie nauki.* Vol. 65, no. 1, pp. 72–90. DOI: 10.30727/0235-1188-2022-65-1-72-90

Введение

Проблема доверия выводам систем ИИ появилась почти одновременно с рождением термина «искусственный интеллект» (ИИ), достаточно вспомнить работу Д. Пойи «Математика и правдоподобные рассуждения» [Polya 1954]. Уже тогда разработчиков систем ИИ интересовали правдоподобные рассуждения и достоверный вывод [Вагин и др. 2004]. Успешного разрешения проблемы «черного ящика ИИ», как она сейчас иногда обозначается, научно-технический мир достиг лет 40 назад, когда при создании средств ИИ основное внимание стало уделяться созданию логических экспертных систем. Тогда система ИИ на основе довольно замкнутой базы знаний, состоящей из 300–400 правил-продукций, выводила результат, но не могла его доходчиво объяснить. Такие системы не вызывали доверие, плохо продавались, разработчики тратили десятки миллионов долларов на гонорар инженерам по знаниям (когнитологам), предметным экспертам и программистам. С этой проблемой справились путем усложнения логик и создания дополнительных баз знаний, помогающих обобщить результат вывода и дать сносное объяснение.

За последние годы сложность ИИ резко возросла, число слоев в нейронных сетях достигает нескольких сотен, число параметров – триллионы, эволюционные методы вычислений порождают тысячи поколений (множеств) решений, объемы больших данных для обучения систем ИИ стремительно растут и пр. Меняются взгляды на семантику моделей систем ИИ, ее экспликация уходит на неформализуемый когнитивный уровень, учитывает квантовые

и релятивистские эффекты [Raikov 2021]. При этом в рамках развития ИИ возникают принципиально новые проблемы: обостряются угрозы в сфере кибербезопасности, создаются квантовые версии вычислителей, возрастает опасность злонамеренного и неэтичного использования ИИ, системы ИИ становятся решающим фактором производства конкурентоспособной продукции и др.

В таких условиях ведущие компании мира (*DARPA, Google, Tencent Research Institute* и др.) увеличивают инвестиции в создание объяснимого ИИ (*XAI, Explainable AI*). Ключевыми идеями исследований сейчас являются: послойный анализ работы ИИ, испытанные ранее продукционные (по сути, логические) схемы вывода, построения на графах знаний и пр. Среди возможных средств порождения объяснений в последнее время все больше используются графы знаний. Разветвленные связи в графах знаний делают их полезными для объяснений выводов систем ИИ. Например, эти графы строятся путем подсчета расстояний между целевым профилем пользователя и элементом в графе знаний, от которого зависит развитие ситуации. Авторы [Liu, Balsubramani, Zou 2019] предлагают генерировать объяснения по графам, сопровождая процесс обучением с подкреплением. В работе [Madry et al. 2017] предпринята попытка сделать так, чтобы нейронная сеть давала объяснение путем выявления причин взаимодействия пользователя с системой. Однако результаты пока не считаются удовлетворительными. Ответ на каждый новый вопрос по созданию объяснимого ИИ, делающий вроде бы «черный ящик» искусственной нейронной сети «белым», на самом деле обнаруживает внутри него множество других «черных ящиков». При этом экспоненциально растет число необходимых для получения объяснений вычислений, что зачастую заставляет отказываться от построения компоненты объяснения и ее включения в систему ИИ.

Внимание исследователей все больше привлекает учет субъектного фактора создания и использования систем ИИ, помещение этих систем в саморазвивающуюся рефлексивно-активную среду, построение онтологий контекста использования [Дубровский 2021; Лепский 2021].

В настоящей работе предлагается поменять акценты в парадигме создания *XAI,* обратить внимание на субъектные аспекты, внешнюю и глубинную стороны традиционных моделей построения ИИ. При этом к внешней стороне может относиться социально-экономическая и субъективная реальность, а к глу-

бинной – флюктуирующая, квантовая и релятивистская природа сознания человека.

XAI: состояние вопроса

Общепризнанного определения и взгляда на реализацию *XAI,* конечно, нет и, по-видимому, быть не может. Вместе с тем потребность в *XAI* нарастает, а значит следует развивать методологию разрешения проблемы. Имеются множественные интенции и предложения ученых и инженеров, которые сопровождают разработку *XAI,* например, объяснительный ресурс ИИ:

• зависит не столько от сложности логики и алгоритмов работы системы ИИ, сколько от пользователя и внешнего контекста, включая социум, экономику, пользователя, целенаправленность и др. [Лепский 2021];

• обеспечивает создание моделей ИИ, которые могут объяснить свои выводы, сохраняя нужную точность прогнозирования, обеспечивая пользователям понимание, доверие и управление новыми объектами [Chen et al. 2020];

• должен гарантировать, что решения и любые данные, обеспечивающие их, могут быть объяснены человеком непрофессионалом [Wang et al. 2020];

• направлен на послойную расшифровку работы системы ИИ для создания моделей и методов, которые одновременно точны и дают удовлетворительное объяснение [Veličković et al. 2019] и др.

