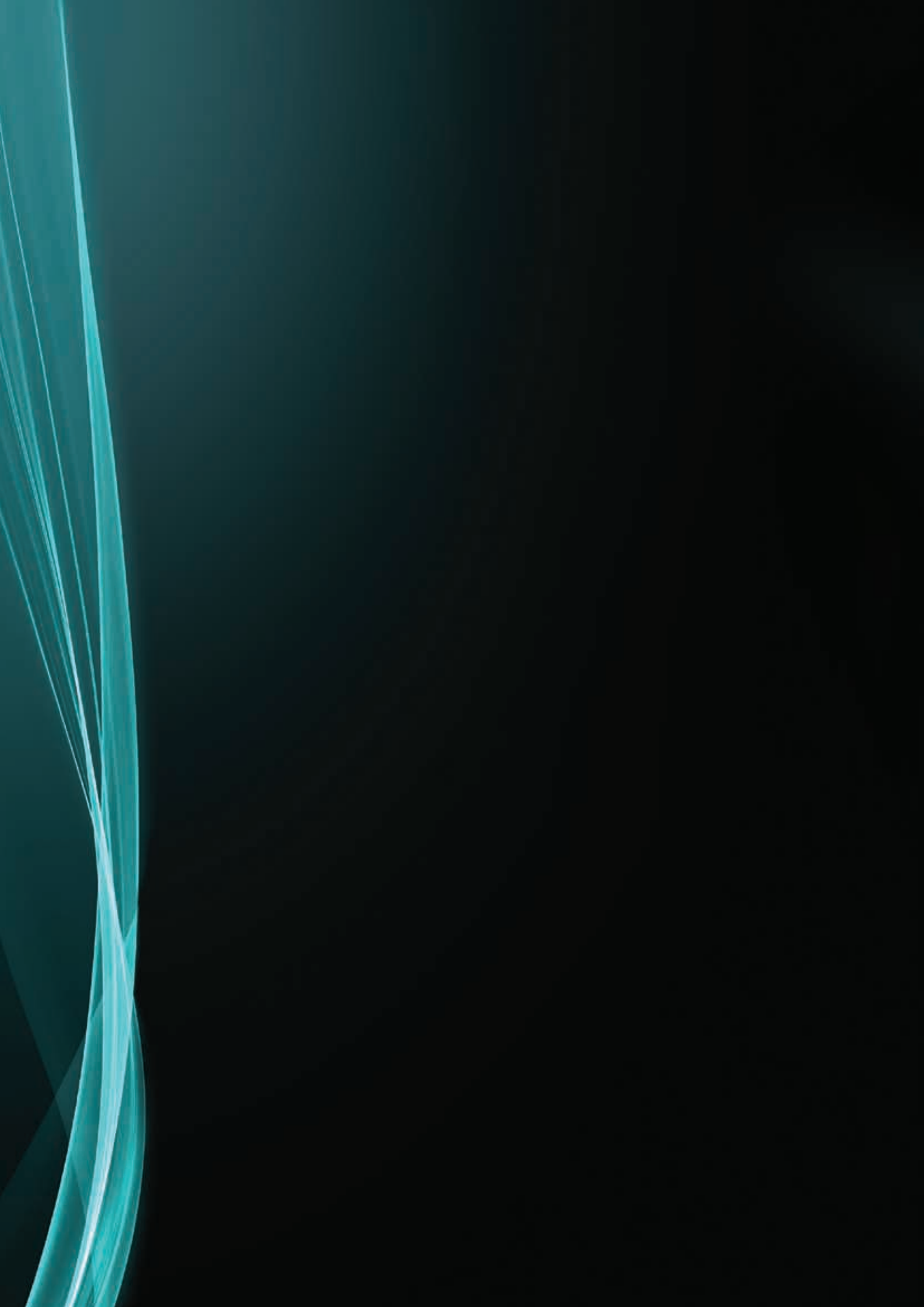
Горизонты искусственного интеллекта:

Какими будут технологии ИИ через 10 лет

Научно-исследовательский проект

Ноябрь 2025



СОДЕРЖАНИЕ

О проекте 3

Методология 8

Направление 1
Архитектуры, алгоритмы машинного обучения, оптимизация и математика 15

Направление 2
Вычисления для ИИ 21

Направление 3
Данные для ИИ 29

Направление 4
Фундаментальные генеративные модели 37

Направление 5
Безопасность, доверие и объяснимость 45

Направление 6
ИИ для узких задач (Narrow AI) 53

Направление 7
Управление, принятие решений и агентные/мультиагентные системы 61

Направление 8
Элементы AGI
69

Направление 9
Взаимодействие человека и машины
77

Направление 10
Общество в эпоху ИИ
85

Заключение
95

Приложение
96

Авторы итогового отчета
106

Команда итогового отчета
111

Благодарности
112

О ПРОЕКТЕ

Сферу искусственного интеллекта (ИИ) сегодня определяет фундаментальное противоречие. С одной стороны, ИИ трансформировался из узкоспециальной технологии в системный фактор, определяющий развитие экономики, государственного управления и основ суверенитета. С другой — его стремительная эволюция, подпитанная генеративными моделями, мультимодальностью и агентными системами, порождает комплекс сложнейших проблем: от запредельных требований к вычислительным ресурсам и энергозатратам до критически важных вопросов безопасности. Именно эта двойственность — колоссальный потенциал, сопряженный с серьезными ограничениями, — выдвигает тему ИИ в число приоритетных. Текущий момент является переломным: решения о масштабировании технологий сегодня закладывают траекторию развития на годы вперед.

Цель данного отчета заключалась в проведении комплексного анализа, результатом которого станет четкое разграничение устойчивых тенденций и краткосрочных колебаний.

В фокусе исследования идентификация перспективных возможностей и разработка обоснованных мер по снижению рисков.

Междисциплинарный и международный характер анализа позволил создавать реалистичные дорожные карты для научных исследований, разработки продуктов и формирования государственной политики.

10 тематических направлений итогового отчета образуют целостную и взаимосвязанную траекторию развития ИИ. Исследование выстроено по принципу сквозного анализа: от исходных данных и применяемых алгоритмов до инфраструктурного обеспечения, оценки рисков, широких социально-экономических последствий и, наконец, перспектив достижения AGI. Отдельное внимание уделено тенденции к автономизации ИИ, напрямую связанной с ростом его агентности.

Фактологической основой для подготовки материалов отчета стал комплексный анализ следующих источников:

21 форсайт-сессия, проведенная в 2025 г.

32 глубинных интервью с ведущими экспертами в области ИИ

В проекте приняли участие:

270+ ведущих ученых в области ИИ приняли участие в подготовке итогового отчета

36 стран — география авторитетных исследователей в области ИИ

Исследование также основывается на всестороннем анализе открытых данных и отраслевых отчетов. Такой ракурс предоставил возможность соотнести академическую повестку с практикой бизнеса, обозначить точки консенсуса и фундаментальные противоречия в области ИИ, а также четко идентифицировать узкие места развития. К их числу относятся проблемы качества данных, нехватки вычислительных ресурсов и обеспечения управляемости сложных моделей. Мы убеждены, что объективные выводы формируются лишь в процессе открытого международного профессионального диалога, который и стал основой для данного отчета.

КАРТА ФОРСАЙТ-СЕССИЙ

Российские сессии 2025 г.

AI Journey AI Horizons
16 июня
Санкт-Петербург

МФТИ
6 августа
Москва

Архипелаг–2025
10 августа
Москва

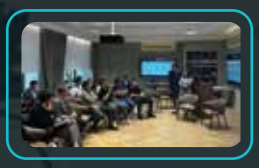
ННГУ им. Н. И. Лобачевского
18 августа
онлайн

МГУ им. М.В. Ломоносова
22 августа
Москва

Университет ИТМО
28–29 августа
Санкт-Петербург

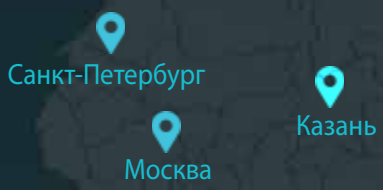
Университет Иннополис
30 августа
Казань

НИУ ВШЭ
10 сентября
Москва



СПбГУ
18 сентября,
Санкт-Петербург

Сколтех
25 сентября
Москва



КАРТА ФОРСАЙТ-СЕССИЙ

Зарубежные сессии 2025 г.

GITEX Африка
14–16 апреля
Марракеш, Марокко

Machines Can See
23–24 апреля
Дубай, Объединенные Арабские Эмираты

Шэньчжэньская глобальная выставка
22–23 мая
Шэньчжэнь, Китай

Партнерская форсайт-сессия
25 мая
Белград, Сербия

4-й Национальный Семинар по ИИ
23 июля
Исламбад, Пакистан

Панельная сессия по изучению будущих тенденций развития ИИ, 2025 WAIC
26–29 июля
Шанхай, Китай

AI Forum CIS 2025
8 августа
Самарканд, Узбекистан

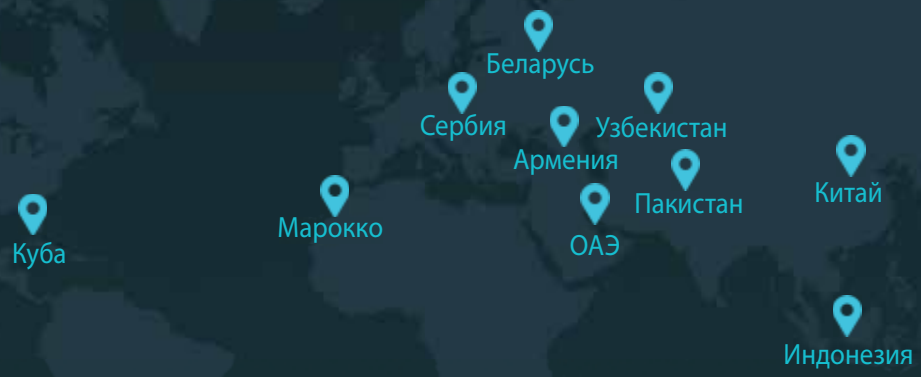
Национальная академия наук Беларуси
25 августа 2025
Минск, Беларусь

International AI Summit
16 сентября
Джакарта, Индонезия

Институт системного программирования им. В. П. Иванникова Российской академии наук (ИСП РАН)
22 сентября, Ереван, Армения

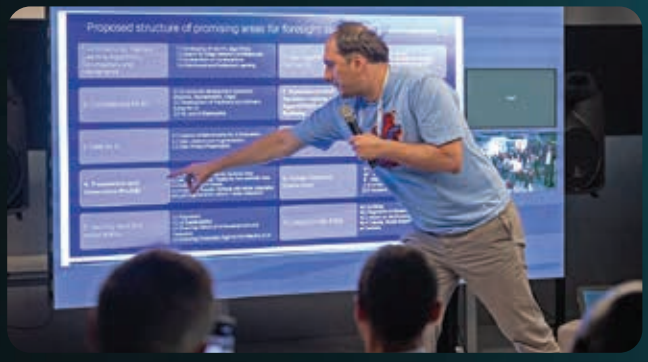
IWAIPR, UCIENCIA 2025
15 октября
Гавана, Куба

ASCOMP 2025
15 октября
Абу-Даби, Объединенные Арабские Эмираты




AI HORIZONS
Closed scientific event to foresee the future of
AI with best AI scientists of BRICS+ countries
in Saint Petersburg – the cultural heart of


AI ALLIANCE NETWORK
AI HORIZONS


Proposed structure of promising areas for foresight


AI HORIZONS

УЧАСТНИКИ ФОРСАЙТ-СЕССИЙ

В научных форсайт-сессиях приняли участие свыше 270 исследователей в области ИИ из 36 стран мира

МЕТОДОЛОГИЯ

Форсайт представляет собой систематический целенаправленный процесс формирования знаний о будущем. В рамках научно-технологического форсайта особое внимание уделяется выявлению приоритетных направлений научных исследований, потенциальных прорывов, а также оценке временных рамок решения тех или иных исследовательских задач. Настоящее исследование формирует прочный фундамент для принятия решений, связанных с планированием исследований, ресурсов и сотрудничества между академическими учреждениями, индустрией и государством как внутри страны, так и с иностранными партнерами.

Специфика сферы ИИ накладывает ограничения на проведения форсайт-исследований. ИИ характеризуется высокими темпами развития: циклы разработки существенно сокращаются, технологии и знания о них стремительно появляются и устаревают, а профиль научных исследований существенно меняется каждый год. В связи с этим методологическим фундаментом проекта стали **экспертные методы**, позволяющие оперативно получать наиболее актуальные данные и быстро адаптироваться к изменяющимся реалиям.

В реализации проекта приняли участие десятки экспертов самого высокого уровня. В качестве формального критерия для выбора экспертов использовался индекс Хирша (не ниже 15), а также наличие не менее двух публикаций на конференциях уровня A* в области ИИ с 2020 г. и/или значительное число публикаций в высокорейтинговых журналах по ИИ.

Форсайт-исследование включало 3 основных этапа:

1. Глубинные интервью | для формирования базы независимых оценок и предложений

Была проведена серия глубинных экспертных интервью с ведущими российскими и международными учеными для сбора независимых экспертных оценок по направлениям поисковых и фундаментальных исследований в сфере ИИ. Структурированные интервью охватывали оценку текущего состояния, перспектив развития и формирование предложений по изменениям перечня поднаправлений и исследовательских задач, формируемого и актуализированного в ходе форсайт-исследования. В результате был сформирован датасет аналитических данных, экспертных оценок и предложений.

2. Форсайт-сессии | для формирования базы консенсусного мнения

Одним из ключевых этапов исследования стали форсайт-сессии, их целью был сбор и формирование коллективных экспертных оценок по направлениям и исследовательским задачам в сфере ИИ. Была проведена 21 форсайт-сессия, на каждой из которых работали несколько рабочих групп, сформированных в соответствии с тематиками направлений научных исследований, идентифицированными в ходе выполнения Форсайта 2024 г. Важнейшие задачи форсайт-сессий: выявление спорных точек зрения и формирование консенсусного мнения. Основной результат каждой сессии — структурированные выводы, относящиеся к актуализации перечня перспективных исследований в области ИИ, особенностям этих исследований и оценкам временных горизонтов достижения важных этапов.

3. Итоговый отчет

Информация, полученная в ходе экспертных интервью и форсайт-сессий, была аккумулирована в единый датасет, после чего была проведена независимая проверка выводов и корректировка на основе дополнительных комментариев и данных.

На базе валидированных данных был сформирован итоговый отчет. На уровне каждого из 10 направлений над отчетом работали по два редактора: выдающиеся ученые по соответствующей тематике. На основе сформированных консенсусных данных редакторы оформили видение каждого из направлений перспективных исследований по единообразной структуре.

Отчет содержит обобщение результатов по каждому направлению, конкретные исследовательские задачи, их решения, временные горизонты, их решения и другие важные аналитические выводы.

В рамках форсайт-исследования подготовлен набор типовых материалов (гайды интервью, типовые сценарии форсайт-сессий, шаблоны форм для заполнения, презентаций для экспертов и др.), формирующих методологическую базу, существенно упрощающую проведение аналогичных исследований в будущем.

В целях подготовки к адресному взаимодействию с учеными и уточнения зон компетенций были проанализированы с использованием инструментов ИИ аннотации и полные тексты всех публикаций исследователей — участников форсайт-сессий.

Для анализа обсуждавшихся тем и научных вызовов сформирован корпус аудиозаписей общей продолжительностью свыше 60 часов (более 400 тыс. слов), включающий материалы форсайт-сессий и интервью с учеными. На основе данного корпуса выполнен глубинный ИИ-анализ, позволивший структурировать тематическое поле и выделить базовые тезисы, что обеспечило комплексную репрезентативность высказанных в ходе обсуждений мнений. Также для проведения форсайт-сессии была подготовлена система, анализирующая рекомендации потенциальных научных взаимодействий на основании анализа публикационного профиля ученого.

ПИРАМИДА НАПРАВЛЕНИЙ ИССЛЕДОВАНИЙ В СФЕРЕ ИИ*

Как связаны направления исследований между собой? Представим их в виде пирамиды, структура которой демонстрирует, как фундаментальные исследования последовательно определяют и обеспечивают развитие прикладных направлений.

Вершину пирамиды занимает **«Интеграция»** — создание элементов AGI и развитие человеко-машинного взаимодействия.

Эту высшую ступень обеспечивает уровень **«Управление»**, отвечающий за безопасное и эффективное применение моделей. Без доверия, объяснимости даже самые совершенные модели остались бы «черными ящиками», непригодными для ответственного применения в реальном мире.

В свою очередь, возможность управления возникает благодаря **«Ядру»** современных систем ИИ — разработке фундаментальных и генеративных моделей, которые являются центральными двигателями прогресса.

Всё это базируется на прочном **«Фундаменте»** — исследованиях в области вычислительных мощностей и архитектурных решений. Именно данный уровень определяет возможности и закладывает основу для всей пирамиды развития.

Интеграция

№ 8 «Элементы AGI»

№ 9 «Взаимодействие человека и машины»

Управление

№ 5 «Безопасность, доверие и объяснимость»

№ 7 «Управление, принятие решений, и агентные/ мультиагентные системы»

Ядро

№ 4 «Фундаментальные генеративные модели»

№ 6 «ИИ для узких задач (Narrow AI)»

Фактор внешней среды

№10 «Общество в эпоху ИИ»

Фундамент

№ 1 «Архитектуры, алгоритмы машинного обучения, оптимизация и математика»

№ 2 «Вычисления для ИИ»

№ 3 «Данные для ИИ»

* Пирамида схематично показывает взаимосвязь направлений форсайта между собой

НЕТЕХНОЛОГИЧЕСКИЕ ФАКТОРЫ, ВЛИЯЮЩИЕ НА РАЗВИТИЕ ИССЛЕДОВАНИЙ В ОБЛАСТИ ИИ

Развитие технологий ИИ и формирование облика исследований в области ИИ зависят не только от технологических достижений, но и от множества нетехнологических факторов, которые влияют на темпы, направленность и устойчивость прогресса. Эти факторы формируют экосистему, в которой технологии рождаются, тестируются, внедряются и регулируются общественным сознанием и государственными структурами. Во многих случаях именно внешние обстоятельства определяют, какие задачи будут ставиться в приоритет, какие данные допустимо использовать, какие методики окажутся приемлемыми с точки зрения этики, доверия и законности, а также как быстро будут создаваться и внедряться новые решения.

Для каждого из направлений были определены 3–5 **ключевых** нетехнологических факторов, влияющих на развитие направления.

- **Регулирование развития технологий** задает рамки для разработки и внедрения ИИ, вынуждая разработчиков и исследователей учитывать вопросы безопасности, прозрачности и ответственность.
- **Запрос общества на этику и доверие** усиливает давление на компании работать открыто и объяснять решения, что в итоге формирует более устойчивые и безопасные системы ИИ.
- **Доступность квалифицированных кадров** определяет темпы инноваций и качество разрабатываемых и внедряемых решений.
- **Спрос на экономическую эффективность и тиражирование** подталкивает к созданию масштабируемых и эффективных решений, которые можно внедрять в разных отраслях без значительных переработок.
- **Доступность и качество данных** напрямую влияет на точность и обобщаемость моделей, а также на возможность прозрачной оценки рисков и предотвращения смещений.
- **Рост спроса на автономность процессов** стимулирует развитие систем, способных принимать решения без постоянного человеческого вмешательства, но при этом сохранять надзор и ответственность.
- **Глобальный спрос на энергоэффективность**, включая экологическую повестку, способствует созданию менее ресурсоемких моделей и более эффективных инфраструктур обработки данных.
- **Потребность в технологическом суверенитете** подталкивает страны к независимой разработке критических технологий и стратегическому инвестированию в собственную экосистему ИИ.

На базе полученных экспертных оценок была построена сводная визуализация, отражающая ландшафт ключевых факторов (рис. 1, с. 12).

На рисунке горизонтальная ось формирует представление о годе, соответствующем пику реализации тренда, размер кругов коррелирует с цветом факторов, а цвета показывают, на какие направления оказывает влияние тот или иной фактор. Следует еще раз отметить, что здесь оценивались только ключевые факторы для каждого направления, и отсутствие указания какого-либо направления для конкретного фактора не означает, что фактор не влияет на развитие этого направления, а всего лишь говорит о том, что влияние других факторов на это направление оценивается экспертами как более сильное.

Наиболее важными факторами, которые сильнее всего влияют на комплексное развитие ИИ, являются регуляторные аспекты, запрос общества на этику и доверие, доступность квалифицированных кадров и спрос на экономическую эффективность и тиражирование.

Все обозначенные факторы взаимосвязаны и создают единую экосистему: регуляторика и этика формируют требования к качеству и ответственности, общественный запрос на доверие задает характер взаимодействия с пользователями, кадровый потенциал может ограничивать скорость развития, а экономическая масштабируемость является двигателем практического внедрения и долгосрочной устойчивости технологий.

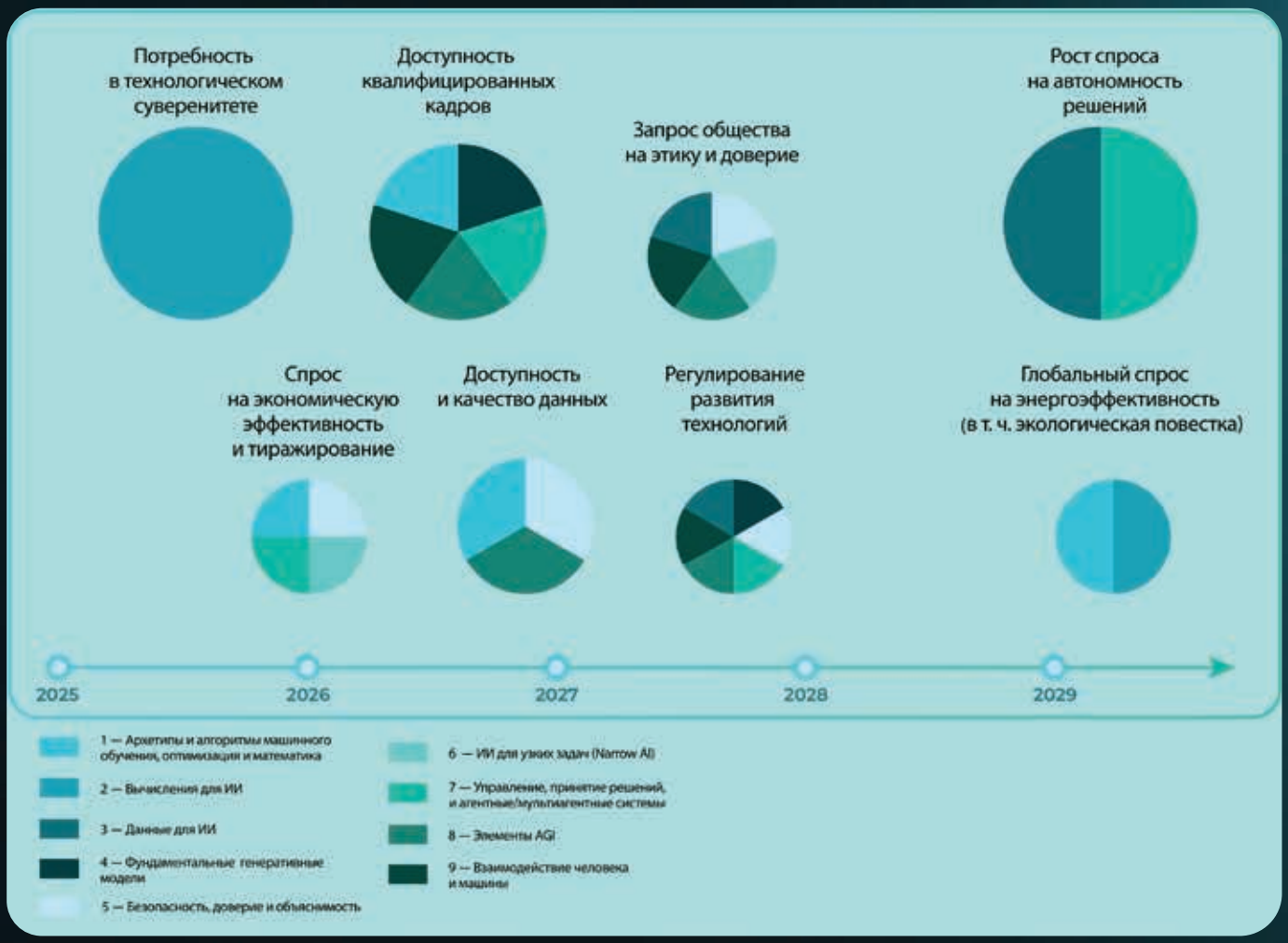


Рисунок 1. Карта ключевых нетехнологических факторов, влияющих на развитие исследований в области ИИ

СТРУКТУРА НАПРАВЛЕНИЙ ИТОГОВОГО ОТЧЕТА

1. Архитектуры, алгоритмы машинного обучения, оптимизация и математика	1.1 Разработка новых алгоритмов машинного обучения 1.2 ИИ-архитектуры 1.3 Ускорение вычислений 1.4 Распределенное и федеративное обучение 1.5 Математические основы ИИ
2. Вычисления для ИИ	2.1 Разработка специализированных вычислителей для ИИ (квантовых, фотонных, нейроморфных и др.) 2.2 Разработка аппаратно-программных комплексов для ИИ 2.3 Фреймворки машинного обучения и ИИ
3. Данные для ИИ	3.1 Разработка бенчмарков для ИИ 3.2 Формирование, преобразование и сопровождение данных 3.3 Обеспечение конфиденциальности и защиты данных
4. Фундаментальные генеративные модели	4.1 Фундаментальные генеративные модели для символьных данных 4.2 Фундаментальные генеративные модели для несимвольных данных 4.3 Мультимодальные фундаментальные генеративные модели 4.4 Трансфер знаний с адаптацией базовой генеративной модели 4.5 Аугментация фундаментальных генеративных моделей
5. Безопасность, доверие и объяснимость	5.1 Alignment 5.2 Объяснимость ИИ 5.3 Обеспечение безопасной разработки и эксплуатации ИИ 5.4 Обеспечение защиты от результатов использования ИИ с целью взлома
6. ИИ для узких задач (Narrow AI)	6.1 Компьютерное зрение (CV) 6.2 Обработка естественного языка (NLP) 6.3 Другие узкие ИИ-технологии (S2T, RecSys, TSA и т. д.)
7. Управление, принятие решений и агентные/мультиагентные системы	7.1 Разработка алгоритмов обучения с подкреплением 7.2 Агентные системы 7.3 Мультиагентные системы
8. Элементы AGI	8.1 Рассуждения и рефлексия 8.2 Lifelong learning 8.3 Гибридный ИИ 8.4 Embodiment (вещественный ИИ) 8.5 Моделирование мозга и психики
9. Взаимодействие человека и машины	9.1 Технические средства прямого взаимодействия с нервной системой человека 9.2 Технические средства традиционного человеко-машинного взаимодействия 9.3 Методы и алгоритмы взаимодействия с человеком
10. Общество в эпоху ИИ	10.1 Механизмы глобального управления ИИ, включая регулирование ИИ 10.2 Этика ИИ 10.3 Изучение эффектов влияния технологий ИИ на общество

НАПРАВЛЕНИЕ 1

Архитектуры, алгоритмы машинного обучения, оптимизация и математика

НАПРАВЛЕНИЕ 1

Архитектуры, алгоритмы машинного обучения, оптимизация и математика

1. Краткое описание направления

Практически весь современный ИИ (как и его предшественник, классический Data Science) стоит на трех китах:

1. Архитектура/модель (разделяющая гиперплоскость, нейросеть, диффузионный процесс и т. п.). Можно рассматривать как функцию, что особенно понятно на примере нейронной сети, где выбор архитектуры соответствует выбору структуры сети. Однако архитектура не обязательно является единственной функцией; это может быть набор функций, например для разных агентов. Функция бывает случайной или адаптивной, меняющейся в процессе обучения. Таким образом, архитектуру можно сравнить с генотипом, чье фенотипическое проявление зависит от различных, в т. ч. внешних, факторов.

