

УДК 004.8

doi: 10.15622/rcai.2025.041

СОЗДАНИЕ ПРОТОТИПА РЕКУРСИВНОГО ОБЪЯСНИМОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА НА ОСНОВЕ ПОВЕДЕНЧЕСКО-ЛОГИЧЕСКИХ МОДЕЛЕЙ¹

Ю.В. Трофимов (*ura_trofim@bk.ru*)
А.Н. Аверкин (*averkin2003@inbox.ru*)
И.П. Муравьев (*mip.22@uni-dubna.ru*)
А.К. Алексеев (*aak.24@uni-dubna.ru*)
Е.М. Кузнецов (*Kot454556@yandex.ru*)

Государственный университет «Дубна», Дубна

В работе представлен прототип рекурсивного объяснимого искусственного интеллекта, сочетающий поведенческо-логические модели мышления с современными нейросетевыми технологиями. Архитектура построена на когнитивной двухсистемной концепции, где Система 1 реализована как гибрид графовой и капсульной нейросети, обеспечивающих первичное интуитивное распознавание. Система 2 построена на базе дифференцируемого логического выводчика Neural Theorem Prover, реализующего формальный вывод в векторном пространстве признаков. Переключение между режимами осуществляется посредством Kolmogorov-Arnold Network (KAN), динамически управляющей распределением доверия между модулями. Дополнительную согласованность обеспечивает нечёткая логика, выступающая связующим звеном между нейросимвольными уровнями. Предложенная архитектура иллюстрирует принципы ХАИ 2.0 и демонстрирует высокую степень интерпретируемости решений без снижения точности. Система сопровождает каждый вывод внутренними логическими обоснованиями, обеспечивая прозрачность работы и устойчивость к неопределённости в сложных когнитивных задачах.

Ключевые слова: ХАИ, графовые нейросети, капсульные нейросети, Neural Theorem Prover, Kolmogorov–Arnold Network, нечёткая логика.

¹ Работа выполнена в рамках государственного задания Министерства науки и высшего образования Российской Федерации (тема № 124112200072-2).

Введение

Развитие объяснимого искусственного интеллекта (ХАИ) требует объединения высокой точности алгоритмов с их прозрачностью и интерпретируемостью. Ещё А.Н. Колмогоров и В.И. Арнольд заложили теоретический фундамент подобных подходов: их знаменитая теорема представления утверждает, что любая непрерывная функция многих переменных может быть представлена в виде суперпозиции конечного числа непрерывных функций одной. [Колмогоров, 1957]

Ученик Колмогорова В. Арнольд предложил конструктивный способ такой декомпозиции, благодаря чему данный результат получил название теоремы Колмогорова-Арнольда. Этот принцип лёг в основу современных ХАИ 2.0-подходов, нацеленных на встроенную интерпретируемость моделей. В манифесте ХАИ 2.0 отмечается, что создание надёжного ИИ требует сочетания субсимвольных нейросетевых методов с символьными алгоритмами объяснения [Longo et al., 2024]. Иными словами, необходимо изначально строить гибридные архитектуры, способные объяснять свои решения на внутреннем уровне, а не полагаться лишь на постфактум "расшифровку" чёрного ящика.

В первой вариации системы был разработан прототип системы ХАИ, сочетающий глубокие нейросети с нечеткой логикой, для решения задачи медицинской диагностики. [Трофимов, 2025] В частности, для выявления коронарных стенозов использовалась многоэтапная архитектура: свёрточная сеть ResNet-34 выделяла диагностически значимые кадры, модифицированная U-Net сегментировала коронарные сосуды с постобработкой CRF (коэф. Dice \approx 0,84; IoU \approx 0,78), а результаты дополнялись ХАИ-визуализацией Grad-CAM, LIME и Score-CAM. Для "прозрачности" заключения выходы нейросети были преобразованы в форму понятийных правил с помощью нейро-нечеткого модуля ANFIS. Такой гибридный подход продемонстрировал высокую точность диагностики (точность выявления стеноза > 90% на уровне кадра) и повысил доверие врачей за счёт интеграции механизмов объяснимости. Более того, включение нечётко-логических правил придало системе надёжность в сложных случаях: граничные ситуации, где нейросеть могла ошибиться из-за низкого контраста изображения, автоматически помечались как неопределённые или корректировались на основе базы знаний. Данный пример иллюстрирует концепцию ХАИ 2.0 на практике: объединение глубокой нейросети и логических правил позволило получить интерпретируемое решение без потери качества.