До недавнего времени основное внимание построению компонент объяснения в системах ИИ уделялось естественно-научным, инженерным и логико-технологическим компонентам. В результате можно выделить следующие основные технологические аспекты XAI, которые определяют текущую реализацию прикладных исследований:

• учет внешнего окружения;

• верифицируемость объяснения на других примерах и моделях;

• зависимость объяснения от объяснительной модели ИИ, например, направленной на визуализацию, построение графики, использование примеров;

• принятие во внимание характеристик субъекта, нуждающегося в объяснении.

Остановимся сначала на первых двух из перечисленных аспектов. Они характеризуют техническую сторону цифровой среды

и практически не затрагивают природу субъектности, сознания, эмоций, духовную сферу. Так, в ряде работ обсуждаются вопросы так называемого обволакивающего интеллекта [Шалова 2016], подразумеваемого в значении окружающего интеллекта (*Ambient Intelligence, AmI*). *AmI* относится к цифровой среде, которая чувствительна и реагирует на присутствие людей. Окружающий интеллект разрабатывается с конца 1990-х годов как проекция электроники, телекоммуникаций и вычислений на будущее [True Visions… 2006; Nolin 2016]. Такие системы сейчас разрабатываются для обеспечения согласованной работы различных технических устройств, чтобы поддерживать людей в их повседневной деятельности, решения задач интуитивно понятным способом, используя информацию и ИИ, которые скрыты в сети, соединяющей эти устройства (например, Интернет вещей).

Построенные объяснения могут быть как релевантными, формализованными, так и смысловыми, неформализованными. Первые лежат в логической плоскости: само объяснение можно проверить, например, послойно анализируя процесс вывода построить логические причинно-следственные цепочки получения вывода. На систему объяснения в этом случае принципиальным образом влияют данные, которые используются для обучения и адаптации системы ИИ [Lin, Hung, Huang 2021; Leavy at al. 2020; Leavy, Siapera, O'Sullivan 2021]. Вторые могут быть оценены только через семантическую интерпретацию пользователем полученного от системы ИИ объяснения, например, пользователь может быть не согласен с результатом, поскольку у него есть свое понимание вопроса. Для второго случая данных недостаточно, поскольку данные представляют собой формализованную компоненту системы ИИ, только косвенно затрагивающую субъектные аспекты системы.

Вместе с тем проблема *XAI* – это не только логическое и технологическое предоставление удовлетворительных объяснений, она тесно связана с неформализуемыми аспектами феномена сознания, субъективной реальности [Дубровский 2021], противоречивого эпистемологического и этического выбора [Kaul 2022]. Ее разрешение должно учитывать способы, которыми люди достигают договоренностей в сферах политических, экономических, социальных отношений, а также соглашаются с тем, чтобы ими впоследствии руководствоваться.

Ссылаясь на требование прозрачности при логическом принятии решений, исследователи отмечают потребность понять, что

такое объяснение вообще [Rauber, Trasarti, Gianotti 2019, 10–11]. Понимание объяснения далеко не однозначно и варьируется в зависимости от ситуации и дисциплины. Довольно емкий обзор по теме объяснения [Mueller et al. 2019] представляет собой широкий набор ссылок на работы по *XAI* в различных дисциплинах. В обзоре анализируются ключевые концепции *XAI* и различные виды систем ИИ (экспертные системы, системы рассуждений на основе прецедентов, системы машинного обучения, байесовские классификаторы, статистические модели и деревья решений). Рассматривается разнообразие приложений: классификации жестов, изображений, текста; отладка программ; синтез музыкальных рекомендаций; финансовый учет; стратегические игры; формирование команд; роботизация; диагностика болезней; построение гипотез, касающихся отношения объяснения к фундаментальным когнитивным процессам и пр. По всей видимости, обзор не столько определяет пути развития *XAI*, сколько демонстрирует неочевидность дальнейшего развития темы.

В большом обзоре работ по *XAI* [Adadi, Berrada 2018] выявляются коммерческие, этические и нормативные причины, по которым необходимы объяснения. Рассматриваются различные цели объяснений, например, чтобы оправдать, контролировать, улучшать и открывать. Обсуждаются методы объяснений с точки зрения локального и глобального, внутреннего и апостериорного, модельно-специфического и модельно-независимого (см. в частности, таблицу, обобщающую ключевые концепции XAI [Adadi, Berrada 2018, 52141]). Облако слов *XAI,* которое они предоставляют [Adadi, Berrada 2018, 52140], имеет интерпретируемый язык машинного обучения (*ML*) и объяснимый ИИ как наиболее известные термины в литературе.