2. ML-алгоритмы/постановки: формализация задачи (обучения) как задачи оптимизации или седловой задачи, составление функционала (для состязательных постановок, в которых ищется седловая точка, термин «функционал» немного может сбивать) невязки (функции риска), введение регуляризации, ограничений и т. п. В функционал закладываются данные и модель. Задача — подогнать модель под данные, формализовав это как задачу оптимизации. Обученная модель должна предсказывать новые данные, предполагая, что закон их порождения неизменен. Нетривиальный пример — обучение сиамской нейронной сети. В общем случае возможны более сложные постановки: седловые задачи (GAN), теоретико-игровые (RLHF), вариационные неравенства и многоуровневая оптимизация. Ключевая особенность — существенные ограничения на доступ к оракулу. Обычно доступен лишь стохастический градиент функционала (или меньше, как в задачах о многоруких бандитах), который вычисляется, например методом обратного распространения ошибки.

3. Оптимизация. Решение задач оптимизации, седловых задач или их комплексов требует создания эффективных распределенных алгоритмов. При этом учитывается как скорость сходимости численных методов стохастической оптимизации, так и оптимизация вычислительного процесса на двух уровнях: алгоритмическом, что включает разработку специализированных методов для частных классов задач с использованием малоранговых аппроксимаций и неточных вычислений, и аппаратном, подразумевающем поиск новых вычислительных архитектур, ориентированных на специфику типовых задач.

Данное направление является крайне динамичным, при этом ключевыми факторами, определившими вехи развития, становятся:

- использование больших языковых моделей (далее — LLM) для автоматизированного поиска архитектур, что может привести к прорыву в их создании;
- развитие саморефлексирующихся систем ИИ, которые могут улучшать себя без постоянного человеческого контроля, ускоряя прогресс;
- рост интереса к мультиагентным системам, где ИИ действует коллективно и автономно в сложных средах;
- прорыв диффузионных моделей в генеративном моделировании, сменивший фокус исследований;
- появление новых оптимизаторов (Shampoo, SOAP, Muon), которые эффективнее классического Adam, они повышают производительность обучения.

Эволюционное развитие направления позволило прийти к следующим ключевым выводам за последние 5 лет:

- законы масштабирования LLM показывают, что при определенном уровне они превосходят человека по многим когнитивным задачам, включая творчество, что требует генерации качественных данных;

Особая благодарность за работу над данным направлением
А. Гасникову, Н. Бачанину, Х. Хуссейну

- в 2024–2025 гг. AI4Science достиг системного уровня генерации новых научных знаний, что может привести к «цепной реакции» ускорения науки и технологий;

- осознано, что эффективные вычислительные архитектуры должны включать разнообразных агентов с разными ролями, что напоминает социальную эволюцию мозга человека и предполагает развитие симбиотических нейроархитектур.

2. Обзор текущего развития направления

В последние 1–2 года значительным изменением стало применение LLM для автоматического предложения и оценки новых архитектур нейронных сетей, что выходит за рамки традиционной числовой оптимизации и включает семантические описания архитектур. Это указывает на самореферентное развитие ИИ, где сами модели ИИ используются для улучшения своих же основ. Усилия по ускорению вычислений сосредоточены на совершенствовании методов дистилляции, прунинга и квантования для уменьшения моделей и их адаптации к различным устройствам. Активно развиваются продвинутые методы, такие как speculative decoding и Mixture of Experts.

Среди ключевых факторов, являющихся драйверами для данного направления, можно выделить:

Растущее энергопотребление ИИ-моделей и экологические/экономические ограничения

Основная проблема ИИ сегодня — это энергопотребление. Экспоненциальный рост энергопотребления крупных ИИ-моделей стал серьезным вызовом, требующим поиска новых принципов и оптимизации. Развитие вычислительно эффективных методов ускоряет совершенствование методов дистилляции, прунинга и квантования, а также разработку новых эффективных алгоритмов численной линейной алгебры и оптимизаторов.

Исчерпание данных, необходимых для обучения

Качество моделей напрямую зависит от объема доступных для обучения данных. Для целого ряда направлений (например, LLM) все доступные данные во многом уже были использованы. Возникает проблема, где брать новые. World Models стимулирует развитие моделей с встроенными априорно знаниями, что позволяет существенно сокращать требования к объему обучающей выборки. Совместные и сохраняющие конфиденциальность экосистемы данных стимулируют развитие федеративного обучения, консорциумов по обмену данными и технологий сохранения конфиденциальности, которые позволяют нескольким сторонам вносить вклад и извлекать выгоду из разнообразных наборов данных, не раскрывая конфиденциальную информацию, тем самым смягчая проблему нехватки данных в критически важных областях.

3. Исследовательские вызовы, стимулирующие или ограничивающие развитие направления

ВЫЗОВ 1.1

Ограниченная обобщаемость и адаптивность ИИ-моделей

Ограниченная способность современных ИИ-моделей к обобщению на новые, ранее не виданные домены или задачи без значительного переобучения. Проблема ограниченной обобщаемости и адаптивности ИИ-моделей фундаментально ограничивает широкое применение ИИ в динамичных реальных сценариях (например, в робототехнике, персонализированной медицине), требующих гибкости и быстрой адаптации к новым доменам. Сегодня активно ведутся исследования в области новых теоретических инструментов для глубокого обучения, «математических основ ИИ», включая математику генеративных моделей (diffusion models, flow matching), оптимального управления и обучения с подкреплением, а также новых численных методов линейной алгебры. В перспективе разработка методов для повышения «эластичности моделей» и «непрерывного обучения» приведет к созданию более универсальных, устойчивых и ресурсоэффективных ИИ-моделей, способных учиться подобно человеку, что значительно расширит их применимость и доверие к ним.

ВЫЗОВ 1.2

Интерпретируемость и объяснимость

Проблема работы нейронных сетей, как «черного ящика», препятствует фундаментальным прорывам и полному пониманию того, как работают глубокие нейронные сети, что приводит к созданию «черных ящиков» и недостаточной теоретической обоснованности новых алгоритмов. В перспективе разработка новых теоретических инструментов и математического фундамента для глубокого обучения, включая соединение статистики и оптимизации, приведет к созданию более обоснованных, робастных и предсказуемых моделей ИИ.

ВЫЗОВ 1.3

Квантификация и вербализация неопределенности в решениях ИИ

Создает риски в чувствительных областях (медицина, автономное вождение), т. к. пользователи могут принимать решения, основываясь на «самоуверенных» ошибочных прогнозах ИИ, снижая общее доверие к ИИ, что в значительной степени ограничивает безопасное и ответственное взаимодействие ИИ с человеком в ситуациях, требующих осторожности и подтверждения решений, особенно когда модели «плохо выражают свою неопределенность». Сегодня исследования этого вызова сосредоточены на использовании внутренней статистики модели (вероятности последовательностей, энтропия), разработке вспомогательных моделей для предсказания ошибочной генерации, архитектурных решениях, а также методах, обучающих модели явно выражать свою уверенность или неуверенность естественным языком. В перспективе разработка надежных методов позволит ИИ «более ответственно взаимодействовать с человеком», сигнализируя о необходимости вмешательства и повышая «доверие к ИИ», делая ИИ более надежным помощником и способным «отклонять ответы».

4. Перспективные исследовательские задачи

Развитие ИИ-систем в сторону большей автономии и сложности порождает новые фундаментальные задачи в области их архитектур, алгоритмов и оптимизации. Эти задачи требуют глубоких исследований, выходящих за рамки уже известных подходов, и затрагивают как теоретические основы машинного обучения, так и прикладные аспекты взаимодействия ИИ с человеком и окружающей средой. Ниже представлены перспективные исследовательские задачи, выделенные на основе экспертных обсуждений.

★ ЗАДАЧА 1.1

Создание «алгебры» над архитектурами, постановками задачи (условно назовем ML-алгоритмами/подходами) и алгоритмами решения возникающих оптимизационных задач

Речь идет об ИИ-селекции — систематическом объединении успешных компонентов и структурных блоков для генерации новых архитектур, постановок задач и алгоритмов оптимизации, применимых к сложным задачам, сводимым к известным элементам. Каждый исследователь интуитивно формирует решения для новых задач оптимизации, но это важно формализовать, например через алгебраический подход к построению эффективных методов на базе классического или стохастического градиентного спуска с использованием конкретных приемов. Несмотря на наличие общих принципов математической формализации задач и понимания основ существующих архитектур, главная задача — формализация этих правил и разработка новых архитектур.

Мультиагентный ИИ рассматривается как перспективный инструмент для создания новых архитектур, тогда как математика обеспечивает формализацию и решение остальных задач. Уже получены значимые результаты по мета-оптимизации, где поиск оптимального алгоритма представлен как задача полуопределенной оптимизации, успешно решенная для важных подклассов выпуклой и стохастической оптимизации.

В результате такой формализации существенно может сократиться поиск подходящих архитектур, настройка их гиперпараметров. В целом процесс решения задачи ИИ станет лучше формализованным и, стало быть, в большей степени может быть автоматизирован.

★ ЗАДАЧА 1.2

Разработка универсальных и адаптивных ИИ-моделей с повышенной обобщаемостью

Основная задача заключается в преодолении ограниченной способности современных ИИ-моделей к обобщению на новые, ранее не виданные домены или задачи без значительного переобучения. Необходима разработка механизмов эластичности для запоминания и забывания, аналогично человеческому мозгу, чтобы модели могли лучше обобщать и улучшать адаптацию к новым областям задач. Это включает «обучение без забывания» (learning without forgetting) и повышение эластичности моделей.

Решение этой задачи может включать несколько подходов:

- разработка «новых функций потерь» (loss functions), учитывающих как эффективность обучения новой задачи, так и сохранение предыдущих знаний;
- исследование и проектирование адаптивных архитектур нейронных сетей, которые могут динамически изменять свою структуру или реконфигурироваться для адаптации к новым данным или задачам;
- применение непрерывного обучения (continual learning) и адаптации доменов (domain adaptation);
- изучение когнитивно-инспирированных подходов для повышения гибкости и адаптивности моделей;
- разработка новых методологий обучения, которые позволят моделям лучше обучаться, например, на малом количестве примеров, подобно тому как

ребенок, увидев всего несколько разных собак (большую, маленькую, лохматую), формирует в сознании абстрактное понятие «собака» и в дальнейшем без труда узнает любую другую собаку, даже незнакомой породы.

В перспективе решение данной задачи приведет к созданию более универсальных, робастных и ресурсоэффективных ИИ-моделей, способных к непрерывному обучению и адаптации к новым доменам. Это существенно расширит практическую применимость ИИ в динамичных реальных сценариях, таких как робототехника, персонализированная медицина или системы управления, и повысит доверие к системам, способным учиться на протяжении всей своей жизни без необходимости полного сброса и переобучения. Возможность обучения только по 10 экземплярам позволит использовать ИИ в областях с ограниченными данными.

★ ЗАДАЧА 1.3

Интеграция научных знаний и «Мировых моделей» в ИИ-системы

Задача заключается в преодолении проблемы «черного ящика» характера ИИ путем эффективного встраивания проверенных научных знаний (из физики, химии, биологии) и человеческого экспертного опыта непосредственно в архитектуры и алгоритмы ИИ. Это включает разработку «Мировых моделей» (World Models), способных изучать базовые законы физики сами по себе без явного программирования. Актуальность усиливается в областях с небольшим количеством данных, таких как геология или биология.

В числе методов решения данной задачи можно выделить:

- разработка гибридных ИИ-моделей, включающих physics informing или chemistry informing of neural networks для встраивания научных знаний и «не только данных»;
- использование теоретически мотивированного дизайна моделей, «априорных знаний», «ограничений и оптимизации»;
- интеграция теоретических моделей «в функции потерь»;
- создание World Models, способных самостоятельно изучать базовые законы физики;
- разработка методов для автоматической сегментации и явного автоматизированного моделирования, например в символической форме (symbolic regression).

В перспективе решение этой задачи приведет к появлению foundation-моделей, инкорпорирующих человеческие знания по физике и другим наукам, что позволит комбинировать теорию и машинное обучение для решения проблем в областях с небольшим количеством данных. Так появятся модели с повышенной робастностью, экстраполяцией, объяснимостью и ясностью, снизится зависимость от огромных объемов данных и уменьшатся затраты на данные и ресурсы. Ускорение научных открытий в химии, биологии и физике, а также создание ИИ, способного «понимать» мир на более глубоком уровне.

5. Важные выводы: Экспертное заключение

Характер развития данного направления сегодня определяется фундаментальным сдвигом парадигм, вызванным экспоненциальным ростом сложности ИИ. Это сопровождается активным поиском новых теоретических инструментов и математических основ для глубокого обучения, а также для явлений вроде «благоприятного переобучения». Прорыв диффузионных моделей показал, как адаптация довольно известных техник к новой области может привести к впечатляющим результатам, что стимулирует поиск новых физических (точнее, естественно-научных) инсайтов.

Параллельно происходит сдвиг в парадигме оптимизаторов от доминировавшего ранее Adam к более эффективным матрично-ориентированным подходам, таким как Shampoo, SOAP и Muon. Всё более заметным становится развитие ИИ, когда большие языковые модели используются для генерации новых архитектур на основе базовых блоков, обещая прорыв в создании архитектур. Кроме того, наблюдается активный рост интереса к мультиагентным системам, которые стремительно выходят на первый план как центральное направление исследований ИИ.

55% исследовательских задач, согласно оценочному прогнозу, не достигнут исчерпания к 2030 г.

НАПРАВЛЕНИЕ 2

Вычисления для ИИ

НАПРАВЛЕНИЕ 2

Вычисления для ИИ

1. Краткое описание направления

На сегодняшний день создание специализированных вычислительных технологий и инфраструктур является залогом успешной эволюции современного ИИ, основанного на больших моделях и данных. В отличие от классических высокопроизводительных вычислений, ориентированных на решение преимущественно задач компьютерного моделирования, специфика расчетов в современном ИИ связана с двумя аспектами:

1. Опора на характерные матричные и тензорные операции, которые эффективно реализуются на SIMD-архитектурах, порождающих взрывной рост потребности в системах на основе GPU и TPU, в отличие от более универсальных MIMD-архитектур на обычных CPU.

2. Использование больших массивов данных в основе вычислительного процесса, требующих не только планирования и балансировки вычислительной нагрузки, но и управления совместным распределением данных и вычислений, обеспечивающим наилучшую производительность.

В настоящее время область вычислений для ИИ охватывает три ключевых поднаправления:

1. **Развитие масштабируемых и энергоэффективных классических вычислительных архитектур, в т. ч:**

- повышение энергоэффективности параллельных вычислительных архитектур для реализации матричных операций (GPU, TPU и пр.) при сохранении линейной масштабируемости по числу ядер. Целью является обеспечения роста размеров моделей ИИ без значимого удорожания их обучения и использования. Более того, новые вычислительные архитектуры также могут стимулировать разработку новых методов и моделей ИИ, которые лучше подходят для использования уникальных возможностей этих архитектур, образуя эффективный цикл взаимного продвижения;

- создание распределенных вычислительных архитектур, динамически масштабируемых под вычислительные задачи ИИ, в первую очередь — федеративного обучения и мультиагентных систем на основе LLM. Данное направление является критическим для развития мультиагентных систем и LLM в целом;

- повышение возможностей специализированных архитектур для эффективной работы с данными и коммуникации агентов в задачах ИИ, в т. ч. туманные вычисления (fog computing), краевые вычисления (edge AI) и пр. В отличие от мультиагентных систем и больших языковых моделей, эта задача направлена на реализацию проблемы воплощенного ИИ (создание автономных роботов-агентов и пр.).

2. **Создание вычислительных архитектур для ИИ на основе новых принципов, в т. ч.:**

- создание нейроморфных оптоэлектронных и фотонных архитектур (NPU), обеспечивающих баланс производительности и энергоэффективности, кратно превосходящих возможности GPU и TPU;

- разработка квантовых вычислительных архитектур, адаптированных под задачи ИИ, в т. ч. кодизайн архитектур и квантово-подобных вычислительных алгоритмов машинного обучения. Несмотря на то что в настоящее время в этой области нет впечатляющего прогресса, исследования квантово-подобных алгоритмов способствуют определению контуров вычислительных систем в будущем;

- разработка специализированных ускорителей для прикладных задач ИИ adhoc, включая гибридные вычислительные систем и конфигурируемые системы на основе FPGA, а также инструменты для работы с ними.

Особая благодарность за работу над данным направлением:
А. Бухановскому, А. Фан, Ф . Цинь.

3. **Создание методических основ, разработка системного и промежуточного ПО (middleware) для эффективной реализации задач ИИ на классических и перспективных вычислительных архитектурах.** Вместе с тем в рамках данного поднаправления необходимо отметить следующие аспекты:

- алгоритмические механизмы отображения (mapping) и оптимизации алгоритмов ИИ с учетом специфики конкретной вычислительной архитектуры;
- автоматизация конструирования новых алгоритмов ИИ для нестандартных вычислительных архитектур, а также кодизайна новых вычислительных архитектур и алгоритмов;
- эффективное управление вычислительными процессами (планирование, балансировка нагрузки, распределение данных в памяти) для характерных алгоритмов машинного обучения и задач ИИ;
- создание специализированных средств реализации задач ИИ, например компиляторы и низкоуровневые фреймворки для TPU;
- разработка инструментальных фреймворков для быстрой разработки систем ИИ для высокопроизводительных и распределенных вычислительных архитектур, включая создание гибридных систем (нейросимвольный ИИ, мультиагентные системы и большие языковые модели и пр.);
- разработка программных платформ для мультиагентных систем, обеспечивающих эффективное сотрудничество и коммуникацию между несколькими агентами. Такие платформы должны поддерживать динамическое распределение задач, управление ресурсами и разрешение конфликтов в распределенных средах;
- разработка специализированных аппаратных платформ для мультиагентных систем, оптимизированных для параллельной обработки данных и передачи данных с низкой задержкой.

2. Обзор текущего развития направления

В настоящее время ландшафт исследований в сфере вычислений для ИИ формируется на основании двух разнонаправленных факторов: развитие агрегированных технологий высокопроизводительных вычислений для больших моделей ИИ, с одной стороны, и технологий распределенных вычислений для реализации мультиагентных систем и специализированных систем ИИ ad hoc, с другой. Как следствие, это связано с интенсивной проработкой следующих вопросов:

Повышение вычислительной эффективности при работе с большими моделями на существующих архитектурах, которое обеспечивается в два этапа: во-первых, за счет высокоэффективных методов обучения и инференса (дистилляция (teacher–student) для сжатия моделей, прореживание (pruning), квантизация, speculative decoding, а также использование Mixture of Experts (MoE) для ускорения вычислений). Во-вторых, путем эффективного отображения (mapping) результирующих алгоритмов на архитектуру GPU/TPU (что ярко проявилось на моделях DeepSeek).

Диверсификация типов вычислительных архитектур для задач ИИ, в т. ч. выходящих за рамки традиционных решений. В настоящее время внимание уделяется нейрографическим спайковым нейросетям, оптическим нейросетям, а также реализующим их нейроморфным компьютерам и ускорителям, которые могут быть как основанными на классических кремниевых технологиях, так и полностью альтернативными.

Эффективные распределенные вычисления для LLM

Обучение современных LLM требует колоссальных вычислительных ресурсов, которые могут быть доступны только в распределенной гетерогенной среде, на кластерах с различными типами GPU. Управление и эффективное использование этого «оркестра GPU» остается серьезной и нерешенной задачей. Проблема оркестрации становится еще более яркой в свете перехода к мультиагентным системам больших языковых моделей.

Использование больших языковых моделей в поиске архитектур и адаптации алгоритмов

Большие языковые модели уже традиционно применяются как мощный инструмент для проектирования и оптимизации нейросетей в рамках подхода Neural Architecture Search (NAS). Однако в настоящее время большие языковые модели могут решать и задачу отображения (mapping) алгоритмов, т. е. проектировать нейросеть, которая оптимально исполняется на конкретной вычислительной архитектуре, а также подбирать параметры архитектуры для конкретного алгоритма машинного обучения.

3. Исследовательские вызовы, стимулирующие или ограничивающие развитие направления

Несмотря на значительный прогресс, существует ряд фундаментальных и прикладных проблем, которые препятствуют дальнейшему развитию и широкому внедрению вычислительных технологий в области ИИ.

ВЫЗОВ 2.1

Отсутствие кардинального решения проблем энергопотребления и устойчивости за счет перехода к новым типам вычислительных архитектур

Стремительно растущие энергозатраты больших моделей ИИ становятся критическим ограничением для их масштабирования. Требуются исследования в области принципиально новых, энергоэффективных парадигм вычислений — включая квантовые нейронные сети, голографические представления данных и фотонные технологии — чтобы радикально снизить потребление энергии.

ВЫЗОВ 2.2

Отсутствие унифицированных фреймворков и постановок задач для новых типов вычислительных архитектур

Несмотря на обнадеживающие разработки в области создания аппаратной базы вычислений для ИИ, их широкое использование упирается в отсутствие зрелых инструментальных фреймворков, а также постановок задач, наглядно демонстрирующих их применимость. Этот пробел сдерживает их практическое применение и широкую интеграцию. К тому же разработчики hardware, как правило, не готовы сами ставить прикладные задачи и часто пользуются бенчмарками, которые в ИИ-сообществе уже считаются устаревшими.

ВЫЗОВ 2.3

Слабость интеграции и совместной оптимизации аппаратного и программного обеспечения

Для достижения оптимальной производительности систем ИИ требуется глубокая оптимизация совмещения аппаратных и программных компонентов. Это требует, в первую очередь, разработку специализированных библиотек, компиляторов и инструментов, которые эффективно связывают новые аппаратные решения с пользовательскими приложениями. Однако задача остается сложной из-за тесного переплетения указанных уровней.

ВЫЗОВ 2.4

Отсутствие понимания свойств сложных мультиагентных систем ИИ в свете использования распределенных вычислительных архитектур

Перспективные MAC LLM могут включать в себя сотни и тысячи различных агентов, из которых в реальных задачах одновременно взаимодействует гораздо меньшее количество. Как следствие, распределенная вычислительная архитектура для таких систем должна быть динамической, с возможностью выделения под задачу требуемых ресурсов и связывания их оптимальной топологией коммуникаций. Однако для этого необходимо уметь предсказывать поведение самой MAC LLM, что в настоящее время еще не обеспечено ни методической базой, ни даже пониманием происходящих процессов — особенно в свете коллективного поведения агентов, приводящего к эмерджентным эффектам.

ВЫЗОВ 2.5

Отсутствие эффективной интеграции между традиционными искусственными нейронными сетями и новыми моделями, основанными на больших языковых моделях

Существующие методы часто используют традиционные нейронные сети в качестве инструментов, вызываемых агентами в системах, основанных на больших языковых моделях, однако не могут полностью интегрировать и использовать сильные стороны обоих. Данная ситуация приводит к ограниченным возможностям в решении сложных задач ИИ, которые могли бы извлечь выгоду из объединения возможностей традиционных нейронных сетей и больших языковых моделей.

4. Перспективные исследовательские задачи

Несмотря на общую значимость перспективных задач, связанных с созданием и воплощением «в железе» новых вычислительных архитектур (таких как квантовые вычислители или нейроморфные процессоры), центр тяжести их решения лежит вне существующего сообщества специалистов в области ИИ. Однако более значимыми могут оказаться задачи (связанные с созданием новых моделей и алгоритмов, определяющих требования к перспективному аппаратному обеспечению), разработка инструментария для массового использования существующих решений, а также применение для решения прикладных проблем. Как следствие, перспективные задачи сводятся к следующему:

★ ЗАДАЧА 2.1

Создание процессоров на нейроморфных принципах и их алгоритмов для целей ИИ; сенсоры, окружение и исполнительные устройства для нейроморфных процессоров: в части создания методической и алгоритмической базы ИИ для вычислительных систем, основанных на новых принципах (в первую очередь, фотонные и оптоэлектронные нейроморфные системы).

Для квантовых вычислителей эта задача пока еще не является столь критической, что связано с отдаленностью появления адекватной аппаратной части. Здесь могут быть рассмотрены разнообразные вопросы, связанные с адаптацией как существующих методов ML и моделей ИИ (например, в части прогнозирования временных рядов или компьютерного зрения), так и постановкой качественно новых задач.