Данная работа развивает эти идеи и предлагает новую архитектуру рекурсивного объяснимого ИИ на основе поведенческо-логических моделей мышления. В основе – когнитивная двухсистемная теория Д. Канемана о Системе 1 и Системе 2 (интуитивные и аналитические процессы), дополненная современными нейросетевыми технологиями (графовые сети, капсульные сети) и механизмами для переключения режимов (Kolmogorov-Arnold Networks) с нечётко-логической интеграцией.

Двухсистемная архитектура интеллектуальной среды

Согласно Канеману, человеческое поведение определяется взаимодействием двух типов процессов: быстрой интуиции (Система 1) и медленно-го логического анализа (Система 2) [Kahneman, 2011]. В прикладном ИИ это находит отражение в гибридных архитектурах, где модуль Системы 1 реализован в виде глубокой нейросети – "черного ящика", мгновенно выдающей ответ, а Системы 2 – в виде более медленного рассуждающего модуля, который подключается при необходимости для углубленной проверки и объяснения решения. Эффективная когнитивная архитектура должна динамически переключаться между этими режимами, подобно тому как мозг человека задействует либо автоматизм, либо сознательный контроль в зависимости от сложности ситуации. В рассматриваемой системе данная парадигма реализована через связку нейросетевого компонента Системы 1 и логического компонента Системы 2, между которыми встроен интеллектуальный переключатель. Нечёткая логика при этом служит "мягким мостом", обеспечивая плавное взаимодействие подсистем. На рис. 1 изображен полный алгоритм работы системы.

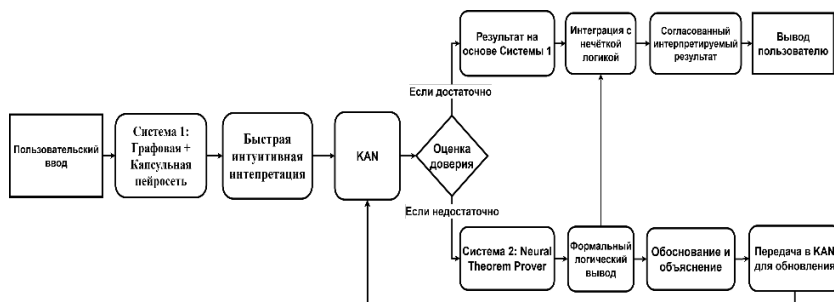


Рис. 1. Алгоритм работы системы

Система 1 – гибридная на основе графовой и капсульной нейросети

В роли Системы 1 выступает модуль глубокого обучения, отвечающий за быстрое "интуитивное" распознавание. Он реализован как гибридная глубокая модель, сочетающая графовую нейросеть с механизмом внима-

ния (Graph Neural Network) [Graph Neural Network, 2008] и капсульную сеть (CapsNet) [Sabour, 2017] в последовательной иерархической связке. Этот модуль отвечает за быстрое, "интуитивное" распознавание входных данных. Классическим выбором для Системы 1 являются сверточные нейросети, однако для более сложных и произвольных по структуре данных перспективнее применение графовых нейросетей. GNN способны эффективно работать с произвольными структурами связей, интегрируя информацию о взаимоотношениях между элементами входных данных. В отличие от данных на регулярных решётках, где связи между элементами фиксированы, графовое представление допускает моделирование неевклидовых структур данных – онтологий знаний.