Работа [Arrieta et al. 2020] представляет обзор концепций, таксономий, возможностей и проблем XAI по таким направлениям, как классификация моделей ML в зависимости от их уровня объяснимости [Arrieta et al. 2020, 90], таксономия литературы и тенденций в сфере объяснимости для моделей машинного обучения [Arrieta et al. 2020, 93]. Авторы отмечают, что интерпретируемость «черного ящика» машинного обучения важна для обеспечения беспристрастности при принятии решений, устойчивости к злонамеренным возмущениям, а также в качестве гарантии того, что только значимые переменные определяют результат [Arrieta et al. 2020, 83]. Авторы пролагают, что

феномен понятности является наиболее важным атрибутом *XAI* [Arrieta et al. 2020, 84–85].

Вместе с тем ответ на вопрос о потребности в объяснении применительно к системам ИИ не является однозначным. Трактовки объяснимости иногда вызывают скептицизм в отношении ее полезности. Предстоит выяснить:

• Может ли высокая прозрачность работы системы ИИ привести к информационной перегрузке?

• Может ли визуализация привести к чрезмерному доверию или неправильному чтению?

• Могут ли системы машинного обучения на самом деле предоставлять хорошие обоснования на естественном языке?

• Действительно ли разные люди в любом случае нуждаются в разных объяснениях? [Heaven 2020].

Ставятся вопросы, связанные с определением роли объяснения при оценке ответственности в законодательстве и судебной практике, отмечается потребность достигнуть компромисса между полезностью и стоимостью объяснений, сложностью и вычислительными ресурсами. Ведь любая уточняющая информация в виде объяснения может быть представлена в виде набора абстрактных причин или оправданий некоего результата, а не описания процесса принятия решения в целом [Doshi-Velez, Kortz 2017, 4]. Для них генерация объяснений – это вопрос дизайна системы, и они предлагают рассматривать системы объяснений отдельно от систем ИИ [Doshi-Velez, Kortz 2017, 16–17], чтобы создать возможности для отраслей, где функционируют системы объяснений в терминах, интерпретируемых человеком, без ущерба для точности исходного предиктора.

В обзоре, посвященном принципам объяснения и человеко-машинным системам ИИ, делается акцент на необходимости ориентации на человека, у которого есть ожидания в отношении объяснения, при создании системы объяснения [Mueller et al. 2020]. Работа в области психологии, посвященная познанию и предубеждениям, использует данные человеческого мышления, чтобы аргументировать их вклад в интерпретируемые модели сложных систем ИИ [Byrne 2019, 6280].

В итоге можно отметить в целом позитивное отношение к дальнейшему развитию *XAI*. Тема постоянно экстраполируется во многие смежные области, включая социально-экономические. Это было ожидаемо, поскольку большинство исследователей в

области ИИ имеют опыт работы в области вычислений, математики, технической кибернетики, естественных наук и, возможно, психологии или философии. Как известно, характер знаний существенно зависит от того, кто является его потребителем, в какой среде оно сформировано и как ведется. Поэтому природа и структура объяснений строятся с учетом не столько вероятности выбора, который делает система ИИ, сколько внешних, в частности социально-экономических, причин его порождения, а также ожиданий и переживаний пользователя системы ИИ.

Каузальность (причинность) как базис объяснения

Особое место в построении компонент объяснения занимает тезис каузальности – причинно-следственной связи событий. Причинность, по-видимому, является наиболее фундаментальным явлением, отражающим всеобщую связь и единство во Вселенной. Она связывает мысли с действиями, а действия с последствиями, движение планет с образующими их атомами, успех лечения людей от используемой методики и др. На частый вопрос «Почему?» помогают ответить причинно-следственные связи между событиями. Подобные связи объясняются научными законами и закономерностями.

Причинность характеризуется частым совместным появлением событий. При этом далеко не всегда по частоте совместного появления событий можно однозначно судить о наличии между ними причинно-следственной связи. Например, традиционный ИИ не может точно преобразовать корреляцию событий в такую связь и не может обеспечить высокий уровень прозрачности или доверия к результатам вывода системы ИИ. Однако именно корреляция и частотность событий, встречающихся в больших массивах информации, являются основой для построения явных и выявления неявных причинно-следственных связей между событиями.

В чем причина того, что коляска движется, если ее толкает человек, или в космическом пространстве звезды вращаются вокруг центра масс и образуются кратные звезды? Философы исследуют этот вопрос на протяжении тысячелетий, по крайней мере со времен Платона и Аристотеля. Проблема рассматривается абстрактно и конкретно, с разных сторон. Число каузальных утверждений явно нарастает, но вопрос остается.