★ ЗАДАЧА 2.2

Фреймворки машинного обучения и ИИ для существующих вычислительных архитектур

При этом наиболее востребованы будут фреймворки для ресурсоемких областей, связанных с иными направлениями форсайта, включая LLM, MAC и элементы AGI, федеративное обучение и пр. В том числе:

- Фреймворки для symbolic and hybrid AI для создания систем гибридного ИИ, эффективно сочетающих LLM и классические технологии работы со знаниями;
- Фреймворки для agent-based schemes and applications (включая Embodied Agents) для построения распределенных систем на основе разнородных интеллектуальных агентов;
- Фреймворки для prompt engineering для тонкой настройки, адаптации и кастомизации LLM.

★ ЗАДАЧА 2.3

Создание фреймворков для обучения и инференса

Несмотря на большое количество общих методов ускорения обучения и инференса, их выбор всецело определяется особенностью вычислительной архитектуры. Потому принципиальным является создание инструментальных фреймворков, которые могут обеспечить адаптацию заданных классов моделей под конкретный вычислитель, исходя из критериев энергоэффективности и производительности. При этом такие фреймворки наиболее востребованы в областях, связанных с использованием специальных вычислителей с ограниченными характеристиками (например, на борту автономного робота).

★ ЗАДАЧА 2.4

Создание системного ПО, повышающего эффективность работы с оборудованием

Интенсивное развитие TPU приводит к необходимости создания для них экосистемы инструментального ПО, аналогичного GPU, а также средств портирования с GPU на TPU. Это включает в себя собственно алгоритмы тензорных компиляторов, промежуточные языки, средства работы с памятью, а также обеспечение кроссплатформенности.

★ ЗАДАЧА 2.5

Распределенные вычисления и LLM в части создания методов, алгоритмов и программного инструментария управления динамическими распределенными инфраструктурами для федеративного обучения, LLM и MAC LLM

Эта задача включает в себя решение вопросов обеспечения многоуровневого параллелизма в моделях и данных для использования на распределенных архитектурах, связанную с этим оптимизацию вычислений и использования памяти, эффективное управление коммуникациями (включая конструирование топологии системы под задачу), а также обеспечение устойчивости и безотказности такой системы в целом.

★ ЗАДАЧА 2.6

Разработка миниатюрных устройств, способных поддерживать большие модели и агентов для интеллектуального принятия решений в сложных средах

Существующие большие модели обычно требуют громоздких устройств для развертывания, что ограничивает их применимость в сценариях, где важна компактность и мобильность, например в робототехнике. Существует необходимость в разработке миниатюрных, но мощных устройств, способных удовлетворить вычислительные потребности крупных моделей и агентов, позволяя роботам принимать разумные решения в сложных и динамичных средах.

5. Важные выводы: Экспертное заключение

В числе важных выводов в рамках данного направления стоит отметить следующие:

- Вычислительные архитектуры для ИИ не являются самостоятельным направлением, их развитие инспирируется созданием новых методов и моделей ИИ, а также постановками прикладных задач. Математические основы ИИ, изобретение новых алгоритмов и архитектур сохраняют непреходящую важность и будут двигать прогресс вперед.

- Будущее ИИ связано с гибридными моделями, которые объединяют сильные стороны машинного обучения с символьными рассуждениями и использованием априорного научного знания. Как следствие, это требует создания гибридных вычислительных систем для ИИ, сочетающих особенности как классических, так и новых вычислительных архитектур, выходящих за рамки универсальных вычислений и оптимизированных под конкретные задачи ИИ.

- Текущее направление развития больших языковых моделей делает неизбежным рост интереса к распределенному и федеративному обучению: как к самой технологии для адаптации и дообучения LLM, так и к более легковесной альтернативе. Это указывает на долгосрочную актуальность таких технологий, особенно для приложений ИИ на персональных устройствах и в чувствительных секторах.

- В отличие от классических алгоритмов высокопроизводительных вычислений, априори не имеющих собственных механизмов отображения (mapping), планирования и балансировки вычислительной нагрузки, ИИ дает возможность строить не только самообучающиеся, но и самоадаптирующиеся системы, способные перестраивать свою работу, исходя из требований к производительности и энергоэффективности на конкретной архитектуре. Это открывает большие возможности для создания технологий воплощенного ИИ, реализуемых посредством ко-дизайна самого алгоритма, вычислительной инфраструктуры для его исполнения, а также иных средств воплощения (сенсоров, актуаторов и пр.).

- Развитие экосистемы вычислений для ИИ чувствительно не только к доступу к аппаратным платформам, но и к инструментальному ПО для их использования. При этом существует напряжение между преимуществами open-source-решений и стратегической необходимостью в национальных проприетарных технологиях (аппаратных и программных), обеспечивающих технологическую независимость.

Сквозное, но несамостоятельное направление, облик которого существенно зависит от характера развития остальных направлений

64% **задач направления не имеют выраженного горизонта исчерпания и будут актуальны еще многие годы**

79% **задач предполагает скорые (в ближайшие 1–2 года) научные достижения. Научные достижения, как правило, носят периодический, повторяющийся характер**

Развитие направления — важный и сквозной приоритет, заделы в области вычислений для ИИ формируют прочный фундамент для развития всего ИИ

Наблюдается развитие вычислительных архитектур на совершенно новых принципах. Будущее ИИ связано с созданием гибридных вычислительных систем, сочетающих особенности как классических, так и новых вычислительных архитектур для ИИ

Развитие новых архитектур сдерживается развитием системного инструментального обеспечения, в связи с этим требуется одновременная поддержка двух направлений

НАПРАВЛЕНИЕ 3

Данные для ИИ

НАПРАВЛЕНИЕ 3

Данные для ИИ

1. Краткое описание направления

Данные являются «цифровой кровью» современного ИИ. Они необходимы как для обучения, так и для оценки качества работы моделей ИИ. По мнению участников форсайта, именно данные, а не модели становятся главным фактором успеха в создании успешных приложений ИИ. При этом наблюдается постоянный дефицит качественных данных для обучения. Недаром на конференции NeurIPS (A* в области ИИ), ежегодно выделяется специальный трек для новых датасетов и бенчмарков.

В настоящее время здесь обозначаются три основных направления исследований.

1. Разработка бенчмарков для ИИ

   Предполагает создание стандартизированных фреймворков, наборов данных и метрик, используемых для измерения и сравнения производительности моделей ИИ в различных задачах. Каждой исследовательской задаче ИИ соответствует некоторый бенчмарк. Более того, новые бенчмарки в современном ИИ — и есть основной способ постановки новых исследовательских задач. К тому же они обеспечивают возможность объективной оценки прогресса ИИ-решений. Как отмечали участники форсайта: «Нужна прежде всего разработка стандартов, создающих единые подходы к испытанию технологий ИИ». Исследования в области бенчмарков включают формирование наборов данных, разработку методик и метрик сравнения моделей, а также создание методов тестирования моделей на данных, возникающих в реальном времени.

2. Формирование, преобразование и сопровождение данных

   Включает широкий круг задач по работе с данными. Генерация и аугментация данных необходимы в тех случаях, когда сбор требуемого объема реальных данных является слишком дорогостоящим, либо получение реальных данных в принципе затруднено. При этом в фокусе должны быть подходы, гарантирующие целостность и достоверность синтетически сгенерированных данных. Помимо генерации и аугментации, в данной области актуальны такие задачи, как активное обучение с использованием краудсорсинга, создание симуляционных сред для обучения и тестирования в реальном времени, а также оценка качества, фильтрация и управление данными на протяжении всего жизненного цикла систем ИИ, включая обнаружение и коррекцию смещений и предвзятостей в данных.

3. Обеспечение конфиденциальности и защиты данных

   Является критически важной и постоянно актуальной областью исследований, которая позволяет развивать ИИ в соответствии с требованиями закона и этическими нормами. Здесь актуальны исследования по использованию технологий, которые скрывают или искажают чувствительную информацию. Также следует использовать возможности синтетических данных для того чтобы уменьшить нашу зависимость от реальных данных, которые могут и должны оставаться конфиденциальными. При этом обнаружение синтетических данных и анализ их влияния на обучение моделей — еще одна из приоритетных задач. В целом необходимо обнаруживать и маркировать различные виды искусственного или незаконно используемого контента.

История развития ИИ показывает, что драйвером появления новых вызовов в машинном обучении всегда являлось создание новых культовых бенчмарков, таких как MNIST, ImageNet, GLUE… Однако со временем бенчмарки теряют свою актуальность. Как было отмечено в рамках одной из дискуссий, *«мы всё еще используем устаревшие тесты, чтобы оценивать модели нового поколения».* Эволюция ИИ постоянно приводит к пересмотру подходов к работе с данными, и за последние годы на развитие области наиболее сильно повлияли:

- развитие технологии «дипфейк» и реакция общества на нее;
- правовые и этические ограничения, включая лицензии, приватность и согласие пользователей на

Особая благодарность за работу над данным направлением
Ю. Визильтеру, С. Линь, Ч. Чэнь

обработку персональных данных, которые формируют барьеры для сбора большого массива данных;

- выявление склонности моделей ИИ к галлюцинациям и смещениям, что привело к необходимости переосмысления понятия надежности данных;
- развитие робототехники и беспилотного транспорта, поставившее задачи обучения в открытом мире и обучения на данных, генерируемых в реальном времени.

2. Обзор текущего развития направления

Сегодня данные рассматриваются не только как технический ресурс, но и как политико-экономическая и даже этическая ценность. Крупные компании владеют уникальными датасетами, что создает информационную асимметрию, с одной стороны, между субъектами рынка и научным сообществом, а с другой стороны, между странами — лидерами в разработке ИИ и другими странами, не обладающими такими масштабными ресурсами. В связи с этим сформировался значимый тренд на создание общедоступных, стандартизированных датасетов под государственным или международным контролем.

Нехватка качественных примеров естественного текста и других типов реальных данных стала сегодня одним из главных барьеров на пути дальнейшего повышения функциональности моделей ИИ: *«Все знают, что у нас не хватает данных. Особенно текста: зачастую у нас просто нет естественных текстов. Поэтому многие современные модели используют синтетические данные».* Во многих областях оказались востребованы исследования по созданию технологий автоматического аннотирования больших объемов реальных данных. В то же время создание полностью синтетических датасетов превратилось в ключевое направление исследований, особенно в специализированных доменах, таких как медицина или наука. Участники международного форсайта также отметили политическую и культурную значимость этого направления: *«Синтетические данные абсолютно необходимы для большинства наших локальных языков, т. к. текстов мало, особенно в научной и других сферах».* При этом особое внимание уделяется валидации синтетических данных. В ходе дискуссий было отмечено, что *«если не контролировать генеративную модель, она может породить артефакты, легко принимаемые за истину».* Проводятся активные исследования в области обеспечения приватности и защиты данных с использованием таких методов, как дифференциальная приватность, распределенное и федеративное обучение. Эти методы позволяют работать с зашумленными данными, хотя качество модели иногда ухудшается, что остается открытой проблемой.

3. Исследовательские вызовы, стимулирующие или ограничивающие развитие направления

Стремительное развитие ИИ выявляет как фундаментальные, так и прикладные вызовы в работе с данными.

ВЫЗОВ 3.1

Надежность и репрезентативность данных

Как отмечают участники форсайта, *«надежность данных — основа надежности моделей»*. Сообщество всё в большей степени озабочено достоверностью данных. Особое беспокойство вызывают галлюцинации ИИ — генерация нефактических или ложных данных. Имеются проблемы с репрезентативностью: большинство датасетов ориентировано на английский язык и западную культуру, что вызывает смещения данных и делает модели менее эффективными в других регионах. В этом контексте, как подчеркнул один из экспертов, *«данные — это форма политической власти».*

Влияние обозначенного вызова связано с тем, что надежные данные — основа доверия к выводам и решениям ИИ. Ненадежные данные могут привести к неточным прогнозам, предвзятым моделям и дорогостоящим сбоям при использовании ИИ. Возможно, самый важный аспект здесь — этический ИИ: устранение смещений (bias) в данных гарантирует справедливость и непредвзятость суждений ИИ.

Для обеспечения надежности данных в мире используются, исследуются и разрабатываются такие инструменты, как методы управления данными, методы обеспечения качества данных (строгие протоколы сбора данных, непрерывная проверка и очистка данных), методы выявления и устранения галлюцинаций, скрытой и явной предвзятости в наборах данных. Создание и использование собственных эталонных наборов данных — также важный инструмент для борьбы с предвзятостью.

ВЫЗОВ 3.2

Потребность в качественных синтетических данных

Одним из центральных вызовов современного ИИ является т. н. проблема стены данных.

Законы масштабирования моделей показали, что для улучшения их работы необходимо увеличить объем данных, используемых при обучении. Однако во многих областях для обучения используются уже все доступные данные из цифрового следа человечества.

От решения данной проблемы непосредственно зависит сценарий дальнейшего развития ИИ: возможно ли дальнейшее улучшение качества работы моделей, или максимальный уровень развития ИИ уже достигнут, поскольку он ограничен имеющимся объемом данных, произведенных человечеством. С практической точки зрения вопрос в том, как выйти за пределы исходной обучающей выборки. С фундаментальной точки зрения вызов почти философский — может ли учитель обучить ученика, который превзойдет его самого? Иными словами, ожидает ли нас в обозримой перспективе ИИ, существенно превосходящий уровень человека (superhuman AI), или для его появления имеется принципиальный запрет.

Последние полтора года усилия сообщества были сосредоточены на решении проблемы «стены данных» за счет генерации синтетических данных, хотя сама задача синтеза новых данных, которые согласуются с распределением реальных данных, является по-прежнему сложной и актуальной исследовательской задачей. Достигнут значительный прогресс в таких областях, как рассуждения, программирование и математика, где возможна объективная проверка качества генераций. В других областях, как отметили участники форсайта, оценка ИИ-сгенерированного контента на предмет фактической точности крайне трудоемка, а отсутствие надежных метрик делает объективную оценку достоверности и качества выходных данных чрезвычайно сложной задачей. Оценка за пределами субъективного мнения человека остается нерешенной проблемой: *«Чаще всего у нас нет объективного механизма. Мы можем опираться на мнение людей, но тогда потолок — человеческая способность».*

В области робототехники, воплощенного ИИ (Embodied AI), инженерных и бизнес-приложений всё большую роль играют среды-симуляторы как специальный источник данных, направленный на моделирование динамических процессов и работы LLM/VLM/VLA-моделей в реальном времени. Используются как физические симуляторы, так и модели бизнес-процессов, расчетное и инженерное ПО (software). При этом главный вызов заключается в том, чтобы обеспечить одновременно реалистичность и скорость работы таких цифровых двойников. Скорость необходима для выполнения многочисленных циклов обучения с подкреплением. Реалистичность важна для переноса обучения из виртуальной среды в реальную. Как отмечали участники форсайта, *«перенос моделей из симуляторов в реальные условия остается серьезным узким местом. Преодоление разрыва sim2real — ключ к внедрению ИИ-моделей».*

ВЫЗОВ 3.3

Обеспечение конфиденциальности и безопасности данных

Как отмечали участники форсайта: *«Конфиденциальность данных — это, безусловно, сейчас самая важная область. Всё больше данных создается, всё больше нарушений приватности происходит. Эта работа по усилению мер защиты данных должна вестись непрерывно, но прорывы возможны уже через несколько лет».*

Конфиденциальность и безопасность данных в ИИ являются ярким примером такой области, где ИИ-технологии одновременно создают угрозы и дают средства защиты от них. Если инструменты ИИ и машинного обучения смогут дать гарантии конфиденциальности данных, это сделает возможным соблюдение нормативных требований и укрепит доверие пользователей, в частности, в таких областях, как здравоохранение, финансы и электронная коммерция, где критически важно соблюдение конфиденциальности данных.

Участники форсайта отмечают, что интеграция методов дифференциальной приватности (differential privacy, DP) и тестирования на утечки (leakage) должны стать стандартом в разработке обучающих пайплайнов. Подобные технологии помогают соблюдать требования законодательства о конфиденциальности (к примеру, GDPR) и стандарты, такие как ISO/IEC 42001, предоставляя количественные гарантии конфиденциальности. При этом необходимо учитывать и исследовать различные схемы атак на приватность данных в моделях машинного обучения. Например, при использовании т. н. атак определения принадлежности (Membership Inference Attacks, MIA) злоумышленник, наблюдая за выходными данными обученной модели, пытается определить, входила ли некоторая запись данных в обучающий набор. В этом контексте актуальной исследовательской задачей является не только борьба с подобными атаками на основе DP и других методов, но и соревновательное создание новых атак, способствующее дальнейшей разработке более совершенных систем защиты.

Генерация синтетических данных на основе приватных реальных данных также может быть способом защиты их конфиденциальности, хотя и за счет неко-

торой потери уникальности данных. Это особенно важно в чувствительных областях, таких как медицина.

4. Перспективные исследовательские задачи

★ ЗАДАЧА 3.1

Управление данными, оценка их качества, а также методы фильтрации, курирования и сортировки

Управление данными служит важнейшей основой для обеспечения качества, безопасности и удобства использования данных при разработке современного ИИ. Поскольку масштабы данных быстро расширяются, а сценарии применения ИИ становятся всё более сложными, проекты ИИ, в которых отсутствует систематическое управление данными, часто сталкиваются с многочисленными проблемами, такими как несогласованность данных, ошибки в маркировке, утечки конфиденциальной информации и риски несоблюдения требований законодательства. Эффективное управление данными не только обеспечивает высококачественные, стандартизированные и отслеживаемые источники данных для обучения моделей, но и устанавливает четкие механизмы подотчетности и стандарты контроля качества на протяжении всего жизненного цикла создания, маркировки, интеграции и использования данных. Управление данными, особенно в средах с несколькими источниками, гетерогенных и крупномасштабных данных, значительно повышает надежность данных, возможность их повторного использования и соблюдение этических норм с помощью таких методов, как управление метаданными, отслеживание происхождения, контроль доступа и мониторинг качества.

Оценка качества данных (DQA) производится по критериям точности, полноты, согласованности, достоверности и уникальности.

Фильтрация данных предполагает удаление нерелевантной, ошибочной или зашумленной информации. Кроме того, фильтрация данных со времен работы Textbooks Are All You Need является одним из мощных инструментов преодоления законов масштабирования, позволяющим существенно уменьшить объемы данных, время обучения и вычислительный бюджет при том же достижимом уровне результатов.

★ ЗАДАЧА 3.2

Конфиденциальность данных в технологиях федеративного обучения

Федеративное обучение (FL) — это специализированная часть более широкой парадигмы распределенного обучения, предполагающая обучение модели одновременно на нескольких серверах без необходимости централизации данных или обмена исходными данными. Локальные серверы обучают локальные копии модели и передают центральному серверу только информацию, необходимую для обновления модели (например, градиенты или веса). Общая модель обновляется, после чего ее копии вновь отправляются обратно на локальные серверы, и цикл обучения повторяется.

Хранение данных на локальных устройствах решает проблему конфиденциальности данных и соответствия требованиям регуляторов по защите данных. Это также снижает риск утечек данных, поскольку создание больших централизованных наборов данных всегда связано с их потенциальной большей уязвимостью.

Федеративное обучение позволяет множеству организаций, имеющих собственные банки приватных данных, участвовать в создании инновационных решений ИИ. При этом риски разглашения конфиденциальной информации для них значительно сокращаются, а при одновременном использовании дополнительных механизмов защиты типа DP могут быть практически сведены к минимуму. Это чрезвычайно важно в таких отраслях, как телекоммуникации, здравоохранение и промышленное производство, где конфиденциальность данных и соответствие требованиям регуляторов имеют решающее значение.

★ ЗАДАЧА 3.3

Конфиденциальность данных в технологиях федеративного обучения

Обнаружение и выделение искусственно сгенерированного контента исторически имеет несколько аспектов: выявление и маркировка сфальсифицированных или искусственно созданных медиафайлов, анализ текстовой информации с целью выявления ее искусственного происхождения и определение наличия синтетической информации в обучающих выборках.

Обнаружение дипфейков традиционно основано на поисках свидетельств манипуляций с данными, таких как визуальные артефакты, временные нарушения, уникальные характеристики камер или микрофонов. Однако всё чаще требуется выявлять, скорее, «авторский стиль» генеративных моделей, которые не оставляют явных несоответствий в генерируемых данных.

Обнаружение сгенерированного текста методами NLP может опираться на статистический анализ,

лингвистические признаки, а также информационные характеристики текста, к примеру сложность и перплексия. В последнее время характеристики текстов, созданных LLM и человеком, всё сложнее различить, поэтому требуется находить новые способы определения синтетики, основанные на методах машинного обучения.

Маркировка синтетического контента т. н. водяными знаками (watermarking) предполагает намеренное встраивание тех или иных незаметных искусственных паттернов в выходные данные генеративных моделей ИИ, что позволяет впоследствии обнаруживать их с высокой точностью.

При этом возможно либо совместное создание данных и водяных знаков, либо их последовательное создание. Первый вариант более перспективен, поскольку генерируемые данные могут сразу содержать информацию о водяных знаках, но второй подход пока используется чаще.

Решение данного круга задач чрезвычайно важно для создания эффективных средств борьбы с дезинформацией, которая может иметь серьезные социальные последствия. Технологии дипфейк также связаны с экономическим мошенничеством (фишинг, фальшивые звонки и видеосообщения), а технологии обнаружения призваны помочь защитить компании и людей от финансовых потерь, утечек данных и ущерба для репутации. Также есть целый ряд областей (например, образование), где важно подтвердить, что текст написан человеком (допустим, в студенческих эссе). Проблема загрязнения обучающих данных связана с тем, что для обучения часто используются данные, извлеченные из интернета. Участники форсайта неоднократно отмечали: *«Крайне важно уметь выявлять синтетические данные и оценивать их воздействие на процесс обучения»*.

5. Важные выводы: Экспертное заключение

Научное сообщество в области ИИ подчеркивает необходимость открытого сотрудничества, расширения исследовательской повестки в работе с данными, а также особое внимание к этическим и регуляторным аспектам для обеспечения ответственного развития и внедрения технологий ИИ.

Большинство участников согласно с тем, что качество данных является важным ограничителем развития ИИ. Надежность данных — основа надежности моделей. Обеспечение конфиденциальности данных — критически важная и чрезвычайно актуальная область исследований. Создание и использование собственных эталонных наборов данных необходимо для проведения исследований и борьбы с предвзятостью. В фокусе должны быть подходы, гарантирующие целостность и достоверность синтетически сгенерированных данных. При этом крайне важно уметь выявлять синтетические данные не только для предотвращения угроз типа дипфейков, но и для оценки их воздействия на процесс обучения.

Существует консенсус о том, что кооперация и совместное использование данных ускоряют развитие и особенно важны для стран с ограниченными цифровыми ресурсами, которые стремятся разрабатывать собственные модели ИИ. Однако ответственное применение технологий ИИ требует формализованных механизмов, таких как сертификация.
Международная стандартизация и сертификация методов бенчмаркинга (тестирования и оценки систем AI) имеет решающее значение для взаимного признания разработок ИИ и эффективной коммуникации в данной области. При этом мягкие регуляторные рамки для технологий ИИ предпочтительнее жесткого законодательства: *«Нужно начинать с создания регуляторных рамок… Может быть, не законы в строгом смысле, а именно framework, который позволит упорядочить взаимодействие с ИИ»*.

Бенчмарки в современном ИИ — основной способ постановки новых исследовательских задач

47% **задач направления непосредственно связаны с вопросами формирования, преобразования и сопровождения данных**

30% **связаны с обеспечением безопасности и конфиденциальности данных**

Качество данных является важным ограничителем развития ИИ, а надежность данных — основа надежности моделей

Критический вызовов современного ИИ — проблема «стены данных». Возможно ли дальнейшее улучшение или максимальный уровень развития ИИ уже достигнут, поскольку он ограничен имеющимся объемом данных, произведенных человечеством

Ответственное развитие направления требует открытого сотрудничества, расширения исследовательской повестки в работе с данными, а также особого внимания к этическим и регуляторным аспектам

НАПРАВЛЕНИЕ 4

Фундаментальные генеративные модели

НАПРАВЛЕНИЕ 4

Фундаментальные генеративные модели

1. Краткое описание направления

Область фундаментальных и генеративных моделей характеризуется стремительным развитием, высокими требованиями к вычислительным ресурсам и быстро меняющимися исследовательскими приоритетами. Ключевые тенденции — развитие способностей генеративных моделей для расширения сфер их применения, а также возрастающая роль синтетических данных для обучения моделей. Этот сдвиг обусловлен дефицитом новых исходных данных и стремлением превзойти человеческий уровень производительности.

В настоящее время особое внимание уделяется следующим поднаправлениям фундаментальных и генеративных моделей:

1. Фундаментальные генеративные модели для символьных данных

Фундаментальные и генеративные модели для символьных данных лежат в основе современного развития больших языковых моделей (LLM). Основное направление связано с развитием систем, способных к осмысленной генерации и интерпретации сложных символьных структур, включая логические рассуждения, программный код и формализованные знания. Исследования ведутся в области повышения фактической точности с помощью таких механизмов, как Retrieval Augmented Generation (RAG), и оптимизации производительности через методы ускоренного инференса (например, speculative decoding).

2. Фундаментальные генеративные модели для несимвольных данных

Ключевую роль играют диффузионные модели и методы на основе нейронных полей, включая NeRF и Gaussian Splatting. Они позволяют моделировать сложные физические и визуальные процессы, создавая правдоподобные изображения, трехмерные сцены и анимации. Наряду с этим развивается направление ко-оптимизации алгоритмов и аппаратных средств — от сенсоров до процессоров.

3. Мультимодальные фундаментальные генеративные модели

Развитие мультимодальных моделей критически важно для создания антропоморфных роботов, способных к естественному взаимодействию с людьми и окружающей средой. Эти модели позволяют роботам одновременно обрабатывать и интерпретировать данные от различных сенсоров (зрение, слух, тактильные датчики), что необходимо для выполнения сложных задач в динамичных реальных условиях. Без подобных технологий невозможно достичь полноценного социального взаимодействия, автономного принятия решений и адаптивного поведения в непредсказуемых ситуациях. Таким образом, мультимодальные модели являются ключевым элементом в преодолении разрыва между ограниченными автоматизированными системами и роботами будущего.