Особый подкласс графовых нейросетей с механизмом внимания дополнительно выделяет наиболее значимые связи в графе, присваивая каждому ребру обучаемый коэффициент значимости. Это позволяет Системе 1 фокусировать "внимание" на ключевых фрагментах входной структуры, подавляя шумовые или малозначимые связи. Фактически GNN на первом этапе выполняет роль интуитивного распознавания шаблонов: за один-два шага распространения активности по графу модель вычленяет главные характеристики входного сигнала и формирует его компактное скрытое представление. Механизм внимания облегчает интерпретацию результатов – можно визуализировать, на какие именно связи графа модель опиралась при распознавании в наибольшей степени. В результате после прохождения через GNN Система 1 выдаёт осмысленное векторное представление наблюдения, дополненное сопутствующей информацией.

Для повышения точности, на следующем этапе Система 1 использует капсульную сеть. CapsNet состоит из вычислительных "капсул", каждая из которых представляет собой группу нейронов, совместно кодирующих все параметры обнаруженного примитива или признака объекта (положение, масштаб и ориентацию отдельного сегмента). Каждая капсула более высокого уровня получает на вход вектор выходных сигналов от нескольких капсул низшего уровня, благодаря чему в представлении явно сохраняется иерархическая структура типа "часть-целое". Так, при обработке изображений капсульная сеть сначала выявляет низкоуровневые капсулы, соответствующие элементарным деталям (границам, углам), затем объединяет их в капсулы более высокого уровня, отображающие крупные фрагменты или части объекта, и в конечном итоге формирует капсулу верхнего уровня, представляющую целостный образ.

Подобное многоуровневое представление обеспечивает две ключевые выгоды с точки зрения объяснимости результатов. Во-первых, оно более интерпретируемо: можно проследить, какие именно составные части образа распознаны моделью и как эти части организованы между собой в рамках целого. Во-вторых, CapsNet обладает устойчивостью к аномаль-

ным конфигурациям признаков: модель способна правильно распознать объект, даже если его составные части расположены в необычных или нетипичных взаимных позициях. Кроме того, капсульная архитектура характеризуется свойством согласованности преобразований: при трансформациях входных данных активации капсул изменяются предсказуемо и пропорционально произведенной трансформации, вместо потери существенной информации, как это происходит при использовании операций пулинга в классических свёрточных сетях. В данной архитектуре капсульная сеть интегрирована сразу после графовой и необходима для повышения точности модели.

Система 1 представляет собой последовательное сочетание GNN и CapsNet, соединяя интуитивное шаблонное распознавание с итеративным уточнением и проверкой структурной согласованности. Выходные данные Системы 1 – компактное векторное описание наблюдения, подкреплённое картой значимых связей (внимания) и согласованной иерархией выделенных признаков.

Система 2 – модуль логического рассуждения на основе Neural Theorem Prover

В роли Системы 2 выступает модуль логического вывода, реализованный на основе подхода Neural Theorem Prover (NTP). NTP был предложен как дифференцируемый аналог классического выводчика Prolog, выполняющего backward chaining – пошаговый обратный вывод с использованием правил базы знаний. [Rocktäschel et al., 2017] В NTP классическая дискретная унификация терминов заменена на вычисление сходства между их векторными представлениями, то есть логический вывод реализуется как непрерывная дифференцируемая процедура на подсимвольных репрезентации. Такой подход позволяет комбинировать нейросетевое представление знаний с формальной логикой, превращая процесс доказательства фактов в end-to-end обучаемую нейросетевую модель. Данный подход был применён и успешно протестирован на широком спектре задач [Концепция иерархически организованных..., 2025].

Благодаря градиентному обучению NTP осваивает ряд важных возможностей логического вывода. Во-первых, модель обучается располагать векторные эмбединги семантически схожих символов близко друг к другу, что реализует механизм "нечёткой" унификации – способность сопоставлять неидентичные, но похожие объекты. Во-вторых, NTP использует такое сходство для многошагового доказательства новых фактов на основе неполной базы знаний, распространяя вывод по цепочке правил. В-третьих, модель способна автоматически индуцировать новые логические правила на основе данных, обнаруживая скрытые закономерности. Наконец, NTP может применять как априорно заданные человеком правила, так и самостоятельно выведенные, комбинируя их для получения

сложных выводов. Заметим, что использование явных правил базы знаний гарантирует логическую корректность рассуждений: система 2 выводит новые утверждения только путём применения допустимых логических трансформаций к уже известным фактам. Более того, за счёт работы в векторном пространстве признаков даже одно заданное правило может быть обобщено на множество конкретных ситуаций с похожими объектами, что повышает эффективность и универсальность логического модуля.