Наиболее распространенное мнение, что в основе причинно-следственной связи лежит регулярность: одно событие (вещь)

постоянно связано с другим. Классическая и наиболее распространенная тока зрения восходит к Дэвиду Юму. Он считал, что если события *A* и *B* постоянно появляются вместе и *A* происходит до *B*, то этого все же недостаточно для того, чтобы заключить, что *A* является причиной *B*. Причина появляется тогда, когда *A* и *B* также должны быть смежными, быть рядом друг с другом, то есть иметь пространственную связь.

Вместе с тем понятие смежности нельзя не подвергнуть сомнению. Например, из фундаментальной физики известен эффект квантовой запутанности (энтэлгмента). Можно считать доказанным, что две частицы, находящиеся в разных концах почти бесконечной Вселенной, могут быть связаны таким образом, что изменение квантового состояния одной из них приводит к мгновенному и независимому от расстояния изменению состояния другой [Einstein, Podolsky, Rosen 1935; The BIG Bell… 2018]. Это, однако, происходит в нарушение законов теории относительности: причинно-следственная связь движется быстрее скорости света. И остается вопрос, является ли изменение состояния одной из частиц случаем реальной причинности. Пока остается непонятной природа явления и причины флюктуации частицы, состояние которой изменилось в первую, вторую очередь или совместно с другой. Появился некий научный парадокс, и, как следствие, временной приоритет и смежность по Юму могут быть оспорены.

Понятие причинности претерпевает изменения со временем и в различных контекстах. Ранее причинность больше мыслились как действия агента-человека или робота, сейчас понятие причинности трансформируется в причинность, которая становится предметом объяснения, свободного от эмоциональной окраски, например восхищения, похвалы или осуждения.

Вопросы причинности классифицируются по нескольким направлениям в зависимости от подхода: плюрализм, примитивизм, физикализм и др. Признается, что переход к плюрализму является просто признанием поражения всех иных, скорее специфичных направлений. Например, плюралист использует различные теории и понимает причинность как множество разных вещей или событий. Причинно-следственная связь в отдельных теориях и подходах не поддается анализу. Возможно, это происходит из-за ограниченности выбранного подхода [Mumford, Anjium 2013].

Вместе с тем в аналитической философии, например с учетом взглядов Локка, что-то должно быть взято за основу

[Mumford, Anjium 2013, 88–89], хотя любую часть аналитического подхода можно подвергнуть сомнению. Если бы в мире была бесконечная сложность, например, доходящая до структуризации фотонов, нейтрино и кварков, то в конце концов не было бы ничего базового. Но это неизвестно наверняка. Возможно, существует базовый, неразложимый далее элемент природы, который мы не знаем, но способны в будущем познать. Этот взгляд может подойти для отмеченного выше примитивистского подхода. Некоторые примитивы, вероятно, будут правильными, так что нет ничего явно неправильного в том, чтобы использовать примитивистский подход. Возможно, это больше, чем просто возможный вариант, особенно для частных утилитарных соображений.

Тема объяснимости выводов ИИ явно шире и выходит за рамки вопроса каузальности, охватывает субъективную реальность в саморазвивающихся междисциплинарных полисубъектных (рефлексивно-активных) средах [Лепский 2021]. Бытие субъектов в таких средах может быть задано системой онтологий, которая обеспечивает сборку пребывающих в среде субъектов в целое. Разработана и опробована система онтологий, в которую входят онтологии обеспечения жизнедеятельности, преодоления точек разрыва, стратегического целеполагания, разработки стратегий и проектов, внедрения и инновационного обеспечения стратегий и проектов [Lepskiy 2019; Лепский 2021]. Показано, что постановка задач для ИИ должна осуществляться с использованием системы онтологий бытия субъектов и в контексте поддержки рефлексивной активности субъектов.

Таким образом, в контексте создания *XAI* можно думать о причинности как об одной из фундаментальных сил и элементов Вселенной. Причинность – это то, что удерживает объекты вместе посредством атомарных и молекулярных связей. Она производит изменение одной вещи с помощью другой. Это придает любому действию значимость. И тогда есть ли основания думать, что мы можем объяснить причинность в некаузальных терминах?

Дисциплинарные и предметные особенности объяснения

Отметим особый подход экономических теорий к феномену объяснения [Kaul 2022]. Экономисты иногда считают объяснения, даваемые экономическими теориями и моделями, формой, которую принимает теория, и воспринимают это как само собой разумеющееся. Большинству людей, которые сталкиваются

с объяснениями, предлагаемыми этими теориями, сама идея объяснения может показаться нелогичной. Экономисты строят модели, которые направлены именно на абстрактное объяснение экономической ситуации и поэтому далеко не всегда могут быть опровергнуты эмпирическими данными. Вместе с тем объяснительный приоритет модельного экономического построения работает совсем не только для получения количественных характеристик динамики и прогноза развития ситуации, которая, по сути, в экономике наиболее важна. В указанной работе [Kaul 2022] отмечаются разнообразные взгляды экономистов на получение ответов на вопросы относительно феномен объяснения:

• примирение противоборствующих сторон относительно роли ценностных или нормативных различий;

• признание необходимости телеологии в объяснении при неспособности эмпиризма определять человеческие цели;

• различие и развитие природы телеологических и каузальных объяснений в конкретных областях экономики; использование модели рационального выбора;

• объяснение событий в различных подотраслях экономики;

• рассмотрение дедуктивного, индуктивное и абдуктивного рассуждения в неоклассической экономике и др.