4. Трансфер знаний и адаптация фундаментальных генеративных моделей

Важным направлением развития становится адаптация и перенос знаний фундаментальных моделей в новые домены и задачи. Исследуются методы дообучения и параметрической адаптации (fine-tuning, LoRA, Mixture of Experts), обеспечивающие эффективное использование уже обученных моделей в областях с ограниченными данными. Развиваются подходы к переносу знаний между модальностями и доменами, в т. ч. в условиях непрерывного обучения (continual learning).

5. Аугментация фундаментальных генеративных моделей

Современные исследования направлены и на расширение возможностей самих генеративных моделей через механизмы аугментации. Это включает подключение внешней памяти и инструментов (tool use, memory augmentation), использование внешних баз знаний и симуляторов, а также создание систем, способных к самодиагностике и самообучению. Аугментация также охватывает использование синтетических данных для расширения обучающих наборов и повышения устойчивости моделей.

Современная эпоха в развитии больших языковых моделей берет свое начало в 2017 г., ознаменовавшись

Особая благодарность за работу над данным направлением
А. Кузнецову, Х. Хаовэнь, Е. Тянь

появлением архитектуры трансформеров и механизма внимания. Эта революционная инновация послужила катализатором для стремительного прогресса в области генерации и обработки текста.

- **Прорыв ChatGPT (2022):** Релиз ChatGPT ознаменовал переход больших языковых моделей. Появление ChatGPT и аналогов (2022) стало поворотным моментом: взрывной рост интереса к генеративным моделям сместил фокус с узких NLP-задач на универсальные диалоговые системы.
- **Мультимодальность LLM:** добавление визуальных, аудио- и видеомодальностей расширило их применение: модели научились распознавать изображения, анализировать видео и рассуждать над сложными вопросами, решая широкий спектр задач.
- **Embodied AI (воплощенный ИИ)** обеспечил применение мультимодальных моделей в робототехнике и автономных системах (Vision Language Action), что открыло путь к более интеллектуальным агентам.
- **Диффузионные модели** стали основой генерации изображений и видео, превзойдя GAN. Модели вроде Stable Diffusion и Sora задали новый стандарт качества и стимулировали развитие направлений вроде генерации синтетических данных и RL-тюнинга визуальных моделей.
- **Рост open source-инициатив** (LLaMA и др.) демократизировал доступ к передовым технологиям, позволив создавать национальные и доменные модели, адаптированные под языковые и культурные особенности.

2. Обзор текущего развития направления

Современный ландшафт фундаментальных и генеративных моделей динамичен. Его отличают стремление к универсальным возможностям, рост мультимодальности и интеграция моделей в широкий спектр приложений.

Большие языковые модели (LLM)

Данный вид моделей ИИ всё шире применяется для обработки текста и взаимодействия с пользователем, включая интеграцию с аппаратными системами для выполнения реальных действий. Развиваются методы расширения их возможностей без полного переобучения, в частности Retrieval Augmented Generation (RAG), обеспечивающий актуальность и точность знаний. Параллельно ведутся исследования по оптимизации prompt-инженерии и ускорению инференса (speculative decoding, Mixture of Experts), что повышает эффективность и масштабируемость систем.

Мультимодальность как ключевое направление

Развитие мультимодальных моделей, способных обрабатывать и генерировать данные в разных модальностях (текст, изображение, видео, аудио), является одной из ведущих тенденций. Сюда входят технологии понимания изображений, аудио и видео, обработки длинного контекста, понимание пространственных и пространственно-временных взаимосвязей объектов, генерация изображений и видео, которые часто интегрируются в единую архитектуру.

Диффузионные и другие модели для мультимедийных данных

Мультимодальные генеративные модели стремительно развиваются, демонстрируя возрастающую способность синтезировать контент на основе различных модальностей, например генерировать высококачественные изображения и видео по текстовым описаниям. Диффузионные модели стали краеугольным камнем этого прогресса, обеспечивая беспрецедентное качество и управляемость генерации мультимедийных данных. Их применение выходит за пределы креативных и медийных сфер — они используются в робототехнике для планирования действий на основе восприятия, симуляций и моделирования окружающей среды. Эти достижения означают трансформационный скачок в способности ИИ интерпретировать и воспроизводить сложные сенсорные данные реального мира.

Мультимодальные модели и антропоморфные роботы

Ключевыми проблемами в разработке гуманоидных роботов являются создание оптимальных мультимодальных сенсорных систем с эффективным слиянием данных и обеспечение обработки информации в реальном времени при ограниченных вычислительных ресурсах. Параллельно актуальными задачами остаются преодоление дефицита специализированных данных, достижение социально приемлемого взаимодействия без эффекта «зловещей долины», а также обеспечение способности к обобщению и адаптации в разнообразных средах без масштабного переобучения.

Современные подходы к моделированию физики и геометрии мира сталкиваются с фундаментальными вызовами, включая нарушение физических законов

data-driven-моделями, ограниченную обобщающую способность для внераспределительных данных и недостаточную масштабируемость для сложных сред. Критически важными направлениями развития являются интеграция физических priors в архитектуры моделей, обеспечение устойчивости к неизвестным сценариям, создание эффективных вычислительных методов для крупномасштабного моделирования, а также достижение интерпретируемости и способности к кросс-модальным причинно-следственным рассуждениям. Интеграция с поиском и гибридные подходы: расширяется применение гибридных подходов, объединяющих генеративные модели с алгоритмами поиска и внешними инструментами для повышения качества рассуждений и генерации данных. Перспективным направлением становятся эволюционные методы, подобные AlphaEvolve, где языковые модели сочетаются с эволюционными алгоритмами для автоматического создания и улучшения кода, что позволяет формировать сложные интерпретируемые решения без участия человека.

3. Исследовательские вызовы, стимулирующие или ограничивающие развитие направления

ВЫЗОВ 4.1

Галлюцинации фундаментальных и генеративных моделей

Одним из главных вызовов остаются галлюцинации — генерация моделями правдоподобной, но неверной информации, что подрывает доверие и ограничивает применение ИИ в ответственных сферах, таких как медицина и госуправление. Преодоление этой проблемы требует разработки объективных методов оценки качества и снижения числа ошибок. В мире активно исследуются подходы, повышающие надежность, включая Retrieval Augmented Generation (RAG), а также гибридные и нейросимвольные методы, направленные на создание прозрачных и доверенных ИИ-систем.

ВЫЗОВ 4.2

Вычислительные затраты и эффективность

Обучение и использование крупных моделей требуют огромных вычислительных ресурсов и энергии, что становится главным барьером для масштабирования и массового применения ИИ. Высокие затраты ограничивают доступ к технологиям и замедляют прогресс, особенно в задачах, требующих работы в реальном времени. В ответ на этот вызов развиваются энергоэффективные архитектуры (например, Mixture of Experts), методы оптимизации (квантизация, дистилляция) и специализированные чипы (TPU, NPU), а также системы распределенных вычислений, направленные на снижение затрат и повышение доступности ИИ.

ВЫЗОВ 4.3

Генерализация и перенос знаний

Несмотря на успехи в решении отдельных задач, модели по-прежнему слабо обобщают знания и плохо работают в новых условиях, не встречавшихся при обучении. Ограниченная генерализация мешает созданию универсальных и автономных систем, способных адаптироваться без переобучения. Для преодоления этого разрыва развиваются методы transfer learning, domain adaptation, гибридные и эволюционные подходы, такие как AlphaEvolve, а также интеграция априорных знаний, включая физические закономерности, для повышения устойчивости и адаптивности моделей.

4. Перспективные исследовательские задачи

★ ЗАДАЧА 4.1

Создание вычислительно эффективных архитектур фундаментальных генеративных моделей

Разработка методов, позволяющих обучать модели сопоставимой сложности при значительно меньших объемах данных и вычислений. Основная проблема современных систем — высокая ресурсоемкость обучения; задача состоит в повышении его экономичности без потери качества. Решение предполагает использование более быстрых оптимизаторов, переноса знаний (knowledge distillation), генерации синтетических данных, архитектур с разреженными параметрами (Mixture of Experts) и непрерывного обучения. Важным направлением остается ко-дизайн алгоритмов и аппаратуры для максимальной производительности. В результате решения станет возможным обучение моделей того же уровня за счет в десятки раз меньших ресурсов, что сделает разработку фундаментальных моделей доступной для широкого круга исследовательских групп.

★ ЗАДАЧА 4.2

Исследование и разработка методов создания генеративных моделей (включая RL в различных приложениях)

Современные фундаментальные модели достигли предела по объему доступных «человеческих» данных, поэтому ключевая задача — научить их самостоятельно генерировать качественные синтетические данные и обучаться на них с использованием обратной связи от среды. Для этого применяются методы обучения с подкреплением (RL), где модели взаимодействуют с симуляторами или реальным миром и улучшают стратегии на основе вознаграждения (пример — DeepSeek R1), а также гибридные подходы с поисковыми алгоритмами (MCTS) и обучение на неразмеченных данных. Решение этой задачи приведет к появлению самосовершенствующихся моделей, способных обучаться без постоянного притока новых данных, что обеспечит качественный рост их способностей к рассуждению, планированию и автономному действию, в т. ч. применение в робототехнике и агентных системах.

★ ЗАДАЧА 4.3

Разработка методов файнтюнинга фундаментальных генеративных моделей (например, LoRA, P-tuning)

Фундаментальные модели содержат обширные знания, но их адаптация к новым доменам (допустим, медицина или право) требует дорогостоящего дообучения и часто приводит к потере ранее усвоенной информации. Необходимо создавать эффективные методы переноса и интеграции знаний без полного переобучения. Для этого применяются адаптеры и LoRA, позволяющие обучать отдельные модули без изменения основной модели, а также подходы вроде RAG, использующие внешние базы знаний. Перспективны дистилляция знаний для переноса информации в компактные модели и методы «разучивания», позволяющие удалять нежелательные данные. Результатом станет развитие гибких и экономичных технологий адаптации, обеспечивающих создание точных и специализированных доменных и национальных моделей при соблюдении требований по защите данных.

5. Важные выводы: Экспертное заключение

Данное направление можно охарактеризовать как один из локомотивов современного ИИ: его достижения быстро транслируются во все остальные области и порождают новые подходы к решению старых задач.

Экспертное сообщество сходится во мнении, что будущее развитие фундаментальных и генеративных моделей будет определяться гибридными подходами. Ожидается, что наибольший прогресс будет достигнут за счет совмещения машинного обучения с символическим рассуждением, а также за счет создания мультиагентных систем, способных сложно взаимодействовать как друг с другом, так и с окружающей средой.

Несмотря на впечатляющую мощь больших фундаментальных моделей, признается критическая необходимость их адаптации и дообучения (локализации). Это касается не только поддержки малоресурсных языков и учета культурных контекстов, но и применения в специализированных областях, таких как медицина или инженерия, где требуются специфические знания.

Основой для устойчивого прогресса считаются стандартизированные и открытые бенчмарки. Они необходимы для объективной оценки моделей, обеспечения воспроизводимости исследований и достоверного сопоставления результатов. Преодоление недостатка надежных метрик, особенно для оценки генеративных текстов, остается серьезным вызовом. Следует также отдельно подсветить проблему оценивания творческих задач без правильного варианта ответа. Каким образом мы можем определить, что один текст лучше другого, или одно изображение качественнее другого. Для решения этой задачи активно развивается обучение моделей-критиков (больших языковых или мультимодальных моделей), выступающих в качестве судей при оценивании сгенерированного контента.

Быстрорастущее направление, потенциал прикладных приложения которого сложно переоценить

Важнейший тренд развития направления — мультимодальность, значимость исследований здесь существенно вырастет в ближайшие годы

34% **задач направления напрямую связаны с текстовыми и символьными данными, еще**

41% **— косвенно**

Важнейшим прорывом будет создание моделей, понимающих физику и геометрию мира, это откроет целый спектр новых возможностей и приложений и может затронуть развитие практически всех остальных направлений

Одним из наиболее серьезных вызовов для генеративных моделей остается проблема галлюцинаций — генерации ложной или неподтвержденной фактами информации

НАПРАВЛЕНИЕ 5

Безопасность, доверие и объяснимость

НАПРАВЛЕНИЕ 5

Безопасность, доверие и объяснимость

1. Краткое описание направления

Направление «Безопасность, доверие и объяснимость ИИ» охватывает комплекс задач, связанных с разработкой и эксплуатацией интеллектуальных систем, обеспечивающих их надежность, предсказуемость и социальную приемлемость. Оно формирует границы применимости технологий ИИ: именно степень доверия со стороны общества, государства и бизнеса определяет возможность их интеграции в критически важные сферы.

В отличие от традиционного понимания безопасности, ограничивавшегося лишь технической исправностью программных средств, современный подход интегрирует вопросы устойчивости моделей, их согласованности с человеческими ценностями и способность предоставлять прозрачные, верифицируемые и воспроизводимые основания для принимаемых решений.

Исследования в данной области носят междисциплинарный характер, объединяя методы информатики, математики, кибербезопасности, права, этики и социальных наук. Взгляд на развитие ИИ через призму доверия позволяет не только выработать прикладные решения, но и сформулировать новые фундаментальные проблемы ИИ, требующие формирования целостной теории и инженерной практики его безопасного функционирования.

В рамках данного направления выделяются следующие поднаправления:

1. Выравнивание целей (Alignment)

Задача выравнивания заключается в предотвращении генерации моделями ИИ вредоносных результатов — недостоверных, противоречащих ценностям общества, нарушающих законодательство или направленных на противоправные действия. Для этого создаются методы дообучения моделей, внедрения дополнительных уровней контроля, тестирования и бенчмарков, а также механизмы формирования ценностных установок в генеративных моделях широкого применения.

2. Объяснимость работы технологий ИИ (XAI)

Объяснимость обеспечивает прозрачность и доверие при использовании ИИ в критически важных сферах, таких как медицина, управление или юриспруденция. Исследования включают разработку методов постфактум-объяснения «черных ящиков», создание прозрачных моделей на основе логики и баз знаний, а также формирование требований и протоколов тестирования систем ИИ на объяснимость.

3. Обеспечение безопасной разработки и эксплуатации технологий ИИ

Для интеграции ИИ в критически важные системы необходимо формирование целостной теоретической и технологической базы, обеспечивающей доверие к интеллектуальным системам. Она должна включать не только проверку программного кода и используемых библиотек на отсутствие уязвимостей, но и выявление и устранение дефектов в наборах данных и моделях, характерных именно для ИИ. Ключевыми направлениями являются защита данных и моделей от атак, развитие методов безопасной разработки (MLSecOps), а также создание специализированных инструментов и бенчмарков для оценки уровня безопасности и степени доверия к системам ИИ.

4. Обеспечение защиты от результатов использования ИИ с целью взлома

ИИ может применяться для взломов, кибератак, социальной инженерии и подделки информации. Важными направлениями исследований являются выявление уязвимостей и дипфейков, разработка методов защиты данных, включая технологии цифровых водяных знаков, и создание средств противодействия противоправному использованию интеллектуальных систем.

Изначально безопасность систем ИИ понималась исключительно как техническая корректность и устойчивость к сбоям. Основные дискуссии сводились к определению специфики ИИ в рамках традиционной парадигмы безопасной разработки программного обеспечения. Однако довольно быстро стало очевидно, что методы анализа программного кода не позволяют выявлять дефекты в наборах данных

Особая благодарность за работу над данным направлением
А. Аветисяну, А. Роше , Е. Рогову

и моделях машинного обучения, что потребовало разработки новых специализированных подходов и инструментов.

С расширением применения ИИ в социально и экономически значимых сферах начали проявляться системные риски, связанные с дискриминацией, уязвимостью к атакам, недостаточной предсказуемостью и непрозрачностью принимаемых решений. Это вызвало расширение фокуса внимания исследователей и разработчиков: от проверки того, работает ли система технически правильно, к более широкому пониманию — обеспечивает ли она безопасность и справедливость для общества.

Среди ключевых событий, ставших основой развития данного направления, выделяются следующие:

- **Массовое внедрение генеративного ИИ** продемонстрировало потенциал технологий, но выявило проблемы галлюцинаций, недетерминированности и злоупотреблений (deepfake, деструктивный контент). Расширение агентных возможностей (инструменты, память, веб-доступ) резко увеличило поверхность атаки и частоту нарушений политик поведения моделей.
- **Аварии автономного транспорта** подтвердили необходимость разработки новых методов обеспечения функциональной надежности в условиях неопределенности.
- **Выявленные случаи дискриминации** в системах подбора персонала и кредитного скоринга обострили внимание к вопросам справедливости и предвзятости алгоритмов.
- **Начало разработки международных стандартов и законодательства (EU AI Act, инициативы IEEE)** усилило институциональное регулирование и сформировало базовые требования к прозрачности и безопасности систем. Развивающиеся требования смещают акцент в сторону аудита, сертификации и эксплуатационного мониторинга ИИ-систем.
- **Ужесточение норм защиты персональных данных (GDPR и аналоги в других странах)** стимулировало развитие методов дифференциальной приватности и федеративного обучения.
- **Недавние исследования в области безопасности ИИ подчеркнули ключевую роль устойчивости к состязательным атакам (adversarial robustness),** показали, что защита моделей от подобных воздействий является необходимым условием их безопасного применения в критически важных областях.
- **Появление скрытых распределенных инструкций во внешних данных** стало одной из новых центральных угроз для автономных агентных ИИ. Высокая эффективность подобных атак требует фундаментальных и многоуровневых адаптаций политик разработки, интеграции и эксплуатации ИИ-моделей.

Направление «Безопасность, доверие и объяснимость ИИ» имеет критическое значение для устойчивого внедрения ИИ в социально и экономически значимые сферы. Доверие общества, бизнеса и государства к интеллектуальным системам определяет границы их применения: без уверенности в надежности и справедливости решений ИИ его потенциал остается ограниченным. Разработка методов безопасного и объяснимого ИИ обеспечивает возможность более широкого внедрения технологий, снижая риски ошибок и злоупотреблений.

2. Обзор текущего развития направления

В последние несколько лет в области доверенного ИИ наблюдается переход от концептуальных дискуссий к внедрению практических инструментов. Широкое распространение генеративных моделей актуализировало задачи фильтрации входных и выходных данных, внедрения цифровых водяных знаков для защиты от злоупотреблений и разработки методов мониторинга моделей в эксплуатации. В направлении объяснимости параллельно развиваются архитектурные подходы (создание интерпретируемых моделей и методов) и методы постфактум-объяснения результатов пользователю (визуализация и анализ решений уже обученных систем).

Создаются и совершенствуются программные инструменты разработки безопасных систем ИИ по аналогии с инструментами разработки безопасного ПО, отрабатываются методики их применения. Эти инструменты включают в себя как средства анализа обучающих данных, так и средства тестирования безопасности обученных моделей. Ведущие организации создают специализированные команды по безопасности ИИ, которые проводят стресс-тестирование моделей, имитируя злонамеренные атаки, попытки обхода ограничений и эксплуатацию скрытых уязвимостей. Цель — выявить и устранить риски до публичного развертывания модели.

Разрабатываются и развиваются средства и методы, обеспечивающие безопасное функционирование моделей ИИ. Сюда входит фильтрация входа и выхода модели (цензурирование), обнаружение и предотвращения атак, аномалий, утечек информации,

попыток кражи (несанкционированной дистилляции) модели ИИ, меры дополняются регламентами реагирования на инциденты и журналированием решений моделей. Могут применяться как классические, неинтеллектуальные средства, так и модели ИИ, специально обученные обеспечивать безопасность других моделей ИИ.

Всё большее внимание уделяется не только итоговому решению модели, но и степени ее уверенности в этом решении. Разрабатываются методы, позволяющие модели калибровать свои предсказания и сигнализировать о низкой уверенности в незнакомых или пограничных ситуациях.

Обучение с подкреплением на основе человеческих предпочтений (RLHF) стало де-факто стандартом для выравнивания крупных языковых моделей. Всё шире применяются технологии обучения на основе предпочтений модели (RLAIF), когда выравнивание одной модели ИИ реализуется другой моделью ИИ. Ведутся исследования в области наделения моделей ИИ способностью оценивать обоснованность, безопасность и этичность собственных суждений в ходе их генерации.

В области объяснимого ИИ выделяют два направления:

1. Интерпретируемые модели, где прозрачность заложена на этапе проектирования: это модели, в которых процесс принятия решений интуитивно понятен человеку. К ним относятся деревья решений, линейные модели и гибридные архитектуры с визуализируемыми внутренними зависимостями.

2. Постфактум-объяснение, ориентированное на разработку методов интерпретации уже обученных моделей, включая сложные нейронные сети. Здесь применяются такие подходы, как SHAP, LIME, визуализация attention-механизмы и генерация объяснений для пользователей и разработчиков.

При этом важно учитывать, что полная и очевидная интерпретируемость работы современных моделей без потери их качества практически недостижима; задача состоит не в полной прозрачности, а в обеспечении достаточной объяснимости и доверия при сохранении высокой эффективности.

Оба направления дополняют друг друга: первое обеспечивает «прозрачность из коробки», а второе позволяет интерпретировать сложные модели, где встроенная интерпретируемость ограничена.

Реальные испытания агентных ИИ показывают критические уязвимости: в ходе крупномасштабного публичного red team не более чем 1,5 млн обращений к модели было зафиксировано свыше 60 тыс. успешных нарушений политик поведения в 44 сценариях для 22 моделей. Особенно эффективны косвенные инъекции через сторонние загружаемые данные, такие как веб-страницы, документы и письма электронной почты.

3. Исследовательские вызовы, стимулирующие или ограничивающие развитие направления

ВЫЗОВ 5.1

Отсутствие строгой формальной теории машинного обучения

Современный ИИ в значительной степени развивается в парадигме поиска новых решений методом проб и ошибок. Отсутствие строгой формальной теории ограничивает возможности прогнозировать границы применимости технологий, определять условия их надежности и безопасности. Без фундаментальной базы невозможно ответить на ключевой вопрос: где и при каких условиях современные методы перестают быть эффективными и безопасными. В ответ на вызов усиливаются исследования в области математических основ машинного обучения, теории генеративных моделей и формализации понятий доверия и устойчивости.

ВЫЗОВ 5.2

Выравнивание целей (Alignment)

Основной вызов связан с обеспечением соответствия поведения генеративного ИИ социальным ценностям и правовым нормам. Угроза заключается в том, что модели могут генерировать недостоверные, запрещенные или опасные результаты, а также использоваться для противоправных действий. Нерешенность этой проблемы подрывает доверие к ИИ и ограничивает его внедрение в социально значимые сферы. В ответ на вызов в мире ведутся работы по созданию методов ценностного выравнивания, систем тестирования и бенчмарков для оценки надежности моделей.

ВЫЗОВ 5.3

Объяснимость и прозрачность решений

Сложность современных нейросетевых моделей препятствует пониманию их внутренней логики. Это вызывает риски при принятии ответственных решений в медицине, юриспруденции и управлении. Отсутствие объяснимости ограничивает практическую применимость ИИ и усиливает общественный скепсис. Для преодоления вызова разрабатываются

методы создания интерпретируемых моделей, методы постфактум-объяснения (SHAP, LIME и др.), а также формируются стандарты и протоколы оценки уровня объяснимости.

⚡ ВЫЗОВ 5.4

Безопасная разработка и эксплуатация (MLSecOps)

Рост масштабов применения ИИ усиливает угрозы, связанные с внедрением уязвимостей в код, библиотеки и наборы данных. Атаки на модели во время обучения и исполнения (отравление наборов данных, состязательные атаки на модели) способны приводить к системным сбоям и компрометации критически важных систем. В ответ формируются подходы безопасной инженерии ИИ (MLSecOps), направленные на обеспечение защищенности всего жизненного цикла моделей. Разрабатываются новые методы противодействия косвенным инъекциям на этапах интеграции инструментов, включая анализ цепочек вызовов и фильтрацию вводимых данных. Активно развивается бенчмаркинг устойчивости к агентным атакам с использованием эталонных наборов атак и регулярными переоценками надежности моделей. Одновременно формируются политики безопасной работы с недоверенными источниками, предусматривающие создание изолированных «песочниц», использование списков разрешенных доменов и внедрение автоматизированных проверок контента перед его загрузкой в модель.

⚡ ВЫЗОВ 5.5

Противодействие злоупотреблениям ИИ

Модели ИИ могут использоваться в противоправных целях — от взлома и социальной инженерии до создания дипфейков и поддельной информации. Это угрожает безопасности личности, организаций и государства. Вызов требует разработки методов защиты данных (например, с использованием водяных меток), алгоритмов выявления фальсификаций и систем противодействия кибератакам.

4. Перспективные исследовательские задачи

★ ЗАДАЧА 5.1

Разработка методов снижения рисков, связанных с неверными или вредоносными данными (неточные данные, данные с ограниченным доступом, включая информацию, распространение которой запрещено в соответствии с законом, данные, не соответствующие ценностям общества)

Для решений поставленной задачи необходима разработка методов для формирования ценностных установок в генеративных моделях, тестов и бенчмарков, а также проверка устойчивости моделей к уязвимостям. Для систем агентного ИИ выравнивание должно распространяться на цепочки действий с инструментами (tool-use), включая оценку безопасности намерений и побочных эффектов до совершения вызовов инструментов. При этом важно учитывать влияние оптимизационных процедур при разработке, а также устойчивость по отдельным доменам знаний при таких действиях.

★ ЗАДАЧА 5.2

Формулировка общих подходов для обеспечения объяснимости, повышения доверия работы ИИ — Explainable AI (XAI)

Наметился тренд по исследованию и созданию методов прозрачного ИИ на основе формальных систем, логики и баз знаний. Также ведется формализация требований, протоколов, показателей оценки и бенчмарков тестирования систем на объяснимость. Важными являются и операционные объяснения для пользователя или аудитора: почему был вызван инструмент, какие ограничения политики учитывались, как оценивались риски.

★ ЗАДАЧА 5.3

Создание методов и инфраструктуры для обеспечения безопасной разработки систем с ИИ (MLSecOps)

В рамках данной задачи ведутся работы по разработке и эксплуатации принципов построения инфраструктуры безопасной разработки (MLSecOps), разрабатываются методы защиты от отравления наборов данных и атак на модели, а также инструменты поиска уязвимостей в сторонних библиотеках. Ведется создание бенчмарков, позволяющих формализовать уровень доверия к системам ИИ.