Ключевым достоинством NTP как Системы 2 является высокая степень объяснимости выводов. Каждый факт доказывается через явную последовательность шагов – дерево доказательства, состоящее из применённых правил и найденных совпадений. Такая цепочка служит естественным объяснением полученного результата для пользователя или разработчика системы. NTP способен в процессе обучения выявлять латентные логические зависимости, которые затем могут быть декодированы в правила легко читаемые человеком. Таким образом, помимо самого ответа система генерирует понятные логические обоснования, демонстрируя на каких правилах и фактах основывается её вывод. Это существенно повышает доверие к модели и прозрачность её работы в контексте объяснимого ИИ.

Модуль NTP органично сочетается с Системой 1. Поскольку NTP оперирует распределёнными векторными представлениями символов, он напрямую потребляет на вход скрытые признаки и отношения, сформированные глубокой моделью первого уровня. Другими словами, выход Системы 1 поступает в NTP как база фактов для рассуждения. Благодаря полному дифференцированному описанию, обе системы могут обучаться совместно, образуя сквозной конвейер нейросимвольного ИИ. Быстрые интуитивные выводы Системы 1 дополняются строгой логической проверкой в Системе 2. В результате интеграции систем итоговое решение основывается не только на статистических шаблонах, но и на формальных знаниях, что обеспечивает и высокую точность, и интерпретируемость.

Стоит заметить, что прямое применение NTP сопряжено с комбинаторным ростом числа возможных путей доказательства, из-за чего наивная реализация плохо масштабируется на большие базы знаний. Для решения этой проблемы был разработан Greedy NTP (GNTP), который при доказательстве ограничивает поиск лишь топ-k наиболее перспективными фактами и правилами на каждом шаге. [Minervini et al., 2018] Такой жадный отбор значительно снижает вычислительную сложность, позволяя успешно применять нейронное доказательство теорем на реальных датасетах.

Kolmogorov-Arnold Network как интеллектуальный переключатель

Для координации взаимодействия между системами используется специальный модуль, в основе которого лежит Kolmogorov-Arnold Network (KAN) [Ziming et al., 2024]. Его роль – это слежение за ходом решения

задачи и выяснение, достаточно ли быстрого ответа или требуется углубленный анализ. В каждый момент KAN получает на вход показатели сложности и неопределённости текущей ситуации и вырабатывает сигнал переключения режима. Если задача простая и Система 1 уверенно справляется, переключатель оставляет основное решение без изменений. Если же входные данные сложны или нейросетевая модель не уверена в ответе, KAN активирует Систему 2. Формально выход переключателя – это коэффициент, лежащий в интервале от 0 до 1, определяющий вклад Системы 2 в итоговый вывод. В простейшем случае можно реализовать двоичный выбор: $g = 0$ допускает только ответ Системы 1, а $g = 1$ запускает полный логический вывод Системы 2. Однако более гибким является мягкое смешивание: итог, вычисляется как $y = g(x) \times y_{S2} + (1 - g(x)) \times y_{S1}$, где y_{S1} – ответ нейросети, а y_{S2} – ответ логического модуля. KAN особенно удобен для задания такой функции, поскольку его обучаемые одномерные элементы способны реализовывать пороговые или сигмоидоподобные зависимости от признаков сложности. По сути это реализует правило типа: "если доверие сети падает ниже X%, то подключить Систему 2", только не жёстко заданное, а выученное на данных и оптимально адаптированное под задачу. Аналогично по семантическим признакам входа KAN может давать системе "сигнал внимания": если обнаружено взаимодействие нескольких объектов, система переходит в режим вдумчивого, пошагового анализа. В итоге достигается баланс скорости и точности: глубокое рассуждение привлекается только при необходимости, что экономит ресурсы и одновременно повышает надёжность вывода.