Все чаще встречаются работы, пытающиеся объяснить экономические явления с помощью инструментов, которые вроде бы далеки от экономики, например с помощью методов квантовой физики [Orrell, Houshmand 2022; Райков 2009]. Однако, по-видимому, вопросы адекватности применения методов физики к экономическим системам все еще нуждаются в осмыслении. Таким образом, экономика имеет очень своеобразный подход к объяснению по отношению к ее регулярной практике построения экономической теории с целью обеспечить достоверность и доверие к экспертизе экономической ситуации, и его нельзя признать удовлетворительным для использования в полной мере при создании *XAI*.

Очевидно, что люди относятся к сверхсложным системам, выходящим за границы, диктуемые технической кибернетикой. В отличие от естественных наук, где формулы, закономерности, физические законы и пр. могут быть получены путем наблюдения, измерены и воспроизведены разными способами, в субъектной рефлексивно-активной среде человеческое поведение внутри коллективов и между коллективами не поддается пря-

мому наблюдению, измерению и формализованному описанию [Лепский 2021], и, как следствие, однозначному объяснению. Поэтому для объяснения явлений привлекаются упомянутые выше онтологии.

В менеджменте, стратегическом планировании, где решение проблем осуществляется с учетом идентификации сотен факторов, материальных и субъектных, для объяснения проблемной ситуации используются методы стратегического анализа и когнитивного моделирования [Raikov 2022]. Как и в подходе с онтологиями, решаемая проблема коллективно декомпозируется на конечное число целей, факторов, препятствий, ограничений, функций, задач и пр., они выстраиваются специальным образом, например с помощью метода анализа иерархий, затем результат загружается в компьютер и проводится моделирование. Результат моделирования не всегда поддается объяснению, однако сам аналитический процесс, состоящий из множества шагов, делается под непосредственным контролем команды пользователей системы ИИ и должен вызывать доверие.

Предметные особенности проблемы объяснения неизменно проявляются в повседневной и деловой среде. Возьмем актуальную задачу борьбы с вредными насекомыми, например с кукурузным мотыльком. От него страдают порядка 250 растений, а урожай может сократиться на 25% и более. По всей видимости, системы ИИ могли бы помочь более успешно справиться с решением проблемы, поскольку обладают способностью собирать, накапливать и анализировать большие объемы данных, делать прогнозы. Однако ИИ далеко не всегда может давать объяснения своим выводам, что резко снижает доверие к ним и может сделать применение ИИ нерентабельным. Достаточно отметить сложность и скрытность поведения мотылька, которое пока не поддается успешному моделированию и прогнозированию [Абросимов, Райков 2022].

Приведенный пример с мотыльком видится много более простым, чем проблема объяснения с учетом субъектных аспектов. Вместе с тем в разрешении проблемы с мотыльком эти аспекты тоже имеют место, поскольку борьбу с вредителем растений запускают люди, а куколка прячется в стволе растения. Процесс едва ли поддается наблюдению, и, как следствие, охвату системами сбора информации, а затем их аналитической обработке, прогнозированию и объяснению с помощью систем ИИ.

Заключение

Объяснительная функция приобретает все более важное значение в дальнейшем развитии и применении систем ИИ, что вызвано их усложнением и, как следствие, снижением доверия к генерируемым ими рекомендациям.

Помимо технологической составляющей, решающее значение для повышения эффективности систем ИИ приобретает учет субъективной реальности, погружение системы ИИ в полисубъектную рефлексивно-активную среду.

Росту доверия к выводам систем ИИ способствует непосредственное включение людей в процесс получения результата. Для такого включения используются специальные подходы, основанные на построении онтологий и структуризации проблем для создания условий их целенаправленного разрешения.

ЦИТИРУЕМАЯ ЛИТЕРАТУРА

Абросимов, Райков 2022 – *Абросимов В.К., Райков А.Н.* Интеллектуальные сельскохозяйственные роботы. – М.: Карьера Пресс, 2022.

Вагин и др. 2004 – *Вагин В.Н., Головина Е.Ю., Загорянская А.А., Фомина М.В.* Достоверный и правдоподобный вывод в интеллектуальных системах / под ред. В.Н. Вагина и Д.А. Поспелова. – М.: Физматлит, 2004.