★ ЗАДАЧА 5.4

Разработка методов обнаружения и защиты от дипфейков, включая водяные знаки

Разработка методов противодействия дипфейкам становится критически важной задачей в обеспечении цифровой безопасности. Массовое распространение генеративного ИИ привело к тому, что создание

реалистичных поддельных видео- и аудиоматериалов стало доступным даже для неподготовленных злоумышленников, что требует срочного развития эффективных методов защиты. Особую актуальность приобретают технологии цифровых водяных знаков, которые позволяют не только обнаруживать подделки постфактум, но и заранее маркировать легитимный контент, создавая основу для верификации его подлинности. Эти решения становятся особенно востребованными в контексте борьбы с дезинформацией и защиты персональных данных, где возможность оперативного выявления фальсификаций напрямую влияет на защиту прав и свобод пользователей.

В области создания теоретических основ доверенного ИИ ведутся отдельные работы по формированию формальной теории, определяющей границы применимости методов ИИ и критерии их надежности. Создание общей теории позволит перейти от эмпирического поиска решений к системному проектированию безопасных алгоритмов и создаст основу для новых фундаментальных исследований.

5. Важные выводы: Экспертное заключение

Безопасность и доверие становятся системообразующими принципами современного ИИ, определяющими границы его применения в критически важных сферах — от медицины и транспорта до государственного управления. В центре внимания — выравнивание целей моделей с человеческими ценностями, устойчивость к атакам, защита данных и формирование прозрачных механизмов принятия решений. Современная парадигма безопасности ИИ выходит за рамки технической корректности и охватывает весь жизненный цикл разработки: от формирования наборов данных и обучения моделей до их эксплуатации. Развитие методов безопасной инженерии (MLSecOps), верификации и тестирования доверенности моделей, а также протоколов объяснимости создает основу для ответственного внедрения ИИ, способного действовать предсказуемо и справедливо.

Формирование доверенного и объяснимого ИИ имеет не только технологическое, но и институциональное значение.

Ведется активная разработка международных стандартов, протоколов аудита и сертификации систем ИИ, направленных на обеспечение прозрачности, устойчивости и подотчетности. Создание объяснимых моделей и инструментов постфактум-анализа решений способствует укреплению общественного доверия, снижению рисков злоупотреблений и дискриминации. В долгосрочной перспективе направление «Безопасность, доверие и объяснимость ИИ» становится ключевым элементом архитектуры цифрового суверенитета, способствующим формированию глобальной культуры ответственного использования ИИ и развитию безопасной инновационной экономики.

Направление является определяющим для массового внедрения ИИ

27% **тематик направления связаны с «выравниванием» целей ИИ с человеческими, формированием корректных ценностных установок, обеспечивающих базу для дальнейшего развития и использования технологий**

НАПРАВЛЕНИЕ 6

ИИ для узких задач (Narrow AI)

НАПРАВЛЕНИЕ 6

ИИ для узких задач (Narrow AI)

1. Краткое описание направления

Узкоспециализированный ИИ (далее — Task-Specific AI, Narrow AI) — это класс систем ИИ, предназначенных для выполнения конкретных, четко определенных функций. В отличие от общего ИИ (AGI), который стремится имитировать широкий спектр когнитивных возможностей человека, узкий ИИ сфокусирован на решении конкретных задач, что позволяет достигать высокой точности и практической полезности в ограниченных областях.

Его ключевое преимущество — практическая применимость уже сегодня: от промышленного внедрения компьютерного зрения до языковых моделей в государственных сервисах. Task-Specific AI стал основным драйвером технологического прогресса в ИИ благодаря устойчивым экономическим выгодам и прямой пользе для общества и науки. Машинное обучение уже широко применяется в самых разных областях — от разработки ПО и планирования в многоагентных системах до медицинской диагностики, что подчеркивает значимое влияние узких ИИ-решений в промышленности и науке.

Ближайшее будущее ИИ видится не монолитным: ожидается отход от эпохи простого линейного масштабирования трансформеров к эре архитектурного разнообразия. При этом трансформеры остаются фундаментом многих передовых систем: современные открытые LLM (Qwen, LLaMa) используют трансформер, всё чаще в варианте «смеси экспертов» (MoE); в компьютерном зрении фундаментальные модели вроде Segment Anything Model 2 (SAM-2), DINOv1-v3 также основаны на трансформерных архитектурах; AlphaFold3 сочетает трансформер с более сложными специализированными компонентами.

Вместе с этим активно исследуются и альтернативные подходы — от моделей состояний (SSM), в т. ч. гибридных архитектур Mamba + Трансформер (например, Nemotron, Granite), до нейросимвольных систем, «мировых моделей» (world models) и графовых нейронных сетей.

В практических сценариях всё чаще применяются узкоспециализированные решения: от простых эвристических систем (к примеру, чат-боты для приемных кампаний в вузах) до более продвинутых инструментов с поддержкой мультиязычности и поисковой интеграции (RAG). В исследовательской сфере изучается потенциал носимых устройств (wearables) для улучшения человеко-машинного взаимодействия, включая мониторинг стресса по физиологическим данным и дополненную реальность (например, Ray Ban Display glasses).

Исследования в области Task-Specific AI критичны и для социальной устойчивости и справедливости. Недостаточная надежность моделей, включая склонность к галлюцинациям, ограничивает их применение в критически важных и высокорисковых сферах (медицина, транспорт, право). Повышение достоверности и прозрачности повысит общественное доверие и обеспечит безопасную интеграцию технологий в чувствительных доменах. В социальной сфере узкие ИИ-решения уже позволяют автоматизировать обработку обращений граждан, снижая нагрузку на персонал и предотвращая профессиональное выгорание. Параллельно развитие направления способствует демократизации доступа к ИИ и расширению охвата технологий за счет поддержки языков с ограниченными ресурсами, что препятствует усилению социального неравенства.

В рамках данного направления были выделены три поднаправления, охватывающие его основу:

1. Компьютерное зрение (Computer Vision, CV)

Технологии компьютерного зрения достигли промышленного уровня зрелости. Появилось новое направление — генерация видео, где активный прогресс начался около 2023 г.: появились коммерческие проекты, время генерации видео сократилось с часов до минут, а качество продолжает расти благодаря новым архитектурам и методам оптимизации. Модель MagicTime способна лучше моделировать физически обоснованные процессы (например, рост растений), благодаря тренировке на видеозаписях с подробными аннотациями (2000+ клипов). Одним из ключевым направлений в компьютерном зрении является создание фундаментальных моделей, которые обучены на больших корпусах данных и способны с высоким уровнем качества решать задачи компьютерного зрения на пользовательских данных. Следует отметить

Особая благодарность за работу над данным направлением
Д. Юдину, Р. Баэза-Йейтс, Е. Тутубалиной

прогресс в области сегментации и трекинга произвольных объектов, в которой появились модели типа Segment Anything 2, SAM, осуществляется обнаружение объектов по запросам на естественном языке, например, на основе моделей YOLOE, Grounding DINO, оценка позы камеры и карт глубин по видео с помощью фундаментальных трансформерных моделей типа VGGT, даются ответы на вопросы (VQA, Visual Question Answering) и рассуждения (Reasoning) по изображениям с помощью визуально-языковых моделей VLM типа открытых моделей Qwen-VL, доступных по API моделей вроде GPT-4o и GPT5 и др.

2. Обработка естественного языка (Natural Language Processing, NPL)

LLM внедряются повсеместно, например, в государственных сервисах для автоматической обработки обращений граждан, снижая нагрузку на сотрудников и предотвращая выгорание. Многоязычная тренировка может повышать эффективность LLM, используя общие семантические структуры разных языков, однако добавление большого числа языков не всегда дает выигрыш в качестве. На практике некоторые компании, к примеру, Cohere, отказались от 27 языков в пользу шести наиболее важных, что может быть обусловлено как сокращением затрат на сбор данных, так и более стабильными финальными метриками модели.

3. Прочие узкие технологии ИИ

Специализированные методы ИИ активно развиваются за пределами CV и NLP. В разработке программного обеспечения применяются модели для оптимизации компиляторов; зарождается концепция AI for Science для решения задач в фармацевтике, физике, химии, биологии, материаловедении.

2. Обзор текущего развития направления

Компьютерное зрение (CV). Современное CV характеризуется развитием эффективных методов анализа визуальной информации и появлением новых генеративных возможностей. Активно развиваются модели генерации изображений и видео по тексту (text-to-image, text-to-video), а также методы создания и редактирования 3D-контента.

«Генеративные исследования и практические разработки в области ИИ *— это в первую очередь модели text-to-image, text-to-video, text-to-visual…»,* — подчеркнуто в экспертной дискуссии, указывая на ключевую роль диффузионных моделей в подобных задачах. В медицинском зрении широко применяются методы самообучения (self-supervised learning), позволяющие использовать неразмеченные данные, например в радиологии для анализа снимков. Последние достижения в области конформного прогнозирования и контроля конформных рисков, интеграция резюме в рабочие процессы врачей (врачи участвуют в принятии решений с помощью моделей ИИ, а врач остается в курсе событий). Практические системы CV уже внедрены в промышленность и медицину: работают автоматизированные линии сортировки на производстве, а в диагностике используются системы «второго мнения», помогающие врачам при анализе снимков. Разрабатываются и новые способы извлечения информации из изображений: например, гиперспектральная реконструкция позволяет получать точные количественные измерения по изображениям, выходя за рамки сугубо визуальных (эстетических) задач.

Обработка естественного языка (NLP). Наблюдаются мощные достижения в мультизадачных языковых моделях — GPT-5, Claude, Gemini (2023–2025) объединяют мультимодальность, улучшенное логическое мышление и few-shot обучение. Большие языковые модели (LLM) внедряются во множестве сфер. В частности, государственные сервисы начали использовать LLM для обработки обращений граждан, что снижает нагрузку на персонал и помогает предотвратить профессиональное выгорание. Узкоспециализированные NLP-модели находят применение в гуманитарных науках: например, ИИ используется для сохранения культурного наследия через распознавание старинных рукописей и изучение эволюции языка по историческим текстам.

RAG (Retrieval-Augmented Generation). Интеграция поиска в процесс генерации текста: модель перед генерацией ответа запрашивает внешние источники знаний (скажем, документы). Этот подход стал фундаментальным для современных LLM, т. к. позволяет актуализировать знания модели и ссылаться на факты, уменьшая число галлюцинаций. Этот подход критически важен в условиях быстро меняющегося мира, поскольку, если модель обучена на данных предыдущего года, то естественно, что актуальные данные мы уже не получим. В качестве примера применения приводится разработка ИИ-агента для приемной комиссии RAGFlow, который успешно использует RAG-модуль для предоставления актуальных ответов на нескольких языках. Это подчеркивает потенциал RAG в сферах, где критически важен доступ к последним данным, например в образовании, юридической сфере или государственном управлении.

Пошаговое рассуждение (Chain-of-Thought, CoT). Техника пошагового рассуждения в процессе

генерации. Модели побуждают сначала явно писать цепочку логических рассуждений, а затем итоговый ответ. Это существенно улучшило способность LLM решать сложные задачи (математика, логика, бытовые рассуждения). Эффект CoT особенно проявляется у крупных моделей (свыше ~100 млрд параметров) и считается эмерджентным свойством — крупные модели внезапно начинают успешно рассуждать, если их правильно подтолкнуть к цепочке мыслей. Однако остаются вызовы, связанные с галлюцинациями и ограниченностью внутреннего «мышления» моделей. Перспективы развития включают интеграцию внешних инструментов и знаний, а также разработку систем, способных к самообучению и самодиагностике, имитируя человеческое «мышление, память, планирование». Интеграция семантических сетей или графов знаний в ИИ-системы может повысить достоверность рассуждений, обеспечивая доступ к структурированным фактам и позволяя фильтровать противоречивые или ложные генерации. Это направление активно исследуется в рамках нейросимвольного ИИ и RAG-систем.

Структурированный вывод (structured output). Формирование строго структурированного вывода (планы действий, машинный код, 3D-молекулярные структуры или таблицы) по запросу. Для бизнес-применений это критично, т. к. позволяет интегрировать LLM в существующие системы. Однако добиться от модели безусловно корректного форматирования сложно: исследования показывают, что даже передовые LLM иногда сбиваются, особенно на сложных структурах, требуя дальнейших улучшений надежности формата. Разработчики внедряют специальные prompt-шаблоны и методы контроля вывода (например, ограниченное декодирование), чтобы повысить консистентность структурированных ответов. Такие модели способствуют широким инновациям и исследованиям. Интеграция машинного обучения со структурированными правилами, научными знаниями и онтологиями (т. н. нейросимвольный ИИ) признается перспективным подходом. Это позволяет ограничить модели и обеспечить более предсказуемое и надежное поведение, особенно в областях, где требуется высокая точность и интерпретируемость.

Мультиагентность. Создаются мультиагентные системы и автономные ИИ-агенты для прикладных задач: например, узкоспециализированные боты-помощники в университетских приемных комиссиях, которые владеют несколькими языками и умеют извлекать нужные факты из баз знаний (воплощение связки LLM + RAG). Также агентные способности LLM и способности к Tool Calling (вызову специализированных инструментов) позволяют языковым моделям решать любые прикладные задачи, для решения которых можно использовать заданные программные инструменты. К примеру, существуют успешные примеры применения LLM в задачах глубокого поиска информации в интернете (Deep Research), где с помощью вызова поисковика языковая модель синтезирует отчет на заданную ей тему. Гибкость механизма Tool Calling позволяет применить LLM к решению большого количества задач, что является привлекательным направлением для дальнейших исследований.

3. Исследовательские вызовы, стимулирующие или ограничивающие развитие направления

ВЫЗОВ 6.1

Ограниченность и синтетичность данных

Узкоспециализированные модели компьютерного зрения, обработки естественного языка сталкиваются с дефицитом качественных данных в специализированных областях. В медицине трудно собрать выборки по редким диагнозам, а в кибербезопасности — достоверные данные об атаках или биометрическом антиспуфинге. Генерация синтетических данных становится критически важной, однако ее долгосрочные последствия остаются неясными. Дефицит данных ограничивает качество и надежность моделей, снижает возможность их внедрения в практику. Если будет найден эффективный способ замещения реальных данных синтетическими, это может открыть путь к масштабируемости и ускорить внедрение ИИ в сложные домены. В связи с этим активно развиваются методы генерации синтетических датасетов, в т. ч. на основе специализированных симуляторов, формируются консорциумы для обмена редкими медицинскими и промышленными данными. Параллельно исследуются подходы к валидации качества синтетики, чтобы минимизировать риски искажения результатов обучения.

ВЫЗОВ 6.2

Создание эффективных пространственных представлений для решения задач компьютерного зрения

Ключевой проблемой остается отсутствие универсальных методов представления пространственных данных от различных сенсоров для задач компьютерного зрения. Несмотря на развитие мультимодальных моделей и трехмерных представлений (в т. ч. NeRF, Gaussian Splatting), существующие подходы не обеспечивают эффективной интеграции семантических признаков и остаются непригодными для задач, тре-

бующих высокого быстродействия. Сохраняется потребность в создании адаптивных методов, способных объединять информацию от разнородных сенсоров и поддерживать сложные пространственные рассуждения, что необходимо для практического применения в реальных условиях.

ВЫЗОВ 6.3

Аппаратно-программная совместная оптимизация систем компьютерного зрения

Для высокоточной работы систем компьютерного зрения (в измерительной технике, медицинской визуализации) требуется не только развитие алгоритмов, но и совершенствование сенсоров, оптики и оборудования. Без комплексной оптимизации невозможно добиться массовой доступности таких систем. Ко-оптимизация аппаратуры и алгоритмов способна превратить дорогостоящие прототипы в широко применимые решения, ускорить внедрение CV в медицине и промышленности.

Это направление критически важно для выхода ИИ за пределы лабораторий. Создаются междисциплинарные проекты, объединяющие инженеров и специалистов по ИИ. Разрабатываются новые сенсоры и специализированные чипы, адаптированные под конкретные задачи компьютерного зрения.

ВЫЗОВ 6.4

Языки с ограниченными ресурсами, в т. ч. редкие языки

Большинство языков мира плохо представлены в цифровых корпусах, что делает их малоохваченными крупными языковыми моделями. Это создает цифровое неравенство и снижает универсальность NLP-технологий. Поддержка малых (редких) языков критична для глобальной доступности ИИ. Решение проблемы позволит значительно расширить охват технологий и снизить социальное неравенство, открыв возможности для новых образовательных и культурных сервисов. Для выработки ответа на вызов создаются проекты по сбору и разметке текстов для языков с ограниченными ресурсами. Исследуются методы синтетического пополнения корпусов и мультиязычного обучения, включая перенос знаний с более распространенных языков.

ВЫЗОВ 6.5

Использование альтернативных архитектур

Несмотря на доминирование нейросетевых трансформеров, исследователи рассматривают нетрадиционные подходы для узких задач CV и NLP. Изучаются, к примеру, нейроморфные спайковые сети и оптические нейронные сети. Пока эти решения находятся на фундаментальной стадии и недостаточно зрелы для конкуренции с основными методами, но они открывают новые направления развития архитектур ИИ. Развитие альтернативных архитектур способно преодолеть ограничения текущих моделей, повысить энергоэффективность и открыть новые сферы применения.

ВЫЗОВ 6.6

Преодоление разрыва между симуляцией и реальным миром (Sim2Real) при решении узких задач в робототехнике и автономном транспорте

В робототехнике сохраняется разрыв между симуляцией и реальным миром (Sim2Real). Алгоритмы, хорошо работающие в симуляторах, сталкиваются с реальным миром без знаний об истинных физических свойствах объектов. Для преодоления этого барьера требуются методы, которые не только устойчивы к вариативности среды и шумам датчиков, но и обладают восприятием и пониманием физических свойств. Существующие подходы интегрируют физические знания о мире через обучение на датасетах в формате «вопрос — ответ» и имеют ограниченную переносимость этих знаний непосредственно на объекты манипуляции. Проблему усугубляют дефицит качественных реальных данных для обучения роботов и ограничения вычислительных мощностей автономных платформ. Обеспечение надежной и стабильной работы роботов в реальных условиях остается нерешенной задачей мирового уровня. Интегрирование достоверной информации о физических свойствах в обучающие данные является важным этапом для того, чтобы робот мог предсказуемо, безопасно и эффективно следовать инструкциям.

4. Перспективные исследовательские задачи

★ ЗАДАЧА 6.1

Развитие методов компьютерного зрения для симуляции сценариев реального мира (в т. ч. воплощенный ИИ)

Особое внимание уделяется созданию методов компьютерного зрения, устойчивых к сложным сценариям реального мира и способных преодолевать разрыв между симуляцией и реальностью (Sim2Real). Для этого разрабатываются методы доменной рандомизации, имитационного и активного обучения, а также архитектуры Vision-Language-Action (VLA), которые позволяют моделям учиться на собственных действиях в симулированной среде и переносить эти знания в

физический мир. Такие системы будут критичны для робототехники, автономного транспорта и промышленных приложений, где требуются не только высокая точность восприятия, но и гарантированная безопасность поведения.

★ ЗАДАЧА 6.2

Создание фундаментальных моделей для различных задач CV (типа SAM, Dyno v2, CLIP, генеративных VLM) (в т. ч. решение задач классификации, обнаружения, сегментации с открытым списком классов)

Такие модели должны обладать мультимодальными пространственными представлениями, объединяющими информацию из изображений, видео, текстов и сенсорных данных (включая лидары и IMU). Развитие идет в сторону контрастивного и маскированного обучения на мультимодальных данных, объединения 2D и 3D-представлений, а также интеграции механизмов физической правдоподобности в обучение. Ожидается, что такие системы смогут понимать контекст сцены, работать с неизвестными ранее объектами и выполнять пространственные рассуждения, что станет шагом к созданию универсальных систем восприятия для робототехники и воплощенного интеллекта.

★ ЗАДАЧА 6.3

Исследование и разработка эффективных методов обучения и выполнения для архитектур обработки естественного языка (включая AutoML)

Исследования направлены на оптимизацию архитектур с помощью AutoML, использование параметро-эффективных подходов (LoRA, адаптеры), дистилляции знаний и методов ускорения инференса — квантизации, обрезки и смешанных типов точности. Особое внимание уделяется снижению энергопотребления и адаптации моделей к распределенным и специализированным вычислительным средам. Решение этой задачи позволит создавать более доступные, быстрые и обновляемые языковые системы, применимые в широком спектре задач — от интеллектуальных ассистентов до специализированных отраслевых ИИ-приложений.

★ ЗАДАЧА 6.4

Исследование и разработка эффективных методов обучения и выполнения для архитектур рекомендательных систем, распознавания речи, анализа временных рядов (включая автоматическое обучение)

Наконец, важным направлением остается унификация подходов к обучению и исполнению моделей в различных областях — рекомендательных системах (RecSys), распознавании речи (S2T) и анализе временных рядов (TSA). Современные исследования сосредоточены на разработке архитектур, обеспечивающих эффективную работу с потоковыми данными и низкую задержку при высоком качестве предсказаний. Используются стриминговые трансформеры, модели состояний и методы контрастивного самообучения, а также параметро-эффективный тюнинг, позволяющий адаптировать крупные модели под узкие домены и языки программирования без полного переобучения. Результатом станет создание универсальных архитектур и алгоритмов, обеспечивающих адаптивность, энергоэффективность и точность в условиях ограниченных ресурсов, что повысит масштабируемость и доступность современных ИИ-решений в разных секторах экономики.

5. Важные выводы: Экспертное заключение

Научно-техническое сообщество демонстрирует широкий спектр взглядов на приоритеты и пути развития узкого ИИ — от базовых теоретических исследований до практических внедрений. Ниже приведены ключевые позиции и разногласия, озвученные экспертами.

Исследователи выделяют два ключевых пути развития узкого ИИ:

1) унификация, полагающаяся на базовые языковые модели, а также сравнительно небольшое дообучение;

2) специализация, с развитием доменно-специфичных моделей.

Второй путь часто позволяет добиться большей надежности, первый путь более масштабируем и легче в поддержке. В связи с этим логично использовать второй подход в областях с высокой ценой ошибки (например, медицинские диагнозы, кибербезопасность), а первый — для менее критичных областей (чат-боты, общая работа с текстом).

Одной из ключевых проблем, ограничивающей применимость подхода на основе LLM, является недостоверность их размышлений (reasoning unfaithfulness). Несмотря на поверхностную правдоподобность, текстовые размышления часто неполно или искаженно отражают действительную процедуру принятия решения — опускают некоторые факторы, повлиявшие на решение, подгоняют объяснения под правдоподобный, но неверный ответ и вводят в заблуждение относительно ключевых факторов, повлиявших на ответ.

Поэтому перспективны исследования, повышающие достоверность размышлений языковых моделей, — это позволит постепенно расширять область их применения в области с более высокой ценой ошибки, в сочетании с доменно-специфичными и более прозрачными подходами.

Фокус на прикладные исследования. Многие эксперты подчеркивают значимость прикладной работы над ИИ. Особое внимание уделяется комплексной оптимизации «сенсоры — нейросети — процессоры» под конкретные задачи. Такой science-informed ML подход нацелен на уменьшение разрыва между лабораторными прототипами и отраслевыми решениями. Практически это означает, что успех видят в тесной связке фундаментальных исследований с потребностями индустрии, чтобы новые модели сразу учитывали особенности данных и оборудования из реального сектора.

Фундаментальные модели и тренд на эффективные пространственные рассуждения. Современные фундаментальные модели (в частности, VLM), в создании которых достигнут значительный прогресс, всё еще плохо решают задачи, связанные с рассуждением, ответами на вопросы, локализацией объектов в трехмерном пространстве. Разработка таких моделей и архитектур, способных к высококачественным пространственным рассуждениям (Spatial Reasoning), является современным трендом, который может привести к созданию высококачественных систем понимания окружающего пространства для роботов, беспилотных автомобилей, систем виртуальной/дополненной реальности, приложений на смартфонах и т. п.

Эффективность и точность. Отдельно подчеркивается необходимость разработки более эффективных методов обучения и инференса. В частности, задачи компьютерного зрения рассматриваются некоторыми учеными как самостоятельная область для оптимизации: требуется добиться приемлемой скорости и энергоэффективности моделей CV без существенной потери точности. Такой баланс важен для внедрения CV-алгоритмов в устройства с ограниченными ресурсами и для обработки видео в реальном времени.

Многоязыковой барьер. Проблема поддержки языков с ограниченными ресурсами признана серьезной всем научным сообществом. Считается необходимым адаптировать существующие архитектуры LLM и формировать новые наборы данных, чтобы включить как можно больше языков мира в сферу ИИ. Одним из ключевых направлений является разработка более качественных моделей-переводчиков — задача, которая остается нерешенной. Такие модели могут использоваться как для генерации синтетических многоязычных датасетов, так и для улучшения общечеловеческого взаимодействия, расширяя доступность и эффективность NLP-технологий. Без этих усилий технологии NLP рискуют усилить неравенство, оставляя большинство языковых сообществ без современных инструментов.

Выход за пределы нейросетей. Часть исследователей призывает не ограничиваться доминирующей сейчас парадигмой глубоких нейросетей. Звучат призывы к более широкому подходу и изучению альтернативных моделей ИИ (символьных, эволюционных и пр.), чтобы не оставить ценные методы за рамками нейросетевого мейнстрима.

Взгляд на интеграцию узких ИИ. Некоторые эксперты предполагают, что дальнейший прогресс специализированного ИИ может быть связан с прорывами в фундаментальных моделях. *«Я думаю, что при настоящем прорыве в моделях обработки естественного языка можно будет решить все задачи в области 6.3. Однако для этого нам нужен именно прорыв в базовой NLP-модели»,* — считает Шунцюань Тан, подчеркивая, что качественное улучшение базовых языковых моделей может автоматически подтолкнуть развитие смежных узких технологий (распознавание речи, рекомендательные системы и др.). Это мнение иллюстрирует одно из разногласий в сообществе: одни видят будущее узкого ИИ в специальных моделях под каждую задачу, другие — в универсализации базовых моделей с их последующей адаптацией под частные применения.

Наиболее прикладное направление, включающее задачи, важные для конкретных приложений и отраслей

Самое динамически меняющееся направление с точки зрения ландшафта исследований — число самих задач непрерывно растет, и их перспективность сильно меняется в условиях изменчивых внешних обстоятельств

62% **задач связаны с технологиями компьютерного зрения**

Сейчас доминирует парадигма глубоких нейросетей. В рамках направления необходим более широкий подход, связанный с альтернативными моделями ИИ (символьными, эволюционными и пр.)