Интерпретируемость переключателя

Важное достоинство KAN – его встроенная прозрачность. В классической нейросети влияние отдельных признаков скрыто в массе численных весов, и для оценки важности приходится применять косвенные методы. В KAN зависимость выхода от каждого признака задана явной функцией – графиком $\phi(x)$ на одном из ребер сети. Проанализировав форму этой функции, исследователь может напрямую оценить, как данный входной фактор влияет на результат. Сеть KAN предоставляет интерпретируемые параметры – кривые зависимостей, которые можно изучать подобно правилам. В контексте переключения режимов KAN позволяет понять, почему система решила привлечь или не привлекать модуль глубокого рассуждения. Иначе говоря, достигается мета-объяснимость: объяснимым становится не только базовое решение, выход Системы 1, но и факт его проверки. В результатах работы система может обосновать: "нейросеть дала неуверенный ответ, поэтому включён логический модуль, который, проанализировав признаки A, B, C, скорректировал вывод...". Такая многоуровневая прозрачность выгодно отличает предлагаемый подход от обычного post hoc XAI.

Ограничения KAN к ряду типов данных

Практическое применение KAN сопряжено с рядом ограничений, в задачах, характеризующихся высокой стохастичностью, структурной дискретностью или контекстной зависимостью.

KAN слабо устойчивы к входным данным, содержащим выраженные высокочастотные компоненты, неструктурированный шум или импульсные артефакты. Поскольку базовые элементы KAN реализуют гладкие одномерные аппроксиматоры, наличие нерегулярных возмущений приводит либо к переобучению, либо к агрессивному сглаживанию входной информации. Особенно это проявляется в биомедицинских задачах (анализ ЭКГ или МЭГ), где шумовые искажения носят негауссовский и высокоамплитудный характер.

Базовая архитектура KAN не предусматривает нативную обработку дискретных входных переменных. Классы, представленные в виде бинарных или категориальных признаков, требуют предварительного кодирования. Отсутствие встроенного механизма дискретизации ограничивает применение KAN в задачах, включающих логику принятия решений на основании номинативных категорий, таких как медицинские диагнозы, социодемографические характеристики и т.п.

Нечёткая логика как "мягкий мост" между модулями

Для интеграции результатов Системы 1 и Системы 2 используется нечёткая логика. Прежде всего, нечёткие продукционные правила позволяют избежать жёсткого разделения на случаи – условия правил могут выполняться с некоторой степенью принадлежности, а не только истина или ложь. Это соответствует характеру человеческих рассуждений, где многие понятия имеют размытие границ. В нашем прототипе выходы глубокой модели конвертируются в термины высокоуровневых признаков через лингвистические переменные и функции принадлежности. Например, числовое значение длины объекта превращается в понятие "большой", степень проявления признака – в "высокий" и т.д. Далее на основе обучающей выборки нейро-нечёткий модуль (тип ANFIS) автоматически извлекает базу правил. Такие правила привязаны к понятным характеристикам задачи. В ранее описанной системе диагностики они имели вид: "ЕСЛИ протяжённость стеноза большая И контрастность поражённого сегмента низкая, ТО степень стеноза значительная" [Трофимов, 2025]. Каждое правило имеет числовую степень активации вычисляемую на основе выхода нейросети. Мягкая интеграция заключается в следующем: даже если Система 2 не включается полностью, её нечёткие правила всё равно могут частично скорректировать или обогатить вывод Системы 1. В рамках формулы смешивания это соответствует случаю $0 < g < 1$. Тогда финальный ответ получается, как быстрый нейросетевой прогноз, дополненный

логическим заключением, взвешенным по степени уверенности. Такой подход реализует идею – мягкого переключения, исключающего резкие скачки в поведении системы. Кроме того, нечётко-логический модуль служит связующим звеном между субсимвольным представлением нейросети и символьной базой знаний. Он сопоставляет распределенные признаки с понятиями из онтологии, активируя соответствующие узлы графа знаний. Далее по графу знаний распространяется активность: если набор признаков соответствует некоторой концепции, она становится гипотезой решения; продукционные правила проверяют целостность этой гипотезы, отсекая противоречивые сочетания. В результате высокоуровневый модуль формирует объяснимое заключение в терминах предметной области – фактически, в читаемой для человека логике, подтверждённой данными. Такой вывод может сопровождаться указанием: какие признаки выявлены нейросетью и как они связаны с выводом через базу правил. Тем самым Система 2 становится "гарантом и пояснителем" Системы 1. Интеграция через нечёткую логику обеспечивает требуемую согласованность между подсистемами и понятность рассуждений для эксперта.