Дубровский 2021 – *Дубровский Д.И.* Задача создания Общего искусственного интеллекта и проблема сознания // Философские науки. 2021. Т. 64. № 1. С. 13–44.

Лепский 2021 – *Лепский В.Е.* Искусственный интеллект в субъектных парадигмах управления // Философские науки. 2021. Т. 64. № 1. С. 88–101.

Райков 2009 – *Райков А.Н.* Протуберанцы макроэкономики // Экономические стратегии. 2009. № 7. С. 42–49.

Шалова 2016 – *Шалова С.Х.* Обзор и анализ исследований в области систем обволакивающего интеллекта // Инженерный вестник Дона. 2016. № 4 (43). С. 125.

Adadi, Berrada 2018 – *Adadi A., Berrada M.* Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI) // IEEE Access. 2018. Vol. 6. P. 52138–52160.

Arrieta et al. 2020 – *Arrieta A.B., Díaz-Rodríguez N., Del Ser J., Bennetot A., Tabik S., Barbado A., García S., Gil-López S., Molina D., Benjamins R., Chatila R.* Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges towards Responsible AI // Information Fusion. 2020. Vol. 58. P. 82–115.

Byrne 2019 – *Byrne R.M.J.* Counterfactuals in Explaining Artificial Intelligence (XAI): Evidence from Human Reasoning // Proceedings of the

Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19). – California: *International Joint Conferences on Artificial Intelligence,* 2019. P. 6276–6282.

Chen et al. 2020 – *Chen M., Wei Z., Huang Z., Ding B., & Li Y.* Simple and Deep Graph Convolutional Networks // Proceedings of Machine Learning Research. 2020. Vol. 119. P. 1725–1735.

Doshi-Velez, Kortz 2017 – *Doshi-Velez F., Kortz M.* Accountability of AI Under the Law: The Role of Explanation // Berkman Klein Center Working Group on Explanation and the Law, Berkman Klein Center for Internet & Society Working Paper. 2017. – URL: http://nrs.harvard.edu/urn-3:HUL.InstRepos:34372584

Einstein, Podolsky, Rosen 1935 – *Einstein A., Podolsky B., Rosen N.* Can Quantum-Mechanical Description of Physical Reality Be Considered Complete? // Physical Review. 1935. Vol. 47. No. 10. P. 777–780.

Heaven 2020 – *Heaven W.D.* Why Asking an AI to Explain Itself Can Make Things Worse // MIT Technology Review. 2020. January 29. – URL: https://www.technologyreview.com/2020/01/29/304857/why-asking-an-ai-to-explainitself-can-make-things-worse/

Kaul 2022 – *Kaul N.* 3Es for Al: Economics, Explanation, Epistemology // Frontiers in Artificial Intelligence. Vol. 5. Article 833238.

Leavy et al. 2020 – *Leavy S., Meaney G., Wade K., Greene D.* Mitigating Gender Bias in Machine Learning Sata Sets // International Workshop on Algorithmic Bias in Search and Recommendation. – Cham: Springer. P. 12–26.

Leavy, Siapera, O'Sullivan 2021 – *Leavy S., Siapera E., O'Sullivan B.* Ethical Data Curation for AI: An Approach based on Feminist Epistemology and Critical Theories of Race // AIES'21: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. – New York: The Association for Computing Machinery, 2021. P. 695–703.

Lepskiy 2018 – *Lepskiy V.* Evolution of Cybernetics: Philosophical and Methodological Analysis // Kybernetes. 2018. Vol. 47. No. 2. P. 249–261.

Lin, Hung, Huang 2021 – *Lin Y.T., Hung T.W., Huang L.T.L.* Engineering Equity: How AI Can Help Reduce the Harm of Implicit Bias // Philosophy and Technology. 2021. Vol. 34. No. 1. P. 65–90.

Liu, Balsubramani, Zou 2019 – *Liu R., Balsubramani A., Zou J.* Learning Transport cost from subset correspondence // arXiv. 2019. – URL: https://arxiv.org/pdf/1909.13203.pdf

Madry et al. 2017 – *Madry A., Makelov A., Schmidt L., Tsipras D., Vladu A.* Towards Deep Learning Models Resistant to Adversarial Attacks // arXiv. 2017. – URL: https://arxiv.org/pdf/1706.06083.pdf

Mueller et al. 2019 – *Mueller S.T., Hoffman R.R., Clancey W, Klein G.* Explanation in Human-AI systems: A Literature Meta-Review Synopsis of Key Ideas and Publications and Bibliography for Explainable AI // DARPA XAI Literature Review. 2019. – URL: https://arxiv.org/pdf/1902.01876.pdf

Mueller et al. 2020 – *Mueller S. T., Veinott E. S., Hoffman R.R., Klein G., Alam L., Mamun T., Clancey W.J.* Principles of Explanation in Human-AI Systems // Association for the Advancement of Artificial Intelligence. 2020. – URL: https://arxiv.org/pdf/2102/2102.04972.pdf

Mumford, Anjium 2013 – *Mumford S., Anjium R.L.* Causation. A Very Short Introduction. – Oxford: Oxford University Press, 2013.