НАПРАВЛЕНИЕ 7

Управление, принятие решений и агентные/ мультиагентные системы

НАПРАВЛЕНИЕ 7

Управление, принятие решений и агентные/мультиагентные системы

1. Краткое описание направления

Область ИИ для задач управления, принятия решений, агентных и мультиагентных систем претерпела стремительную эволюцию: от программ, основанных на правилах, к сложным автономным системам, самообучаемым и способным к восприятию, рассуждению и действию в сложных средах. Эта трансформация во многом обусловлена прогрессом в глубоком обучении, обучении с подкреплением (Reinforcement Learning, RL) и координации в мультиагентных системах. Суть направления состоит в разработке систем ИИ, способных автономно принимать решения, управлять процессами и взаимодействовать с другими агентами или со средой для достижения конкретных целей, зачастую оптимизируя долгосрочные вознаграждения. Границы области постоянно расширяются: от кооперативной робототехники и роевого интеллекта до игрового ИИ и сложного принятия решений в различных секторах.

Ключевыми поднаправлениями данного направления являются следующие:

1. Разработка алгоритмов обучения с подкреплением

Предполагает, что агент осваивает оптимальные стратегии, максимизируя суммарное вознаграждение в ходе взаимодействия со средой. Это базовая парадигма создания систем ИИ для управления и принятия решений.

2. Агентные системы

Представляют собой целостные автономные ИИ-сущности, способные воспринимать, учиться и адаптивно действовать.

В настоящее время за счет дизайна и использования сложных архитектур агентов удается решать широкий круг прикладных задач.

3. Мультиагентные системы

Реализуют взаимодействие, координацию и кооперацию между несколькими агентами при решении сложных задач, включая методы коммуникации, соревновательности и коллективного принятия решений.

Особая благодарность за работу над данным направлением
Е. Бурнаеву, А. Дарвишу, Д. Юдину

Исторически ИИ для задач управления и принятия решений начинался с классической теории управления и экспертных систем, опиравшихся на предопределенные правила и программирование. Появление машинного обучения, в особенности глубокого обучения и обучения с подкреплением, ознаменовало существенный переход к ориентированным на данные и адаптивным подходам, а также обучению методом проб и ошибок с использованием Марковских процессов принятия решений (Markov Decision Process, MDP) и Марковских игр (Markov Games, MG) для моделирования взаимодействий «агент — среда» и мультиагентных взаимодействий. В последние годы внимание сместилось к масштабируемости в больших пространствах «состояний-действий», обеспечению безопасности в реальных приложениях и развитию обобщенных способностей к обучению. Наблюдается явный тренд к созданию ИИ-агентов, способных к самообучению и работе в разнообразных, открытых и динамических средах. От таких агентов ожидается оркестрирование внешних сервисов и расстановка приоритетов при решении задач. Интеграция в задачи управления больших языковых (Large Language Models, LLM) и фундаментальных моделей (Foundation Models, FM) породила целый ряд новых исследовательских направлений, конечной целью которых является наделение агентов расширенными способностями к рассуждению, планированию и коммуникации.

«В последние несколько лет заметен тренд на создание интеллектуальных агентов для робототехники и автономного транспорта, появление и развитие фундаментальных VLA (Vision–Language–Action) моделей, например, π0, Gemini Robotics, Gr00t постепенно приближают создание роботов общего назначения, которые могут стать универсальными помощниками по дому, помогут автоматизировать рутинные операции на производстве и т. п.», — Дмитрий Юдин, ведущий научный сотрудник Лаборатории когнитивных систем искусственного интеллекта AIRI.

Среди ключевых событий, ставших основой развития данного направления, выделяются следующие:

- прорыв DeepMind's AlphaGo: продемонстрировал мощь RL в сложных стратегических средах и вдохновил дальнейшие исследования в области общего игрового ИИ и систем управления;
- разработка больших языковых и фундаментальных моделей. Стремительное развитие и широкая доступность LLM и FM радикально преобразовали агентные и мультиагентные системы, обеспечив мощные возможности для наделения моделей способностями к рассуждению, планированию и коммуникации на естественном языке;
- прорывы в робототехнике (продолжаются): роботы и беспилотные аппараты осваивают всё более сложные манипуляции, ловкие передвижения и умение работать в неструктурированных средах, что выводит ИИ из симуляций в физический мир и имеет прямые экономические последствия;
- усиление фокуса на мультиагентной кооперации и федеративном обучении: растущая потребность в распределенном интеллекте и конфиденциальном ИИ стимулировала активные исследования в области федеративного мультиагентного обучения с подкреплением (Federated MultiAgent Reinforcement Learning, FMARL) преимущественно в таких секторах, как здравоохранение и финансы;
- разработка генеративных ИИ-агентов: создание фреймворков и технологий для формирования и обучения LLM-агентов ускорило широкое внедрение ИИ в различных отраслях;
- разработка интеллектуальных агентов для научных исследований, способных автоматизировать генерацию гипотез, проведение экспериментов, написание кода и научных статей (AlphaEvolve от Google DeepMind, AI Scientist от Sakana AI). Такие решения могут ускорить научные открытия в областях математики, компьютерных наук, инженерии и т. д.

ИИ для управления, принятия решений и агентных систем является одним из наиболее значимых направлений, поскольку его развитие лежит в основе следующего поколения автономных технологий, трансформирующих практически все секторы промышленности, экономики и жизни людей:

- влияние на промышленность, транспорт, безопасность за счет прогресса в области робототехники и автономных систем: ускорение внедрения автономных транспортных средств, роевых систем БПЛА и передовой робототехники, что ведет к радикальной трансформации транспорта, промышленности и действий в опасных средах;
- влияние на качество жизни за счет трансформации персонализированных сервисов, цифровых экосистем и кибербезопасности: агентные технологии ИИ структурно изменяют цифровые экосистемы, обеспечивая развитие продвинутых игровых ИИ, интеллектуальных интерфейсов, систем предоставления персонализированных услуг — от адаптивных образовательных платформ до клинических ассистентов. В сфере кибербезопасности они могут внести вклад в увеличение проактивности выявления и нейтрализации угроз, способствуя защите критически важных цифровых активов;
- влияние на экономику в качестве драйвера экономического роста за счет повышения производительности, снижения операционных затрат, формирования новых отраслей и сервисов, высокого уровня автоматизации в производстве, логистике и управлении инфраструктурой. Ожидается, что экономические эффекты будут широкомасштабными, затрагивая глобальные рынки и новые конкурентные механизмы;
- влияние на рынок труда: широкое внедрение ИИ-агентов и автономных систем, по прогнозам, трансформирует рынки труда, автоматизируя рутинные операции и формируя спрос на новые компетенции в области разработки, сопровождения и надзора за ИИ. Хотя ряд рабочих мест может быть вытеснен, ожидается появление новых возможностей, что потребует масштабных программ переподготовки и повышения квалификации.

2. Обзор текущего развития направления

Текущее состояние ИИ для задач управления, принятия решений, агентных и мультиагентных систем характеризуется стремительным прогрессом в области глубокого обучения с подкреплением (Deep Reinforcement Learning, DRL) и нарастающей интеграцией LLM. Глубокое обучение с подкреплением демонстрирует высокую эффективность в контексте принятия решений в сложных средах — от игр и управления роботами до управления ресурсами и здравоохранения. В последние 1–2 года фиксируется выраженный переход к разработке унифицированных ИИ-агентов, интегрирующих в единую архитектуру самообучение, планирование и взаимодействие. Приоритетные направления исследований включают повышение устойчивости, безопасности и обобщающей способности у агентов с подкреплением. В частности, происходит развитие обучения с подкреплением с риск-ограничениями (Risk-Constrained RL),

ориентированное на построение стратегий поведения с учетом ограничений и минимизацией отказов, что является особенно критичным для небезопасных приложений, таких как автономное вождение.

Кроме того, набирает значимость концепция «агентного ИИ» (Agent AI) — систем, интегрирующих большие фундаментальные модели в поведенческий контур агента, ведущих к формированию целостного интеллекта. Методологии многозадачного обучения агентов используются для создания адаптивных и универсальных агентов для робототехники, игрового ИИ и здравоохранения. В мультиагентных системах особое внимание уделяется координации, коммуникации и децентрализованному принятию решений, чему способствуют достижения в LLM-ориентированном мультиагентном обучении с подкреплением (MARL) и федеративном обучении. Разработка мультиагентных архитектур, опирающихся на LLM-ориентированные агентные схемы, включая схемы типа Actor-Critic и ролевые модели, является важным трендом и открывает путь к масштабируемым решениям задач, требующих кооперации и конкуренции.

3. Исследовательские вызовы, стимулирующие или ограничивающие развитие направления

ВЫЗОВ 7.1

Масштабируемость и сложность в мультиагентных системах

Одной из ключевых проблем в мультиагентном обучении с подкреплением является работа с большими пространствами состояний и действий, размерности которых экспоненциально растут с увеличением числа агентов и сложности каждого из них. Эта проблема, известная как «проклятие размерности», существенно снижает эффективность существующих подходов по мере усложнения сред. При этом исследования мультиагентных архитектур выходят за рамки инженерной задачи — они касаются фундаментальных вопросов о природе коллективного интеллекта. Феномены фазовых переходов и самоорганизации в мультиагентных системах могут дать ключ к пониманию того, как из взаимодействия относительно простых агентов возникают качественно новые формы интеллектуального поведения.

Решение данной проблемы привело бы к прорывам в области роевого интеллекта, кооперативной робототехники и крупномасштабной оптимизации ресурсов.

В контексте данной проблемы исследователи активно изучают возможности иерархического обучения с подкреплением (Hierarchical Reinforcement Learning, HRL) для декомпозиции сложных задач на поддающиеся решению подзадачи. ИИ-агенты всё шире используют модули генерации, дополненные поиском (Retrieval-Augmented Generation, RAG). LLM также применяются итеративно для улучшения моделей посредством кода и промптов. ИИ-агенты проектируются всё более проактивными и способными к принятию решений, включая оркестрацию внешних сервисов и приоретизацию задач. Разрабатываются модели, такие как MapGPT, для децентрализованного мультиагентного планирования. Существенной тенденцией является самовоспроизведение (self-play): модели ИИ генерируют большие объемы данных, взаимодействуя друг с другом, что ведет к улучшению показателей, превосходящих обучение на данных с участием человека.

ВЫЗОВ 7.2

Безопасность, надежность и доверие при работе ИИ-агентов

Несмотря на эмпирические успехи RL, перенос этих результатов в реальные приложения остается затруднительным из-за сложности обеспечения безопасности, робастности и доверия. Это особенно значимо в критических областях повышенной ответственности, таких как автономное вождение или промышленные системы управления. Первостепенное значение для ответственного развертывания и общественного принятия более автономных систем ИИ имеет установление четких регуляторных рамок, границ ответственности, механизмов надзора и методов аудита решений ИИ.

Общественная реакция на ошибки ИИ асимметрична: автономная система может быть статистически безопаснее человека, но единичный отказ способен подорвать доверие ко всей технологии. Поэтому решение данной проблемы откроет значительные возможности для реального применения ИИ на транспорте, в здравоохранении и критической инфраструктуре.

Значительные исследовательские усилия сосредоточены на обучении с подкреплением с риск-ограничениями и формальной верификации стратегий на безопасность, использовании распределенного обучения с подкреплением для более полного представления о неопределенности и риске, безопасного обучения с подкреплением (Safe-RL) с ограничениями по затратам (cost constraints) для обеспечения безопасности действий агентов за счет двукритериальной оптимизации указанных критериев.

ВЫЗОВ 7.3

Разрыв между симуляцией и реальностью для воплощенных агентов (Embodied Agents)

Перенос обучения между виртуальными (симуляционными) и реальными средами остается существенным препятствием для воплощенных ИИ-агентов, особенно в робототехнике. Стратегии поведения, выученные в высокоточных симуляциях, нередко демонстрируют недостаточную эффективность при развертывании в физическом мире из-за расхождений между моделируемой и реальной физикой, шумом в данных сенсоров и непредсказуемости среды. Этот «разрыв с реальностью» делает обучение роботов непосредственно в реальных условиях сложным и ресурсозатратным.

Данная проблема существенно замедляет разработку и внедрение интеллектуальных роботов и физических автономных систем. Решение этой проблемы требует обширного и нередко рискованного сбора данных в реальной среде и последующей тонкой настройки, что ограничивает масштабируемость исследований в робототехнике и практических приложений. Ее преодоление радикально трансформировало бы робототехнику, обеспечив более быстрые циклы разработки и расширив спектр применений в производстве, исследованиях и сфере услуг. Для ИИ в целом это стимулирует исследования в областях адаптации между доменами, разработки робастных методов обучения непосредственно на основе ограниченного взаимодействия с реальным миром.

Такие методы, как рандомизация домена (domain randomization), при которых параметры симуляции варьируются в широких пределах, применяются для обучения более робастных политик, лучше обобщающихся на реальность. Изучаются мета-обучение и обучение с переносом знания (transfer learning), позволяющие агентам быстро адаптироваться к новым условиям реального мира при минимальном объеме реального опыта. Исследования универсальных моделей физических действий и мультимодальных моделей класса Vision-Language-Action (VLA) нацелены на формирование более обобщенных представлений, менее чувствительных к расхождениям между симуляцией и реальностью.

4. Перспективные исследовательские задачи

★ ЗАДАЧА 7.1

Создание универсальных мультимодальных моделей, объединяющих работу с текстом и другими модальностями: Vision-Language-Action (VLA)

Мультимодальные системы, особенно в робототехнике, станут одним из ведущих направлений на ближайшее десятилетие. Цель состоит в создании по-настоящему универсальных мультимодальных агентных моделей, способных бесшовно объединять восприятие (например, зрение), понимание и генерацию естественного языка и физические действия (Vision-Language-Action, или VLA-модели). Такие модели должны демонстрировать многозадачность, способность к самообучению и эффективную работу в открытых, неструктурированных средах без необходимости специального дообучения под конкретные задачи.

Для решения задачи требуется крупномасштабное предобучение по разнородным модальностям. Среди методов решения необходимо отметить самообучение (self-supervised learning), автокодировщик с маскированием (masked autoencoders) и контрастное обучение (contrastive learning), применяемые к интегрированным потокам визуальных, языковых и поведенческих (action) данных. Обучение с подкреплением будет критически важно для обеспечения самообучения и адаптации в динамических средах, причем иерархическое RL потенциально позволит структурировать сложные сценарии поведения. Кроме того, для данной задачи характерен дефицит данных, в связи с чем потенциальным решением является генерация синтетических данных и использование самовоспроизведения (self-play) для выхода за рамки наборов данных, сформированных человеком.

Решение задачи откроет путь к высокоуниверсальным роботам для бытовых, промышленных и исследовательских задач, а также к интеллектуальным персональным ассистентам, способным к более «человекоподобному» физическому и речевому взаимодействию.

★ ЗАДАЧА 7.2

Создание эффективных методов получения знаний агентами посредством взаимодействия с окружающей средой для достижения целей

Задача сфокусирована на разработке алгоритмов и фреймворков мультиагентных систем обучения с подкреплением, обеспечивающих формальные гарантии безопасности, робастности и соответствия этическим требованиям, особенно в реальных и критически важных приложениях. Речь идет о выходе за рамки чисто эмпирической эффективности, чтобы многоагентные системы могли работать в пределах заданных границ риска, избегали катастрофических отказов и демонстрировали предсказуемое поведение даже в сложных, неопределенных и враждебных сценариях.

Решение задачи позволит получить глубокое понимание того, как формально гарантировать свойства в сложных адаптивных системах. Это в свою очередь

обеспечит массовое внедрение автономных мультимодальных систем в особо чувствительных областях, таких как автономные транспортные системы, управление умными энергосистемами и медицинская робототехника, где отказ недопустим.

★ ЗАДАЧА 7.3

Исследование и создание системы агентов и изучение явления фазовых переходов в мультиагентных системах

Целью является проектирование мультимодальных архитектур, способствующих возникновению развитого коллективного интеллекта и согласованного поведения за счет иерархической организации и децентрализованного обучения, нередко вдохновленных биологическими системами или теорией сложных адаптивных систем. Это включает разработку «мастер-систем», способных динамически создавать и управлять специализированными узконаправленными агентами, а также изучение явлений фазовых переходов по мере роста числа агентов.

Иерархическое обучение с подкреплением (Hierarchical Reinforcement Learning, HRL) позволит агентам обучаться на разных уровнях абстракции и временных масштабах. Генетические алгоритмы и эволюционные методы можно использовать для эволюционного отбора и оптимизации ролей агентов и коммуникационных протоколов в пределах иерархии. Современные коммуникационные протоколы и эмерджентные языки в мультиагентных системах позволяет усилить координацию и обмен информацией. Теория игр и дизайн механизмов могут быть использованы для стимулирования кооперативного поведения и управления конкуренцией.

В результате решения задачи станет возможным создание масштабируемых, гибких и робастных мультиагентных систем, способных адаптироваться к радикальным изменениям среды и решать задачи, существенно выходящие за рамки возможностей отдельных агентов. Это приведет к появлению продвинутых роёв БПЛА для комплексного наблюдения и поисково-спасательных операций, интеллектуальных систем управления трафиком с адаптацией в реальном времени и высокораспределенных робототехнических комплексов для крупномасштабного строительства или мониторинга окружающей среды.

5. Важные выводы: Экспертное заключение

Развитие ИИ для задач управления, принятия решений, агентных и мультиагентных систем характеризуется существенным движением в сторону интегрированного, автономного и адаптивного интеллекта. Стремительная эволюция глубокого обучения с подкреплением (DRL) и интеграция фундаментальных моделей, в частности LLM, выступают ключевыми тенденциями. Это ведет к созданию агентов, обладающих развитым восприятием, рассуждением и способностью работать с различными модальностями в сложных средах.

Акцент сместился от специализированных, основанных на правилах систем, к универсальным обучаемым архитектурам, способным работать в условиях неопределенности и динамики среды.

Возрастающая сложность и масштаб реальных задач обуславливают необходимость мультиагентных решений, стимулируя исследования в направлении устойчивой координации, коммуникации и децентрализованного принятия решений. В целом траектория развития направлена на создание более общих, интеллектуальных и автономных ИИ-сущностей, способных эффективно обучаться, адаптироваться и взаимодействовать, сокращая разрыв между теоретическими достижениями и практическим внедрением в реальном мире.

Кроме того, решение проблем безопасности в многоагентных системах, особенно в отношении сговора (collusion) и скоординированных атак, становится всё более важным, поскольку агенты взаимодействуют, в т. ч. на различных интернет-платформах.

Критическим фактором становится не только технологический прогресс, но и решение вопросов ответственности и регуляторики. Успех направления будет определяться способностью исследовательского сообщества преодолеть разрыв между впечатляющими демонстрациями в симуляциях и надежной работой в реальности.

Сильный акцент на практичные, эффективные и надежные решения для автономных систем

Существенные тренды развития направления связаны с переходом в сторону интегрированного, автономного и адаптивного интеллекта

40% основ направления — исследования, связанные с обучением с подкреплением (reinforcement learning)

Агентные и мультиагентные системы стремительно развиваются, что обусловлено ростом общественного спроса на автоматизацию и автономность. Для целых классов задач агенты могут стать чем-то вроде новой эволюционной формы «традиционного ИИ»

Важный вызов для направления с учетом спроса на автоматизацию и автономность — обеспечение робастности и надежности разрабатываемых технологий и систем

НАПРАВЛЕНИЕ 8

Элементы AGI

НАПРАВЛЕНИЕ 8

Элементы AGI

1. Краткое описание направления

Элементы AGI — это поле исследований, в котором ИИ движется от узких задач к системам, способным рассуждать, учиться на протяжении всей жизни, интегрировать знание в разных формах и действовать в сложных средах. В практическом фокусе — формирование устойчивых свойств будущих систем: достоверные рассуждения и самопроверка, непрерывное обучение во взаимодействии со средой, гибридные нейро-символьные подходы, воплощенность и мультиагентность, нейровдохновленные архитектуры. На основе форсайтов и интервью экспертов выделяется консенсус: определение AGI остается подвижным, а прогресс задается не одним скачком, а конвергенцией перечисленных линий исследований. В ближайшие 3–5 лет наибольшую динамику ожидают в агентных/мультиагентных системах и непрерывном обучении; в более длинном горизонте — в нейро-символике и моделях мозга.

Суть направления — в переходе от «моделей как инструментов» к системам, демонстрирующим адаптивную универсальность: способность надежно рассуждать, планировать, учиться на опыте и переносить знания в новые домены. Эти свойства не сводятся к одному алгоритму: они складываются из взаимодействия reasoning/рефлексии, памяти и персонализированного контекста, обучения в среде, гибридных представлений знаний и агентного действия.

Границы направления проходят там, где заканчивается узкая оптимизация под фиксированный датасет и начинается устойчивое поведение в открытом мире: работа «вне распределения», объяснимость и проверяемость вывода, способность к длительной памяти и корректной самооценке результатов. Практический критерий — способность систем решать сложные прикладные задачи (наука, медицина, инженерный дизайн) при ограничениях ресурсов и данных, сохраняя безопасность и доверие.

В настоящее время выделяются следующие основные поднаправления исследований:

1. Рассуждения и рефлексия

Акцент на повышении качества рассуждений (в т. ч. мультимодальных), причинно-следственном понимании, планировании и механизмах самопроверки — снижение галлюцинаций, калибровка уверенности, проверка цепочек вывода. Это также включает практики объяснимости и сертифицируемой устойчивости вывода.

2. Непрерывное обучение

Переход от статического обучения к непрерывному: активное и учебное планирование (active/curriculum learning), онлайн/офлайн RL, self-play и добыча данных из среды. Цель — устранить насыщение готовых датасетов и добиться адаптации к изменяющимся условиям без катастрофического забывания.

3. Гибридный ИИ

Интеграция глубинного обучения с онтологиями, графами знаний и логикой для интерпретируемости и устойчивости. Нейро-символический подход сочетается с практиками использования инструментов и извлечения знаний (retrieval), что повышает точность и управляемость вывода.

4. Embodiment и мультиагентные системы

Обучение действием в реальных/виртуальных средах, координация нескольких агентов и коммуникация между ними. Базовым слоем выступают foundation-модели (LLM/VLM), обеспечивающие восприятие и планирование; ключевая проблема — перенос из симуляции в реальность (Sim2Real) и безопасность поведения.

5. Моделирование мозга и психики

Нейровдохновленные модели — от импульсных нейросетей до моделирования мозговых цепей — рассматриваются как долгосрочный источник энергоэффективности, устойчивости к шуму и новых принципов памяти/обобщения. Параллельно используется ИИ для интерпретации нейроданных (ЭЭГ и др.), хотя данные скудны и шумны.

Историческое развитие направления движется от символьного ИИ (экспертные системы, онтологии) через статистическое МО к глубокому обучению и трансформерам — такова траектория, на которой возникла надежда на универсальные модели. Однако масштабирование выявило ограничения (галлюцинации, слабая рефлексия, дороговизна), что вернуло интерес к гибридности, непрерывному обучению и

Особая благодарность за работу над данным направлением
И. Оселедцу, Е. Тянь, С. Колюбину

агентности. В играх self-play (например, в го) показали силу обучения «из среды», а ранние нейровдохновленные идеи — потенциал биологических принципов для эффективности и долговременной памяти.

За последние 5 лет наибольшее влияние на развитие направления оказали:

- **ChatGPT и последующие LLM-агенты.** Момент массового признания потенциала foundation-моделей: диалог, инструментальное использование, цепочки рассуждений. Это резко подняло требования к надежности, объяснимости ИИ (XAI) и управлению рисками, сформировав запрос на элементы AGI;
- **RAG/Tool-use как стандарт гибридности.** Широкое внедрение retrieval-подходов и интеграции с внешними инструментами/базами данных закрепило нейро-символическую практику в индустрии. Это улучшило фактическую точность и управляемость вывода, подтвердив значимость гибридного ИИ;
- **Chain-of-Thought и практики рефлексии.** Распространение техник явного рассуждения, самопроверки и калибровки уверенности задало новую норму для задач reasoning. В ответ усилились исследования сертифицируемой устойчивости и методов снижения галлюцинаций;
- **Embodied/мультиагентные системы и миры обучения.** Рост платформ и подходов, где агенты учатся действовать и координироваться в динамических средах, показал практичность перехода от статических датасетов к опыту. Проблема Sim2Real безопасность поведения стали центральными точками повестки;
- **Спайковые и мозговые нейронные сети как долгосрочный вектор.** Возврат интереса к энергоэффективным и биологически мотивированным архитектурам, а также к использованию ИИ для чтения/интерпретации нейроданных. Несмотря на шум и дефицит данных, направление укрепилось как перспективное.

2. Обзор текущего развития направления

В поднаправлении «Рассуждения и рефлексия» акцент смещается к методам повышения достоверности вывода: структурированные цепочки рассуждений, самооценка, верификация фактов, XAI-подходы и практики сертифицируемой устойчивости.

В поднаправлении «Непрерывное обучение» усиливаются онлайн-обучение и self-play, активно обсуждается преодоление насыщения данных за счет добычи опыта и генерации качественных синтетических данных.

В поднаправлении «Гибридный ИИ» практики RAG/Tool-use и интеграции с онтологиями/графами знаний стали де-факто стандартом для задач, где важны точность и прослеживаемость (auditability).

В поднаправлении «Embodiment и мультиагентные системы» наблюдается рост платформ для обучения действием и координации агентов; ключевые барьеры — безопасность и перенос в реальный мир.

В поднаправлении «Моделирование мозга и психики» продолжается поиск энергоэффективных архитектур и методов интерпретации нейроданных; направление развивается, хотя ограничено доступностью данных.

За последние 1–2 года усилилась работа над надежностью reasoning и уменьшением галлюцинаций; RAG и инструментальные агенты (tool-use) закрепились в продуктах; активно растет интерес к мультиагентным системам и обучениям в средах с обратной связью. Параллельно формируется повестка по эффективности: требуется уменьшать стоимость обучения и инференса, улучшать управление памятью и персонализацией.