Такой двухсистемный подход был реализован с помощью альтернативных моделей в работах [Автоматизация анализа рентгеновских снимков..., 2025], [Кузнецов, 2025], посвящённых комплексному анализу рентгеновских снимков грудной клетки.

Рекурсивность алгоритма

Рекурсивность алгоритма строится вокруг KAN-переключателя, определяющего качество объяснений и заключений Системы 1. При удовлетворительном результате работы первой системы, переключатель «пропускает» на выход результат работы данной системы, но как только результаты работы падают и становятся некорректными, KAN-переключатель запускает Систему 2 для генерации новых правил и корректировки Системы 1. Система 1 проходит валидацию с новыми корректировками, и если система улучшается, то изменения вносятся в базу знаний. После применения корректировок, Система 1 перезапускается и происходит повторная оценка результатов. Таким образом, данный сегмент алгоритма рекурсивно дообучает Систему 1 за счет Системы 2, до тех пор, пока KAN-Переключатель не пропустит результат работы Системы 1. Фрагмент алгоритма представлен на рис. 2, где KAN-переключатель принимает оценку сложности n и оценку неопределенности m , а $const_1$, $const_2$ – допустимые значения для данных оценок соответственно. Такой подход позволяет реализовать адаптивность на уровне архитектуры ко внешним условиям и данным, что требует манифест XAI 2.0.

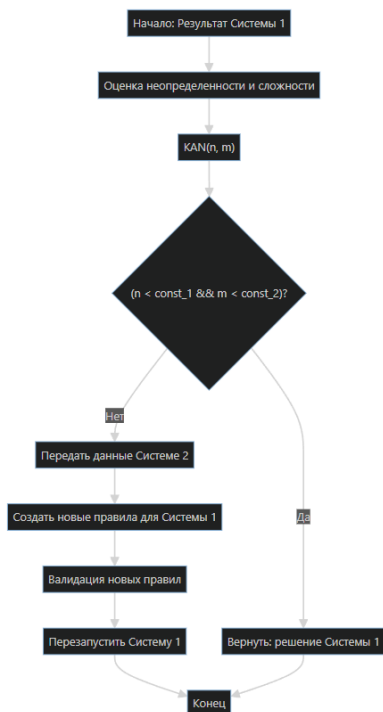


Рис. 2. Фрагмент алгоритма

Заключение

В работе была представлена архитектура прототипа рекурсивного объяснимого ИИ, сочетающая поведенческо-логические модели с современными нейросетевыми методами. Показано, что двухсистемный подход в сочетании с "переключателем" KAN и нечёткой логикой позволяет достичь одновременно высокой точности и интерпретируемости вывода. Система 1 обеспечивает быстрое распознавание образов и отношений, Система 2 вносит смысловой контроль и объяснение решений. KAN-модуль гибко управляет их взаимодействием, включая "медленное мышление" лишь тогда, когда это действительно нужно, а нечёткие правила служат мягким связующим звеном, переводя числовые активации в понятия и выводы на естественном языке.

Полученный прототип соответствует концепции XAI 2.0. Такая система не только выдаёт прогноз, но и сопровождает его развернутым самообъяснением, понятным для пользователя. В перспективе разработка подобных рекурсивных объяснимых систем открывает путь к более дове-

ренному ИИ, которому можно поручать задачи в критически важных областях без опасений непрозрачности решений. Кроме того, заложенные идеи могут быть развиты в сторону интеграции с нейросетями нового поколения. В частности, комбинация KAN-переключателей с большими языковыми моделями позволит строить саморефлексирующие агенты. Модель не только решает задачу, но и динамически формирует объяснение своих шагов, приближаясь к уровню осознанного человеческого рассуждения. Таким образом, подход, основанный на теоретических принципах Колмогорова и Арнольда, прокладывает новую траекторию к созданию гибридного ИИ, способного учиться, рассуждать и объяснять свои решения в едином интеллектуальном цикле.