Nolin 2016 – *Nolin J., Olson N.* The Internet of Things and Convenience // Internet Research. 2016. Vol. 26. No. 2. P. 360–376.

Orrell, Houshmand 2022 – *Orrell D., Houshmand M.* Quantum Propensity in Economics // Frontiers in Artificial Intelligence. 2022. Vol. 4. Art. 772294.

Polya 1954 – *Polya G.* Mathematics and Plausible Reasoning. – Princeton, NJ: Princeton University Press, 1954.

Raikov 2021 – *Raikov A.* Cognitive Semantics of Artificial Intelligence: A New Perspective. – Singapore: Springer, 2021.

Raikov 2022 – *Raikov A.* Automating Cognitive Modelling Considering Non-Formalisable Semantics // Intelligent Sustainable Systems (Lecture Notes in Networks and Systems. Vol. 334). – Singapore: Springer, 2022.

Rauber, Trasarti, Gianotti 2019 – *Rauber A., Trasarti R., Gianotti F.* Transparency in Algorithmic Decision Making // ERCIM News. 2019. Vol. 116. P. 10–11. – URL: https://ercim-news.ercim.eu/en116/special/transparency-in-algorithmic-decision-making-introduction-to-the-special-theme

The BIG Bell… 2018 – *The BIG Bell Test Collaboration.* Challenging Local Realism with Human Choices // Nature. 2018. Vol. 557. No. 7704. P. 212–216.

True Visions… 2006 – True Visions: The Emergence of Ambient Intelligence / ed. by E.H.L. Aarts, J.L. Encarnação. – Berlin: Springer, 2006.

Veličković et al. 2019 – *Veličković P., Ying R., Padovano M., Hadsell R., Blundell C.* Neural Execution of Graph Algorithms // 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, 2019. – URL: https://grlearning.github.io/papers/88.pdf

Wang et al. 2020 – *Wang J., Li Z., Long Q., Zhang W., Song G., & Shi C.* Learning Node Representations from Noisy Graph Structures // 2020 IEEE International Conference on Data Mining (ICDM). – Los Alamitos, CA: IEEE Computer Society. P. 1310–1315.

REFERENCES

Abrosimov V.K. & Raikov A.N. (2022) *Intelligent Agricultural Robots.* Moscow: Kar'era Press (in Russian)

Vagin V.N., Golovina E.Y., Zagoryanskaya A.A., & Fomina M.V. (2004) *Authentic and Plausible Inference in Intelligent Systems* (V.N. Vagina & D.A. Pospelov, Eds.). Moscow: Fizmatlit (in Russian).

Dubrovsky D.I. (2021) The Task of Creating General Artificial Intelligence and the Problem of Consciousness. *Russian Journal of Philosophical Sciences = Filosofskie nauki.* Vol. 64, no. 1, pp. 13–44 (in Russian).

Lepskiy V.E. (2021) Artificial Intelligence in Subjective Control Paradigms. *Russian Journal of Philosophical Sciences = Filosofskie nauki.* Vol. 64, no. 1, pp. 88–101 (in Russian).

Raikov A.N. (2009) Prominences of Macroeconomics. *Ekonomicheskie strategii.* 2009. No. 7, pp. 42–49 (in Russian).

Aarts E.H.L. & Encarnação J.L. (Eds.) (2006) *True Visions: The Emergence of Ambient Intelligence*. Berlin: Springer.

Adadi A. & Berrada M. (2018) Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access.* Vol. 6, pp. 52138–52160.

Arrieta A.B., Díaz-Rodríguez N., Del Ser J., Bennetot A., Tabik S., Barbado A., García S., Gil-López S., Molina D., Benjamins R., Chatila R. (2020) Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges towards Responsible AI. *Information Fusion.* Vol. 58, pp. 82–115.

Byrne R.M.J. (2019) Counterfactuals in Explaining Artificial Intelligence (XAI): Evidence from Human Reasoning. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*, pp. 6276–6282.

Byrne R.M.J. (2019) Counterfactuals in Explaining Artificial Intelligence (XAI): Evidence from Human Reasoning. In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)* (pp. 6276–6282). California: *International Joint Conferences on Artificial Intelligence*.

Chen M., Wei Z., Huang Z., Ding B., & Li Y. (2020) Simple and Deep Graph Convolutional Networks. *Proceedings of Machine Learning Research.* Vol. 119, pp. 1725–1735.