3. Исследовательские вызовы, стимулирующие или ограничивающие развитие направления

ВЫЗОВ 8.1

Надежность рассуждений и снижение галлюцинаций (Reasoning and Reflection)

Суть: верифицируемые цепочки вывода, калибровка уверенности, устойчивость к дистрибутивным сдвигам. Масштаб: критично для медицины, права, науки; ответ — XAI, процедуры валидации, рефлексивные циклы и сертифицируемые методы. Меры: развитие стандартов оценки и аудита, бенчмарки и практики промышленной валидации.

ВЫЗОВ 8.2

Насыщение данными и непрерывное обучение

Современные ИИ-системы переходят от статической схемы «обучение — развертывание» к модели постоянного накопления данных и опыта. Это позволяет преодолеть насыщение готовых датасетов и избежать деградации знаний при дообучении. Ключевая цель — научить модели извлекать информацию из среды, обогащать собственный опыт и сохранять ранее усвоенные навыки. Такой подход лежит в основе

создания устойчивых и долговременных ИИ-агентов, способных учиться на протяжении всего жизненного цикла. Для этого применяются методы online-обучения с подкреплением (online RL), самоигры (self-play), активного и curriculum-обучения, а также генерация качественных синтетических наборов данных, обеспечивающих постоянное обновление знаний.

⚡ ВЫЗОВ 8.3

Sim2Real и безопасность embodied/мультиагентных систем

Одним из центральных вызовов становится перенос навыков, полученных в симуляторах, в реальный мир. Для автономных и робототехнических систем критически важно гарантировать безопасность, предсказуемость и воспроизводимость поведения агентов вне лабораторных условий. Исследования сосредоточены на разработке методов доменной рандомизации, создании высокореалистичных симуляторов, тестовых полигонов и протоколов сертификации, которые позволяют минимизировать риски некорректных действий систем при взаимодействии с физической средой и людьми.

⚡ ВЫЗОВ 8.4

Дефицит и защищенность нейроданных

Продвижение нейровдохновленных архитектур и технологий brain-reading ограничено недостатками, шумностью и чувствительностью нейроданных. Эти данные трудно получать и анонимизировать, что замедляет прогресс в создании энергоэффективных и когнитивно мотивированных моделей. Решение задачи требует выработки стандартов безопасного сбора и анонимизации, разработки методов сжатия и построения компактных представлений сигналов, что позволит использовать нейроданные без нарушения этических и правовых норм.

⚡ ВЫЗОВ 8.5

Вычислительная эффективность и масштабируемость

Рост размеров моделей и стоимости их обучения делает критическим вопрос энергоэффективности и оптимального использования вычислительных ресурсов. Цель — снизить затраты на обучение и инференс без потери качества, обеспечив масштабируемость технологий и доступность элементов AGI. Это направление включает оптимизацию вычислительных процессов, разработку новых типов аппаратуры и архитектур (включая распределенные и энергоэффективные решения), а также создание методов динамического распределения ресурсов в сложных вычислительных системах.

4. Перспективные исследовательские задачи

Развитие ИИ-систем в сторону большей автономии и сложности порождает новые фундаментальные задачи в области их безопасности, надежности, доверия и объяснимости. Эти задачи требуют глубоких исследований, выходящих за рамки уже известных подходов, и затрагивают как теоретические основы машинного обучения, так и прикладные аспекты взаимодействия ИИ с человеком и окружающей средой. Ниже представлены перспективные исследовательские задачи, выделенные на основе экспертных обсуждений.

★ ЗАДАЧА 8.1

Повышение обобщаемости и адаптивности ИИ-моделей через непрерывное обучение

Основная задача заключается в преодолении фундаментального ограничения современных ИИ-моделей: их неспособности к эффективному обобщению на новые, ранее не виданные домены или задачи без полного и дорогостоящего переобучения. Как отмечает эксперт, *«если мы обучаем нейронную сеть для одной области, не факт, что она будет работать так же хорошо в других областях»*. Необходимо разработать механизмы «эластичности в забывании и запоминании знаний», которые позволили бы моделям постоянно адаптироваться к новым задачам, сохраняя при этом ранее полученный опыт, что известно как «обучение без забывания».

Для решения обозначенной задачи исследуются методы continual learning, новые функции потерь, учитывающие сохранение прежнего опыта, доменная адаптация и когнитивно-инспирированные подходы, основанные на принципах памяти и забывания человека. Реализация таких решений обеспечит создание универсальных, робастных и адаптивных ИИ-моделей, способных учиться на малых данных и непрерывно совершенствоваться в динамичных средах, что особенно важно для робототехники и персонализированных систем.

★ ЗАДАЧА 8.2

Надежность рассуждений: количественная оценка неопределенности и снижение галлюцинаций

Современные большие языковые модели, несмотря на способность отвечать на сложные вопросы, «часто делают очень, очень глупые ошибки» и склонны к галлюцинациям — генерации фактически неверной, но уверенно подаваемой информации.

Задача заключается в разработке методов, которые позволят ИИ-моделям адекватно оценивать, выражать и вербализовать степень своей уверенности или неопределенности в ответах. Цель — научить модели «отклонять ответы» или сигнализировать о неуверенности, что является критически важным для их безопасного применения.

Существуют неконтролируемые (unsupervised) методы использования внутренней статистики модели (энтропии, вероятностей) для оценки уверенности, контролируемые (supervised) методы: обучение вспомогательных моделей для выявления ошибок. Использование подходов, основанных на анализе согласованности ответов (методы «черного ящика»), прямой доступ к внутренним параметрам модели (методы «белого ящика») и обучение явному выражению степени уверенности. Это позволит создать более надежные и самокритичные ИИ-системы, способные безопасно взаимодействовать с человеком, отклонять сомнительные ответы и повышать достоверность рассуждений в критически важных областях.

★ ЗАДАЧА 8.3

Развитие мультиагентных систем и оптимального управления для сложных задач

Эта задача направлена на создание сложных ИИ-систем, состоящих из множества взаимодействующих агентов, способных к координации, сотрудничеству и адаптации в динамичных средах. Задача выходит за рамки классического обучения с подкреплением (RL) и включает в себя разработку методов оптимального управления для решения задач сэмплирования, генерации и тонкой настройки больших моделей. Необходимо создать теоретические и практические основы для управления поведением таких систем, включая обеспечение их предсказуемости и безопасности.

Методы решения включают развитие алгоритмов обучения с подкреплением и оптимального управления для динамичных сред.

Это позволит создавать автономные мультиагентные системы, способные решать сложные задачи, ускорять научные открытия, повышать производительность и обеспечивать более точное управление и настройку больших моделей.

5. Важные выводы: Экспертное заключение

Характер развития направления определяется фундаментальным сдвигом от создания узкоспециализированных моделей к разработке более универсальных, адаптивных и автономных систем. Ключевым трендом является конвергенция нескольких исследовательских линий: надежные рассуждения, непрерывное обучение, гибридные подходы и мультиагентные системы. Наблюдается переход от статического обучения на фиксированных датасетах к непрерывному обучению во взаимодействии со средой, что вызвано проблемой насыщения существующих данных и необходимостью адаптации моделей к динамичным условиям. Всё большее значение приобретают гибридные нейро-символические подходы, такие как RAG, и использование внешних инструментов, которые повышают точность и управляемость выводов ИИ.

Растет понимание, что масштабирование существующих архитектур выявило их фундаментальные ограничения, а именно склонность к галлюцинациям и слабую рефлексию, что сделало количественную оценку неопределенности одной из центральных и очень активных областей исследований. В ответ на это усилились исследования в области самопроверки и калибровки уверенности, что является необходимым шагом к созданию надежных и безопасных AGI-систем. Кроме того, активно развивается направление мультиагентных систем и воплощенного ИИ (embodiment), где фундаментальные модели используются как базовый слой для восприятия и планирования, а основной проблемой становится перенос навыков из симуляции в реальность (Sim2Real).

Определение границ направления затруднено не до конца сформированным определением AGI

19% **перспективных исследований направления связаны с моделированием мозга и психики**

Характер развития направления определяется фундаментальным сдвигом от создания узкоспециализированных моделей к разработке более универсальных, адаптивных и автономных систем

Устойчивое развитие направления обусловлено в т. ч. усиливающимися общественными и регуляторными запросами на этически ответственный и надежный ИИ

Важный вызов для направления с учетом спроса на автоматизацию и автономность — обеспечение робастности и надежности разрабатываемых технологий и систем

НАПРАВЛЕНИЕ 9

Взаимодействие человека и машины

НАПРАВЛЕНИЕ 9

Взаимодействие человека и машины

1. Краткое описание направления

Взаимодействие человека и машины (HMI) — это междисциплинарная область, ориентированная на исследование, проектирование и реализацию интерфейсов для коммуникации и эффективного сотрудничества между человеком и системами ИИ. Его границы выходят за рамки традиционных графических интерфейсов и охватывают все формы двустороннего обмена мультимодальной информацией, включая речь, жесты, взгляд, тактильные ощущения, а также прямое считывание и декодирование сигналов активности нервной системы и стимуляционное воздействие на нервную ткань. Важным аспектом представляется создание средств для интеграции и коллективного взаимодействия человека и ИИ.

В рамках данного направления в настоящее время выделяются три поднаправления:

1. Технические средства прямого взаимодействия с нервной системой человека

Данное поднаправление посвящено системам, взаимодействующим с нервной системой человека посредством непосредственного контакта с ней. Включает в себя разработку систем считывания сигналов мозга, а также создания для этого специализированных аппаратных и низкоуровневых программных средств, поддерживающих двунаправленное взаимодействие с нервной тканью. Особое внимание уделяется созданию биосовместимых средств считывания сигналов активности нейронных популяций с максимальной детализацией и охватом. В сочетании с современными технологиями ИИ разрабатываемые в этом поднаправлении технические решения не только революционизируют медицинскую реабилитацию, но и формируют основу для полноценной интеграции естественного и машинного интеллекта.

2. Технические средства традиционного человеко-машинного взаимодействия

Включают разработку иммерсивных сред (виртуальных, смешанных и реальных) с мультимодальным взаимодействием, а также технических средств доставки обратной связи по естественным сенсорным каналам (зрительному, слуховому, тактильному и обонятельному). Дополнительно в этом поднаправлении исследуются и разрабатываются средства и алгоритмы обработки и интерпретации человеческих действий за счет распознавания жестов, эмоций, речи, а также прогнозирование намерений пользователя и адаптивные интерфейсы.

3. Методы и алгоритмы взаимодействия с человеком

Поднаправление посвящено созданию методологии и алгоритмической составляющей обеспечения совместной работы человека и ИИ в контурах технических средств из первых двух направлений. Подразумевается создание технологий для расширения человеческих возможностей и эффективного сотрудничества в человеко-машинных командах, коллаборативных роботов, поддержки процессов принятия решений. Отдельной важной составляющей являются исследования по расшифровке «кода мозга» и интерпретации сигналов активности головного мозга в инвазивных и неинвазивных двунаправленных интерфейсах «мозг — компьютер». Особый акцент — на машинном обучении и адаптивных системах, способных к самонастройке и персонализации взаимодействия. Важным аспектом выступает создание комплексной системы многокритериальных показателей (точность, задержки, утомляемость), физиологических маркеров и методов проверки конфиденциальности, что позволит объективно оценивать эффективность и безопасность систем взаимодействия и обеспечит поступательное развитие систем человеко-машинного взаимодействия.

Эволюция человеко-машинного взаимодействия прошла путь от интерактивных текстовых интерфейсов (command line interface, CLI), требующих знание команд, к интуитивным графическим интерфейсам (graphical user interface, GUI). Ключевой переход осуществили разработки Xerox PARC, популяризированные Apple Macintosh, которые ввели парадигму WYSIWYG и сделали манипулятор «мышь» основным средством навигации. Следующей революцией стало повсеместное внедрение сенсорных экранов с multi-touch, принцип которых был разработан еще в 1970-х гг., но массовое распространение получил благодаря iPhone, определив современный стандарт Touch User Interface (TUI). Совершенствование технологий

Особая благодарность за работу над данным направлением
А. Осадчему, А. Савченко, И. Лаврову и К. Гугеру

распознавания речи и развитие LLM, выступающих алгоритмической основой современных ассистентов, не только открыли возможность естественного голосового общения через VUI, но и радикально повысили эффективность традиционного CLI-взаимодействия, особенно при программировании и работе с операционными системами.

Недавний прогресс в технологии прямого взаимодействия с мозгом при помощи интерфейсов мозг — компьютер (далее — BCI или ИМК) обеспечил декодирования не только моторных интенций, но и речи из активности мозга, а также возможность прямой модуляции нейрональных процессов, что, с одной стороны, способно вывести обмен информацией между человеком и ИИ на качественно новый уровень, но, с другой — влечет за собой этические риски, связанные с несанкционированным считыванием семантического контента и прямым воздействием на мозг пользователя.

Можно выделить следующие 5 событий, оказавших наибольшее влияние на развитие области человеко-машинного взаимодействия за последние 5 лет:

- **взрывной рост и доступность генеративного ИИ,** в частности больших языковых моделей (2022–2023), кардинально изменили сам принцип взаимодействия с машиной — с текстового ввода и кликов на естественный разговор. Голосовые помощники (Alexa, Siri) получили мощнейший апгрейд, что сделало речевое взаимодействие намного более осмысленным и полезным;
- **создание инвазивных речевых BCI (2024–2025),** оперирующих с почти естественной скоростью, подчеркнуло потенциал технологии прямого взаимодействия с компьютером и возможность декодирования семантического, а не только моторного речевого контекста;
- **выход Neuralink на стадию клинических испытаний на людях (2024),** а именно начало экспериментов с имплантацией чипа в мозг человека стало ключевым событием, привлекшим огромное внимание публики и инвесторов к инвазивным ИМК;
- **развитие метавселенной и расширенной реальности (XR), легитимизация пространственных вычислений и успехи неинвазивных ИМК.** Анонс Meta и метавселенной (2021) запустило гонку технологий за иммерсивное взаимодействие, что стимулировало развитие VR/AR-гарнитур (Meta Quest, Apple Vision Pro) с улучшенными системами отслеживания взгляда, жестов и мимики, а также технологиями тактильной обратной связи (хаптики) и декодирования электромиографической активности (Ctrl-labs) необходимыми для взаимодействия с цифровыми объектами, вписанными в физический мир пользователя;
- **консолидация принципов этики ИИ (2019–2023):** Инициативы разных стран и международных организаций (ОЭСР, ЕС) закрепили прозрачность, справедливость и accountability как основы доверия к ИИ, стимулировав исследования в XAI. Запуск инициатив по регулированию ИИ (ИИ-Акт ЕС, 2021–2024): создание правовых рамок напрямую формирует приоритеты HMI, требуя focus на robustness, прозрачности и human oversight. Особенно важно в контексте пандемии COVID-19 (2020–2022), которая выступила катализатором внедрения инструментов удаленной работы, виртуализации человеческого присутствия.

По мере роста возможностей ИИ и вычислительных систем именно HMI превращается в узкое место. Без кардинального улучшения интерфейсов — повышения их скорости, надежности и естественности — человечество не сможет полноценно использовать потенциал ИИ. В долгосрочной перспективе HMI открывает путь к принципиально новым возможностям — от «вычислительных расширителей» когнитивных способностей человека до медицинских нейропротезов, требующих естественного обмена информацией с мозгом. Однако стремительное развитие этой области порождает и серьезные риски, связанные с нейроприватностью, защитой данных и злоупотреблением технологиями. Поэтому развитие подотчетных и объяснимых систем HMI является не только технологическим, но и стратегическим приоритетом, от которого зависят национальная безопасность и социальная стабильность.

2. Обзор текущего развития направления

Современное развитие человеко-машинного взаимодействия (HMI) характеризуется конвергенцией ИИ, робототехники и нейротехнологий, сопровождающейся фундаментальным пересмотром парадигм взаимодействия. Генеративный ИИ и LLM революционизировали естественно-языковое взаимодействие человека и машины. Наблюдается последовательный переход от интерфейсов, требующих специального обучения, к системам, основанным на моделях реального мира, и далее — к проактивным мультимодальным интерфейсам, способным предвосхищать намерения пользователя.

В области нейротехнологий прогресс особенно заметен: современные речевые ИМК достигли декоди-

рования нейросигналов с почти естественной скоростью, а бионическое протезирование демонстрирует успехи в создании интерфейсов с обратной связью. Преклиническая практика обогатилась первыми имплантами Neuralink, а в клинике уже применяются системы адаптивной глубинной стимуляции мозга для восстановления моторной функции у больных паркинсонизмом. Особое внимание уделяется разработке технических средств считывания активности мозга с высоким пространственным и временным разрешением, продемонстрированы устройства, способные регистрировать активность коры с плотность 44 контакта на квадратный миллиметр, проводятся первые испытания подобных систем на человеке (Precision neuroscience Inc.) и развиваются ИИ-алгоритмы декодирования таких сигналов. Технологии создания контакта с мозгом с использованием наночастиц открывают возможность неинвазивной и высокоточной модуляции активности мозга, что, в свою очередь, и в сочетании с другими трендами поднимает ряд этических вызовов, преодоление которых необходимо для устойчивого развития области HMI.

Особую значимость в этом процессе приобретает Explainable AI (XAI), формирующий алгоритмическую основу доверенного взаимодействия человека и машины. Объяснимость решений ИИ трансформируется из дополнительной функции в базовый принцип проектирования HMI-систем, обеспечивая прозрачность процессов принятия решений, интерактивные возможности для совместного рассуждения ИИ и человека, включая визуальную поддержку процесса логического вывода.

Прорывы в интерфейсах «мозг — компьютер» (BCI):

- речевые BCI: демонстрация декодирования нейросигналов в речь с почти естественной скоростью;
- бионическое протезирование: успешные реализации протезов с очувствлением, как использующих прямое взаимодействие с корой, так и управляемых периферическими электромиографическими сигналами посредством стимуляции периферических нервов;
- первые пациенты с имплантом Neuralink и публичные демонстрации контроля курсора/ввода. Факт первой имплантации и устойчивого управления указателем/набором текста усилил интерес к полностью имплантируемым беспроводным BCI (включая поднятые вопросы безопасности, стабильности сигнала и реабилитационных сценариев и др.);
- отдельная категория OTC-слуховых аппаратов. Создание рынка слуховых помощников резко снизило барьеры доступа и стимулировало инновации в пользовательских аудио-интерфейсах (самоподгонка, интеграция с мобильными устройствами), задавая прецедент для потребительских нейро- и сенсорных устройств;
- первый эндоваскулярный BCI в США от Synchron как новый класс минимально инвазивных интерфейсов, фактически мост между неинвазивным EEG и кортикальными массивами.

3. Исследовательские вызовы, стимулирующие или ограничивающие развитие направления

Стремительное развитие направления HMI формирует ряд вызовов, на решение которых должны быть направлены исследовательские усилия сегодня.

ВЫЗОВ 9.1

Достижение устойчивого и синхронизированного мультимодального слияния

Критически важно сделать это слияние устойчивым (robust) и способным учитывать социальный контекст и намерения пользователя в многопользовательских сценариях HMI. Требуется эффективная интеграция асинхронных и разнородных данных (речь, зрение, жесты) без конфликтов. Актуальность вызова усиливается развитием мультимодальных моделей (Vision Language Models, VLM). Неудача в синхронизированном слиянии приводит к созданию ненадежных систем, степень ненадежности растет экспоненциально с ростом числа потенциально конфликтующих модальностей интерфейсов. Это ограничивает применение ИИ в критически важных и чувствительных сферах. Успешное преодоление данного вызова позволяет создавать «невидимые» интерфейсы и проактивные, предвосхищающие интерфейсы, которые способны понимать пользователя и работать на опережение.

ВЫЗОВ 9.2

Преодоление биосовместимости и долговременной стабильности инвазивных имплантов

Иммунная реакция организма (глиоз) на имплант приводит к деградации качества нейронного сигнала со временем, требующей повторных операций по замене или делающей долгосрочное использование невозможным. Главное технологическое препятствие для коммерциализации и широкого клинического применения хронических инвазивных ИМК делает терапию потенциально опасной и экономически нецелесообразной. Сегодня активно проводятся исследования новых материалов (гибкая электроника,

гидрогели), биоинертных покрытий, миниатюризации электродов для снижения иммунного ответа. Развиваются синаптические нейроинтерфейсы, в которых контакт с нервной тканью устанавливается за счет формирования естественного синапса на электродной площадке.

⚡ ВЫЗОВ 9.3

Декодирование семантического намерения против моторных команд

Большинство успешных ИМК декодируют моторные команды (преднамерение пошевелить рукой). Гораздо сложнее декодировать абстрактные мысли, внутреннюю речь (интенцию) или эмоциональное состояние напрямую, без опоры на моторные корреляты. Это ограничивает применение ИМК для восстановления высших когнитивных функций (например, общения у полностью парализованных пациентов) и создание интерфейсов следующего поколения. Сегодня для решения данной проблемы используются LLM для контекстуальной интерпретации нейросигналов, исследования в области декодирования аудиторного и визуального воображения (например, речевые ИМК).

⚡ ВЫЗОВ 9.4

Исследование принципов кодирования для формирования воздействия на мозг

Вызов заключается в расшифровке фундаментального «языка» мозга — нейронного кода, — который преобразует информацию в паттерны электрической и химической активности. Решение этой задачи позволит перейти от простого наблюдения за мозговой деятельностью к целенаправленному программированию нейронных ансамблей, открывая путь к созданию принципиально новых систем: от медицинских нейропротезов, восстанавливающих утраченные функции (зрение, слух, движение), до прямых интерфейсов «мозг — компьютер» для управления сложными системами силой мысли.

⚡ ВЫЗОВ 9.5

Преодоление административно-правовых барьеров для клинических испытаний

Жесткие регуляторные требования и этические комитеты затрудняют проведение исследований с инвазивными ИМК на людях. Дополнительным барьером является отсутствие устоявшейся практики и инфраструктуры для таких рискованных экспериментов с пациентами, которые в них больше всего нуждаются. Сложившаяся ситуация замедляет темпы исследований и переноса технологий из лаборатории в клинику. Пациенты лишаются доступа к потенциально прорывной терапии. В мире практикуется создание «хабовых» клиник при ведущих университетах (например, Mass General Brigham при Гарварде), разработка ускоренных регуляторных процедур для медицинских устройств категории breakthrough therapy, привлечение пациентских сообществ в дизайн исследований.

4. Перспективные исследовательские задачи

★ ЗАДАЧА 9.1

Создание интуитивных агентов, понимающих запросы и ожидания пользователя, его эмоциональное состояние и т. п.

В отличие от классических интерфейсов, взаимодействие с такими системами строится на принципе целеполагания — когда человек задает не действие, а желаемый результат, а агент сам выбирает оптимальный способ его достижения. Для этого применяются мультимодальные модели восприятия речи, жестов, взгляда и физиологических реакций, а также архитектуры типа Vision-Language-Action и memory-augmented-агенты, способные учитывать контекст, социальные роли и эмоциональное состояние пользователя. Такие системы позволяют существенно сократить когнитивную нагрузку и время на взаимодействие, что особенно важно при управлении роботами, автономными системами и комплексными цифровыми средами. Перспективно проектирование систем, поддерживающих сетевое взаимодействие смешанных команд (люди + ИИ-агенты) с динамическим распределением инициативы, разработка моделей для учета социального контекста и алгоритмов для эффективного коллективного мышления.

★ ЗАДАЧА 9.2

Исследование и разработка двунаправленных интерфейсов «мозг — компьютер», расшифровка «кода мозга» и создание фундаментальных моделей данных функционального картирования мозга

Создание замкнутого контура «намерение — действие машины — обратная связь» с использованием мультимодальных методов регистрации активности мозга, адаптивного декодирования на основе foundation-моделей и алгоритмов обучения с подкреплением для подстройки в реальном времени. Ключевые требования — обеспечение минимальных задержек (<100 мс), высокой точности декодирования и повышение скорости информационного обмена в неинвазивных системах.

★ ЗАДАЧА 9.3

Установление метрологической базы для оценки HMI

Сегодня отсутствуют единые стандарты измерения таких параметров, как доверие пользователя, когнитивная нагрузка, вовлеченность или уровень автономии агента. Формирование набора метрик и протоколов испытаний позволит объективно оценивать эффективность и безопасность систем взаимодействия. В качестве инструментов предполагается использование VR/AR-песочниц, многокритериальных метрик (точность, задержка, утомляемость), физиологических показателей и методов тестирования приватности. Разработка открытого бенчмарка HMI и сертификационных критериев создаст основу для сопоставимости решений, ускорит вывод инновационных продуктов на рынок и повысит доверие к технологиям взаимодействия человека и ИИ.

5. Важные выводы: Экспертное заключение

Взаимодействие человека и машины становится стратегическим узким местом цифровой трансформации: качество, скорость передачи информации, естественность таких интерфейсов и степень доверия к ним определяют, в какой мере потенциал ИИ будет реализован. Это относится как к системам, обеспечивающим осознанное взаимодействие человека и машины, так и к нейроинтерфейсам, реализующим возможность прямого контакта с мозгом и нервной тканью и служащим для замещения утраченных функций у пользователей-пациентов или их аугментации у здоровых пользователей.

Траектория развития традиционных средств взаимодействия человека и машины смещается к мультимодальности и нейрокогнитивным системам: комбинирование текста, речи, зрения и жеста фактически стало стандартом. Происходит развитие «невидимых» интерфейсов и умных сред (Ambient/IoT), обеспечивающих контекстность и ненавязчивость взаимодействия. Важным трендом является переход от модели «один пользователь — одна машина» к многопользовательским сценариям, партнерству «человек — ИИ» и коммуникации «машина — машина» с модерацией человеком.

В нейроинтерфейсах зафиксированы точечные прорывы: инвазивное декодирование речи со скоростями, близкими к естественным; успешные попытки семантического декодирования, протезирование конечностей с сенсорной обратной связью через периферическую стимуляцию при управлении по ЭМГ; применение ИМК-опосредованной стимуляции спинного мозга; коммерческая доступность состояние-зависимых RNS-систем для терапии эпилепсии, ригидности и тремора; демонстрация информативности сверхвысокоплотной ЭЭГ (порядка 4К каналов).

Тем не менее основным источником информации об активности нервной ткани, обеспечивающим необходимое количество информации, пока являются инвазивные системы электродов, в идеале обеспечивающие доступ к активности большого числа (более 1000) отдельных нейронов. Для устойчивого развития этого направления необходимо создание таких имплантируемых систем электродов, обладающих долговременной биосовместимостью и регистрирующих активность десятков тысяч нейронов в пространственно распределенных участках коры головного мозга. Интересным направлением представляется формирование контакта с нервной тканью за счет синаптического интерфейса.