Список литературы

- [Автоматизация анализа рентгеновских снимков..., 2025] Беляев М.И., Аверкин А.Н., Трофимов Ю.В., Шевченко А.В., Муравьев И.П. Автоматизация анализа рентгеновских снимков грудной клетки с использованием методов глубокого обучения и объяснимого искусственного интеллекта // Системный анализ в науке и образовании. – 2025. – № 2. – С. 32-41.
- [Колмогоров, 1957] Колмогоров А.Н. О представлении непрерывных функций нескольких переменных в виде суперпозиций непрерывных функций одного переменного и сложения // Доклад АН СССР. – 1957. – Т. 114. – С. 953-956.
- [Концепция иерархически организованных..., 2025] Трофимов Ю.В., Шевченко А.В., Аверкин А.Н., Муравьев И.А., Кузнецов Е.М. Концепция иерархически организованных объяснимых интеллектуальных систем: синтез глубоких нейронных сетей, нечеткой логики и инкрементального обучения в медицинской диагностике // VI Международная конференция по нейронным сетям и нейротехнологиям (NeuroNT) (Санкт-Петербург, 4-5 июнь 2025 г. Труды конференции. – Санкт-Петербург: IEEE, 2025. – С. 147. – doi: 10.1109/NeuroNT66873.2025.11049976.
- [Кузнецов, 2025] Кузнецов Е.М., Трофимов Ю.В., Муравьев И.П. Объяснимый инкрементный подход в диагностике пневмонии: от свёрточных нейросетей до генеративных текстовых заключений / (науч. рук. Аверкин А.Н.) // XIV Конгресс молодых ученых ИТМО (НИУ ИТМО, 07-11 апрель, 2025 г.) : Сборник тезисов докладов конгресса молодых ученых. [Электронное издание] . – URL: <https://kmu.itmo.ru/digests/article/15266> (дата обращения: 14.06.2025).
- [Трофимов, 2025] Трофимов Ю.В. [и др.] Нечёткие продукционные правила и нейросети глубокого обучения: объяснимый искусственный интеллект 2.0 для диагностики коронарных стенозов // Системный анализ в науке и образовании. – 2025. – № 2. – С. 73-82.
- [Graph Neural Network, 2008] Scarselli F., Gori M., Tsoi A.C., Hagenbuchner M., Monfardini G. The graph neural network model // IEEE Transactions on Neural Networks. – 2008. – Vol. 20, No. 1. – P. 61-80. – doi: 10.1109/TNN.2008.2005605.
- [Kahneman, 2011] Kahneman D. Thinking, Fast and Slow. – New York: Farrar, Straus and Giroux, 2011. – 499 p.

- [Longo et al., 2024] Longo L. [et al.]. Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions // Information Fusion. – 2024. – Vol. 106.
- [Minervini et al., 2018] Minervini P. [et al.]. Towards Neural Theorem Proving at Scale [Электронный ресурс] // arXiv. 2018. – URL: <https://arxiv.org/abs/1807.08204> (дата обращения: 14.06.2025).
- [Rocktäschel et al., 2017] Rocktäschel T., Riedel S. End-to-End Differentiable Proving [Электронный ресурс] // arXiv. 2017. – URL: <https://arxiv.org/abs/1705.11040> (дата обращения: 14.06.2025)
- [Sabour, 2017] Sabour S., Frosst N., Hinton G.E. Dynamic routing between capsules // Advances in Neural Information Processing Systems 30 (Long Beach, 4–9 Dec. 2017 г.): Труды конференции. В 1-м томе. Т. 1. – Нью-Йорк: Curran Associates, Inc., 2017. – С. 3856-3866. – doi: 10.48550/arXiv.1710.09829.
- [Ziming et al., 2024] Ziming L. [et al.]. KAN: Kolmogorov-Arnold Networks [Электронный ресурс] // arXiv. 2024. – URL: <http://arxiv.org/abs/2404.19756> (дата обращения: 14.06.2025).