Doshi-Velez F. & Kortz M. (2017) Accountability of AI Under the Law: The Role of Explanation. *Berkman Klein Center Working Group on Explanation and the Law, Berkman Klein Center for Internet & Society Working Paper.* Retrieved from http://nrs.harvard.edu/urn-3:HUL.InstRepos:34372584

Einstein A., Podolsky B., & Rosen N. (1935) Can Quantum-Mechanical Description of Physical Reality Be Considered Complete? *Physical Review.* Vol. 47, no. 10, pp. 777–780.

Heaven W. D. (2020) Why asking an AI to explain itself can make things worse. *MIT Technology Review.* January 29. Retrieved from https://www.technologyreview.com/2020/01/29/304857/why-asking-an-ai-to-explainit-self-can-make-things-worse/

Kaul N. (2022) 3Es for Al: Economics, Explanation, Epistemology. *Frontiers in Artificial Intelligence.* Vol. 5, article 833238.

Leavy S., Meaney G., Wade K., & Greene D. (2020) Mitigating Gender Bias in Machine Learning Sata Sets. In: *International Workshop on Algorithmic Bias in Search and Recommendation* (pp. 12–26). Cham: Springer.

Leavy S., Siapera E., & O'Sullivan B. (2021) Ethical Data Curation for AI: An Approach based on Feminist Epistemology and Critical Theories of Race. In: *AIES'21: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 695–703). New York: The Association for Computing Machinery.

Lepskiy V. (2018) Evolution of Cybernetics: Philosophical and Methodological Analysis. *Kybernetes*. Vol. 47, no. 2, pp. 249–261.

Lin Y.T., Hung T.W., & Huang L.T.L. (2021) Engineering Equity: How AI Can Help Reduce the Harm of Implicit Bias. *Philosophy and Technology*. Vol. 34, no. 1, pp. 65–90.

Liu R., Balsubramani A., Zou J. (2019) Learning transport cost from subset correspondence. *arXiv*. Retrieved from https://arxiv.org/pdf/1909.13203.pdf

Madry A., Makelov A., Schmidt L., Tsipras D., & Vladu A. (2017) Towards deep learning models resistant to adversarial attacks. *arXiv*. Retrieved from https://arxiv.org/pdf/1706.06083.pdf

Mueller S.T., Hoffman R.R., Clancey W. Klein G. (2019) Explanation in Human-AI systems: A Literature Meta-Review Synopsis of Key Ideas and Publications and Bibliography for Explainable AI. *DARPA XAI Literature Review.* Retrieved from https://arxiv.org/pdf/1902.01876.pdf

Mueller S.T., Veinott E.S., Hoffman R.R., Klein G., Alam L., Mamun T., & Clancey W.J. (2020) Principles of Explanation in Human-AI Systems. *Association for the Advancement of Artificial Intelligence*. Retrieved from https://arxiv.org/pdf/2102.04972.pdf

Mumford S. & Anjium R.L. (2013) *Causation. A Very Short Introduction.* Oxford: Oxford University Press.

Nolin J. & Olson N. (2016) The Internet of Things and Convenience. *Internet Research*. Vol. 26, no. 2, pp. 360–376.

Orrell D. & Houshmand M. (2022) Quantum Propensity in Economics. *Frontiers in Artificial Intelligence.* Vol. 4, art. 772294.

Polya G. (1954) *Mathematics and Plausible Reasoning.* Princeton, NJ: Princeton University Press.

Raikov A. (2021) *Cognitive Semantics of Artificial Intelligence: A New Perspective.* Singapore: Springer.

Raikov A. (2022) Automating Cognitive Modelling Considering Non-Formalisable Semantics. In: Nagar A.K., Jat D.S., Marín-Raventós G., & Mishra D.K. (Eds) *Intelligent Sustainable Systems* (Lecture Notes in Networks and Systems. Vol. 334). Singapore: Springer.

Rauber A., Trasarti R., & Gianotti F. (2019) Transparency in Algorithmic Decision Making. *ERCIM News.* Vol. 116, pp. 10–11. Retrieved from https://ercim-news.ercim.eu/en116/special/transparency-in-algorithmic-decision-making-introduction-to-the-special-theme.

The BIG Bell Test Collaboration. (2018) Challenging Local Realism with Human Choices. *Nature*. Vol. 557, no. 7704, pp. 212–216.

Veličković P., Ying R., Padovano M., Hadsell R., & Blundell C. (2019) Neural Execution of Graph Algorithms. In: *33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, 2019.* Retrieved from https://grlearning.github.io/papers/88.pdf

Wang J., Li Z., Long Q., Zhang W., Song G., & Shi C. (2020) Learning Node Representations from Noisy Graph Structures. In: *2020 IEEE International Conference on Data Mining* (pp. 1310–1315). Los Alamitos, CA: IEEE Computer Society.