Принципиальным является использование интерфейса с нервной тканью не только для считывания информации, но и для стимуляции, что необходимо для замыкания петли «обратной связи» (тактильной/сенсорной) в целях обеспечения ощущения агентности, повышения функциональности протезов и снижения когнитивной нагрузки при их использовании. Для решения этой задачи критически важным является использование сложных контекст-зависимых нейросетевых алгоритмов кодирования, распределенных в пространстве и времени паттернов стимуляции, обеспечивающих максимально естественную обратную связь. Оптогенетические и термогенетические технологии также находят применение для высокоточной и пространственно-избирательной стимуляции нервной ткани. Кроме того, наночастицы в настоящее время рассматриваются как альтернативное малоинвазивное средство формирования двунаправленного контакта с нервной тканью. Важнейшим организационным аспектом, определяющим конкурентоспособность и возможность развития нейроинтерфейсных технологий, является создание отечественной и международной правовой базы, легитимизирующей эксперименты по замещению утраченных функций у реальных пациентов с использованием прототипов нейроинтерфейсных систем.

Путь развития технологий человеко-машинного взаимодействия лежит в направлении создания целостных, доверительных и адаптивных экосистем. Ключевой чертой этих экосистем станет их способность к синергии с человеком: от динамических интерфейсов, предвосхищающих намерения, до нейрокогнитивных технологий, обеспечивающих глубокую интеграцию на биологическом уровне. Дальнейший прогресс будет определяться не только техноло-

гическими прорывами, но и успешным решением комплексных задач на стыке регуляторики, этики и обеспечения кибербезопасности. Формирование такой среды, где технологии не просто предоставляют инструменты, но партнерски усиливают человеческий потенциал, является центральной задачей на пути к реализации полного спектра возможностей цифровой трансформации.

Стратегическое узкое место цифровой трансформации: именно качество, скорость и естественность интерфейсов определяют, в какой мере потенциал ИИ будет реализован

Важнейший вызов направления — разработка моделей, понимающих социальный контекст, роли и иерархии

Прорывы в области Generative AI и LLM существенно изменили парадигму взаимодействия машины и человека и исследовательский ландшафт всего направления

Важный вызов для направления с учетом спроса на автоматизацию и автономность — обеспечение робастности и надежности разрабатываемых технологий и систем

80% **задач так или иначе связаны с расширением человеческих возможностей за счет разного рода коллабораций с ИИ**

НАПРАВЛЕНИЕ 10

Общество в эпоху ИИ

НАПРАВЛЕНИЕ 10

Общество в эпоху ИИ

1. Краткое описание направления

Наряду с преимуществами, порождаемыми технологиями ИИ, наблюдается их комплексное трансформирующее воздействие на социальные институты. Данное влияние носит сквозной характер, проявляясь в различных сферах человеческой деятельности. В экономической плоскости это выражается в структурной трансформации традиционных секторов промышленности и формировании принципиально новых, т. н. отраслей будущего. В социальной плоскости происходит глубинная перестройка уклада жизни, трансформация моделей занятости и парадигм образования. Культурная сфера сталкивается со сменой режимов генерации и дистрибуции контента, что ведет к пересмотру самих основ культурного производства. В научно-технической области ИИ порождает новые исследовательские парадигмы, знаменуя наступление новой эпохи в организации научного познания.

«Влияние ИИ в ближайшие годы приведет к новой эволюции, что подчеркивает необходимость внедрять алгоритмы ИИ для улучшения качества жизни и развиваться параллельно с ИИ», — отметил профессор Азидин Гезаз, доцент кафедры компьютерных наук и математики Высшей технологической школы Эс-Сувейры, Университет Кади Айиад (Марокко).

Экспертное сообщество в ходе участия в коллективных обсуждениях неоднократно подчеркивало, что технологические прорывы сопровождаются серьезными социальными и этическими рисками, а темпы внедрения ИИ часто опережают разработку благоприятного нормативно-правового регулирования технологии как на международном, так и на национальном уровнях. Исследователи отметили: *«Мы строим технологии, не дождавшись "глобального общественного договора" об их допустимости».*

Данное направление концептуализирует взаимодействие ИИ и общества как трехстороннюю взаимосвязь, охватывающую сферы управления, этики и социально-экономической трансформации, и подчеркивает, что эти области взаимно усиливают друг друга и должны развиваться в единстве для обеспечения ориентированного на человека развития. Таким образом, в рамках данного направления по итогам международных форсайт-сессий выделяются следующие поднаправления, ставшие его основой:

10.1 Механизмы глобального управления ИИ, включая регулирование ИИ:

- **Национальные и глобальные системы управления ИИ.** Последние годы большинство юрисдикций озабочено выстраиванием системы управления ИИ как на национальном, так и на глобальном уровнях. Неоднократно отмечается и реализуется на практике тот факт, что управление ИИ должно выходить за национальные границы, формируя глобальную рамку на основе многостороннего сотрудничества. Несмотря на существенную вариативность национальных подходов к регулированию ИИ, дальнейшее развитие системы управления ИИ диктует необходимость выработки более многогранной и комплексной парадигмы, выходящей за узкие пределы национальных регуляторных моделей. При этом многие исследователи указывают на феномен систематического отставания правового поля от скорости технологического прогресса, что порождает регуляторный вакуум и, как следствие, усугубляет общественные опасения и социальную настороженность. Тем не менее также примечательно развитие международного сотрудничества в области создания регуляторных песочниц для ИИ на национальном уровне — от Бразилии до Мозамбика и Индонезии, а также по всей территории Европейского континента, что помогает правительствам и предпринимательским экосистемам найти оптимальный баланс между регулированием ИИ и инновациями (Европейский союз в ближайшее время планирует выпустить соответствующие руководящие принципы по их внедрению).
- **Международная кооперация.** За последний год наблюдается тенденция к международному сотрудничеству в сфере ИИ. Исследователи отмечают, что критически важно выработать широкий международный диалог и консенсус по вопросам управления и развития технологий ИИ с целью продвижения безопасных, ответственных и заслуживающих доверия технологий ИИ. В этом направ-

Особая благодарность за работу над данным направлением
С. Белуали, А. Незнамову, Л. Чжэн, Ж. Пита Коста

лении работают все: начиная от международных организаций, например, Генеральная ассамблея ООН, формирующая на своей площадке «Глобальный диалог по управлению ИИ» и научную панель, которые ежегодно будут собирать страны для выработки правил разработки безопасных и подотчетных систем ИИ, заканчивая призывами на национальном уровне. В их числе стоит отметить инициативу Китая по созданию Всемирной организации сотрудничества по ИИ, в рамках которой каждая страна могла бы участвовать на равных и совместно формировать международные правила развития технологий ИИ.

10.2 Этика ИИ

Этические вопросы занимают центральное место в дискуссиях об обществе в эпоху ИИ. Устойчивое развитие технологий невозможно без соблюдения таких принципов, как недискриминация, прозрачность и объяснимость алгоритмов, защита данных, безопасность и надежность, подотчетность и контроль. В экспертном сообществе закрепилось понимание необходимости «этики по дизайну», когда перечисленные ориентиры закладываются в систему еще на стадии ее разработки. Это особенно важно на фоне растущего внимания к рискам предвзятости и несправедливости, которые могут воспроизводиться и усиливаться через технологии ИИ. Этика ИИ не должна ограничиваться процедурными концепциями справедливости и прозрачности, а должна включать в себя социально-технические аспекты этики, охватывая такие вопросы, как алгоритмическая справедливость, эпистемическое разнообразие и неравноправные властные структуры, заложенные в процессах сбора данных и проектирования систем.

10.3 Изучение эффектов влияния технологий ИИ на общество

- **Экономика и рынок труда.** В экономической сфере технологий ИИ представляют собой мощный драйвер, способный повысить производительность и изменить рынок труда. Однако вместе с этим возникают риски увеличения социального неравенства, вытеснения профессий и концентрации капитала в руках крупных корпораций. *«ИИ ускоряет производство, но необязательно справедливо делит выгоду»*, — отмечает научное сообщество. В международной дискуссии поднимается вопрос о справедливом распределении прибыли, и выдвигаются идеи компенсаторных мер, например, налоги на автоматизацию, перераспределение выгоды, а также даже концепция универсального базового дохода как одного из возможных инструментов смягчения социальных последствий. Анализ экономической трансформации, обусловленной ИИ, должен осуществляться через призму политической экономии технологий, учитывая то, как автоматизация перераспределяет стоимость, права на данные и переговорную силу, а также порождает новые дискуссии вокруг дивидендов от ИИ, концепции данных как труда и справедливых механизмов распределения создаваемой стоимости.
- **Культурная идентичность.** В культурном аспекте развития технологий ИИ следует уделить особое внимание принципам культурной идентичности технологий ИИ: язык и стиль общения, обучение ИИ на исторических и региональных выборках, противодействие культурной стандартизации, модели вежливости и «табу», политические и общественные контексты. Эксперты подчеркивают необходимость сохранения культурного разнообразия, развития цифровой грамотности и критического восприятия результатов, создаваемых технологиями ИИ. Без этого общество рискует столкнуться с унификацией культурных кодов и усилением феномена «пост-правды». Данный аспект может быть переосмыслен в контексте цифрового суверенитета и плюрализма, что подчеркивает важность лингвистического и эпистемического разнообразия в разработке ИИ для предотвращения алгоритмической унификации и сохранения культурного наследия в цифровую эпоху.
- **Изучение границ приемлемости уровня автономности ИИ для людей в различных сферах.** Исследование пределов приемлемости автономности ИИ для людей как пользователей систем ИИ в различных социальных контекстах является важным аспектом, поскольку именно этот параметр определяет готовность общества делегировать системам принятие решений в таких сферах, как здравоохранение, транспорт и финансовые услуги и др. Данное исследование должно быть сфокусировано на комплексном анализе взаимодействия между техническими возможностями алгоритмов, психологическими факторами доверия и прозрачности, а также контекстуальными переменными уровня риска и культурных особенностей.

Профессор Суянто, ректор Университета Телком (Индонезия) отметил: *«Исследования должны быть направлены на сохранение суверенитета данных, человека и культуры в эпоху “пост-правды”».*

История взаимодействия общества и технологий показывает, что каждое крупное внедрение автоматизации сопровождалось не только ростом производительности, но и серьезными социальными вызовами.

Индустриализация XIX в. породила движение луддитов, роботизация второй половины XX в. дискуссии о занятости, а цифровая революция начала XXI в. открыла вопросы приватности и контроля над данными. Данные примеры показывают, что технологический прогресс неизменно требует от общества адаптации — социальных, правовых и культурных механизмов, способных смягчить противоречия.

Современный этап развития технологий ИИ повторяет эту закономерность. Первые решения внедрялись локально и точечно, постепенно трансформируя целые отрасли. Массовое распространение систем машинного обучения и нейросетей выявило не только потенциал ускорения процессов и роста эффективности, но и новые риски: от дискриминации в алгоритмах до угроз приватности и прозрачности, от вытеснения профессий до изменения структуры рынка труда, от стандартизации культурных кодов до формирования новой идентичности. В ответ на эти вызовы начали появляться этические кодексы, нормы «этики по дизайну», стандарты оценки воздействия ИИ и нормативно-правовые рамки, однако все они формировались постфактум, что подтверждает устойчивую тенденцию отставания регулирования от прогресса технологий.

В ходе научных дискуссий экспертами была высказана позиция, что будущее должно строиться не через замену человека, а через усиление его возможностей.

Таким образом, история внедрения автоматизации и технологий ИИ демонстрирует повторяющийся сценарий: технологический скачок открывает новые горизонты, но одновременно ставит вопросы о социальной справедливости, регулировании и культурной адаптации и необходимости международной кооперации.

Среди ключевых событий социального, экономического характера, оказавших наибольшие влияние на формирование и развитие направления, можно выделить следующие:

- **Широкомасштабная коммерциализация и интеграция генеративных моделей ИИ** в публично доступные сервисы и потребительские продукты, наблюдаемая в последние годы, формирует принципиально новую технологическую парадигму. Несмотря на потенциал для оптимизации пользовательского опыта и расширения функциональных возможностей, данное событие актуализирует комплекс серьезных вызовов, выходящих за рамки чисто технических задач. К их числу относятся этические дилеммы, связанные с авторством контента, фундаментальная проблема галлюцинаций ИИ — порождения моделями правдоподобной, но фактически недостоверной информации, а также риски принятия необъективных решений.
- **Риски, порождаемые технологиями ИИ**, требуют выработки подходов к регулированию технологий ИИ как на национальном уровне, так и на международном уровнях:
  - **Принятие этических документов по всему миру**. За последние годы произошел сдвиг от частных инициатив отдельных компаний к разработке этических рамок на уровне отраслевых ассоциаций, национальных стратегий и международных организаций. Если в начале 2020-х гг. речь шла о корпоративных принципах «ответственного ИИ», то к 2021–2023 гг. появились документы, принятые государствами и межгосударственными структурами. Среди них выделяются Рекомендация ЮНЕСКО по этике ИИ (2021), Кодекс этики в сфере ИИ в России (2021), принципы OECD и G20 и др. В последние 5 лет наметился переход к закреплению этических принципов в национальных стратегиях и к формированию глобальных рамок, что отражает рост зрелости дискуссии и попытку превратить ценностные ориентиры в политические и правовые механизмы.
  - **Активизация законодательных инициатив.** С 2020 по 2025 гг. наблюдается резкий рост правовых инициатив, направленных на регулирование использования ИИ. Согласно Stanford AI Index 2025, число упоминаний ИИ в законах увеличилось на 21,3 % по сравнению с 2023 г. Мировое сообщество поэтапно приходит к выработке моделей правового регулирования технологий ИИ. На сегодняшний день явно можно выделить ограничительную, гибридную, проинновационную модели регулирования ИИ.
- **Цифровой разрыв между развитыми и развивающимся странами.** Отсутствие равного доступа к технологиям ИИ снижает конкуренцию как на национальном, так и на международном уровне. Преобладание крупных игроков на рынке ИИ оказывает негативное влияние на общие показатели развития мира в области ИИ. В то время как передовые экономики продолжают внедрять ИИ в промышленность, государственное управление и систему образования, повышая производительность и эффективность, развивающиеся страны рискуют столкнуться с еще большим отставанием. Возникающий в результате дисбаланс усугубляет глобальное неравенство в уровнях доходов, качестве образования и инновационном потенциале,

порождая ощущение цифровой колонизации, при котором развивающиеся страны не участвуют в разработке технологий ИИ, а выступают лишь в роли поставщиков данных.

- **Экономика и рынок труда.** Обострение дискуссий о влиянии ИИ на рынок труда стало одним из главных событий последних лет. Массовое распространение генеративных моделей в 2022–2023 гг. ускорило автоматизацию интеллектуальных задач, вызвав волну дебатов о будущем профессий. В ходе научных дискуссий участники форсайта неоднократно отмечали, что сотни миллионов рабочих мест подвергнутся трансформации или исчезнут, а параллельно появятся новые виды занятости. Значимость данного направления определяется тем, что именно оно задает рамки общественного доверия к ИИ и определяет баланс между инновациями и рисками. От способности государства и социальных институтов обеспечить прозрачность, подотчетность и справедливое распределение выгод зависит социальная стабильность, легитимность институтов и устойчивость рынков труда. Пренебрежение этими вопросами может привести к росту дискриминации, нарушению конфиденциальности и усилению социального неравенства, тогда как системное внедрение этики «по дизайну» и продуманное регулирование минимизируют издержки и сформируют основу для долгосрочного доверия. В экономике корректные правовые рамки и социально ответственные стандарты позволяют снизить транзакционные издержки, сделать внедрение предсказуемым и уменьшить риск чрезмерной концентрации рыночной власти. В культурной и образовательной сферах сохранение идентичности и разнообразия при распространении ИИ повышает общественную приемлемость технологий, укрепляет ценностные ориентиры и поддерживает творческое развитие. В совокупности это направление выступает гарантом того, что переход к эпохе ИИ будет происходить не только в логике эффективности, но и в логике устойчивости, справедливости и человеческого достоинства.

2. Обзор текущего развития направления

Многими экспертами была выражена обеспокоенность о том, что регулирование технологий ИИ отстает от темпов его внедрения и не успевает отвечать на негативные последствия для общества. За последние годы наблюдается рост количества регуляторных инициатив в области ИИ, однако все они так или иначе основываются на разных фундаментальных подходах. Так, Закон Европейского союза об ИИ (EU AI Act) демонстрирует консервативный подход к регулированию технологий ИИ, вводя риск-ориентированный подход и обременительные законодательные требования для разработчиков моделей ИИ, разверстывающих их на территории ЕС. Другая категория стран, включая Китай и Россию, ориентированы на более гибкий подход, в рамках которого регулирование комбинируется из актов стимулирования с точечными нормативными ограничениями и саморегулированием. Некоторые страны, в числе которых можно отметить Великобританию, Сингапур, Южную Корею, Японию, формируют проинновационную модель регулирования технологий ИИ, которая вводит минимальные законодательные ограничения, делает упор на поддержку научных исследований и инвестиций в ИИ.

Лю Шу, исполнительный генеральный секретарь Шэньчжэньской ассоциации индустрии ИИ, в ходе одной из научных дискуссий, высказывая свое экспертное мнение по вопросу регулирования технологий ИИ, сравнила подходы Европейского союза и Китая: «Ключ к созданию эффективного регулирования в его применении. Европейский союз, который уделяет особое внимание безопасности, этике и соответствию технологий, установил довольно много ограничений, что, однако, привело к относительно медленному прогрессу в применении ИИ. В то же время Китай, обеспечивая быстрое развитие технологий, посредством соответствующих законодательных актов защищает ключевые направления их использования, что позволяет быстрее трансформировать технологии в практические приложения и способствует широкомасштабному внедрению ИИ».

Несмотря на разные подходы стран на национальном уровне, система управления ИИ требует выработки более многогранного и комплексного подхода к ИИ, выходя за регуляторные рамки на национальном уровне.

Вместе с тем наблюдается активное развитие и внедрение инструментов оценки этичности моделей ИИ как на уровне международных организаций, национального законодательства стран, так и на уровне отдельных ведущих корпораций с целью выявления и устранения предубеждений и ошибок в алгоритмах ИИ, усовершенствования механизмов защиты данных, формирования человекоцентричных моделей ИИ, а также повышения доверия со стороны общества.

По итогам международного форсайт-исследования был подтвержден тезис о том, что на практике применяются разные подходы к мониторингу и оценке

систем ИИ, что порождает разногласия в вопросах контроля и ответственности за действия систем ИИ.

Экономическая трансформация, инициированная распространением технологий ИИ, характеризуется не только ускорением роста производительности, но и заметным перераспределением финансовых потоков в пользу крупных корпораций, что усугубляет проблему социального неравенства. В ответ на эти вызовы на глобальном уровне активизировались дискуссии о поиске механизмов справедливого распределения выгод, включая введение налога на автоматизацию и рассмотрение моделей безусловного базового дохода.

Параллельно в культурно-социальной сфере наблюдается переосмысление традиционных парадигм под влиянием ИИ: трансформируются понятия творчества, образовательных процессов и даже личной идентичности. Данные изменения актуализируют вопросы этической ответственности создателей алгоритмов, а также острую необходимость в развитии у граждан цифровой грамотности, включая навыки критической оценки результатов работы ИИ. Указанные комплексные проблемы требуют для своего решения междисциплинарного подхода, объединяющего компетенции как технических специалистов, так и экспертов в области гуманитарных наук.

3. Исследовательские вызовы, стимулирующие или ограничивающие развитие направления

В числе исследовательских вызовов, стимулирующих или ограничивающих развитие направления, особо стоит выделить следующие:

ВЫЗОВ 10.1

Выработка системы управления технологиями ИИ и преодоление цифрового разрыва в области ИИ

Становление глобальной системы управления искусственным интеллектом находится на начальной стадии, что проявляется в многообразии инициатив, законодательных предложений и концепций, выдвигаемых на национальном уровне, международными организациями и корпорациями. Ключевой задачей является консолидация этих усилий для создания равноправной, устойчивой и заслуживающей доверия модели, основанной на международном сотрудничестве. Эксперты подчеркивают, что управление ИИ должно быть ориентировано на равный доступ к технологиями и способствовать сокращению глобального цифрового разрыва. Технологии ИИ не должны усугублять существующее неравенство, а, напротив, стать катализатором развития для всех стран, включая уязвимые в технологическом отношении регионы. Следовательно, развитие ИИ требует синхронизации с целями устойчивого развития и базовыми ценностями человеческого общества.

ВЫЗОВ 10.2

Международная кооперация

Мировое сообщество поэтапно движется к выработке признанных на международном уровне стандартов в области ИИ. Ключевой задачей на этом пути является организация широкого межгосударственного диалога для достижения глобального консенсуса по вопросам управления ИИ. Особую значимость приобретает переход от декларативных заявлений к практическим действиям, направленным на обеспечение равных возможностей для всех стран.

ВЫЗОВ 10.3

Разработка унифицированных стандартов измерения этических характеристик моделей ИИ

Существующие метрики, составляющие основу методологии оценок этичности моделей ИИ, часто не охватывают в полной мере все аспекты соответствия моделей ИИ принципам этичности, начиная от недискриминации алгоритмов ИИ и избегания случайных и нежелательных корреляций и заканчивая подотчетностью и контролем за деятельностью моделей ИИ. В этой связи экспертным сообществом подчеркивается важность создания комплексных оценочных фреймворков, позволяющих сравнивать модели ИИ между собой, формируя тем самым общепризнанный стандарт оценки, вносящий ясность в набор метрик, необходимых для соответствия принципам этичности ИИ.

Следует добавить необходимость разработки и институционализации методик оценки этических последствий и воздействия на права человека (ОВПЧ), чтобы превратить принцип «этики по умолчанию» в конкретную и проверяемую практику. Данные инструменты обеспечивают эффективное внедрение ценностей на всех этапах жизненного цикла систем

ВЫЗОВ 10.4

Потеря рабочих мест и трансформация занятости

Автоматизация всё большего числа функций, включая интеллектуальные и креативные задачи, вызывает

серьезные опасения относительно будущего рынка труда. Многие профессии находятся под угрозой исчезновения, одновременно возникают новые формы занятости, требующие переквалификации. Научный вызов состоит в прогнозировании этих процессов, разработке моделей адаптации и формировании образовательных систем, готовых к новым условиям.

⚡ ВЫЗОВ 10.5

Культурный плюрализм и идентичность

Культурно-этические особенности стран и регионов существенно различаются и могут вступать в противоречие при создании универсальных международных стандартов. Это проявляется в языке, стилях общения, нормах вежливости, исторических и политических контекстах. Вызов заключается в разработке подходов, которые позволят сохранить культурное многообразие при формировании минимально необходимых универсальных стандартов для глобального использования ИИ. Повышение значимости межкультурного взаимодействия до уровня безопасности человека: крупные языковые модели (LLM) транслируют специфические культурные репрезентации, способные угрожать культурной идентичности. Исследования должны определить, как фильтрация, аннотирование и проектирование наборов данных могут гарантировать, что ИИ будет уважать социальные связи и коллективное достоинство, а не только индивидуальное.

⚡ ВЫЗОВ 10.6

Воздействие сложных технологий, включая ИИ-агентов, на когнитивные функции человека

Новые поколения ИИ, работающие как автономные агенты, усиливают риски от потери контроля над действиями систем до роста непрозрачности их решений. Они меняют формы коммуникации, делегирования полномочий и структуру социальных институтов. Вызов заключается в изучении долгосрочных последствий внедрения ИИ-агентов и выработке рамок, которые позволят использовать их безопасно и с пользой для общества.

⚡ ВЫЗОВ 10.7

Изучение границ приемлемости уровня автономности ИИ для людей в различных сферах

Стремительное развитие технологий ИИ, постепенное появление общего ИИ (Artificial General Intelligence) ставит необходимость определения пределов делегирования функций принятия решений автономным системам. Данный вызов фокусируется на комплексном исследовании социальной приемлемости различных уровней автономии ИИ в рамках всего спектра жизненно важных сфер человеческой деятельности. Предметом изучения являются не столько технологические параметры систем, сколько комплекс социально-психологических, этических и культурных детерминант, формирующих отношение пользователей к системам ИИ.

4. Перспективные исследовательские задачи

★ ЗАДАЧА 10.1

Формирование подходов к глобальной системе управления ИИ

В настоящее время наблюдается постепенное развитие международного сотрудничества в сфере ИИ. Происходит становление подходов к глобальному регулированию ИИ, направленных на преобразование технологии в инструмент решения общечеловеческих проблем, а не в источник новых конфликтов. Для достижения этой цели требуется систематизация данных о национальных правовых системах, а также культурных, социальных и политических особенностях различных государств через разработку унифицированных инструментов оценки. Это позволит создать инклюзивные рамки глобального соглашения с участием всех стран, включая страны Глобального Юга. Необходимо создать механизмы межкультурного этического аудита для обеспечения справедливости технологий, развертываемых в уязвимых регионах. Их цель заключается не только в обеспечении доступа, но и в достижении корректности результатов работы ИИ в условиях локальных контекстов.

Проектируемым результатом решения данной задачи является достижение глобального соглашения по ИИ и/или учреждение международного независимого органа, координирующего функционирование глобальной системы управления ИИ.

★ ЗАДАЧА 10.2

Формирование национальных систем регулирования технологий ИИ

В последние годы наблюдается активный поиск странами эффективных моделей регулирования технологий ИИ, учитывающих национальную специфику. Разработка нормативно-правовых рамок ведется с акцентом на культурные особенности, стратегические приоритеты и существующие правовые системы каждой страны. Значимым аспектом этого процесса становится определение «красных линий» — законодательно установленных границ применения ИИ,

отражающих общественные ценности и этические принципы.

Проектируемым результатом должно стать создание сбалансированных национальных систем регулирования, обеспечивающих как развитие инноваций, так и защиту общественных интересов. Эти системы призваны гармонично интегрировать международные стандарты с национальной спецификой, формируя предсказуемые и прозрачные условия для разработки и внедрения ИИ-технологий.

★ ЗАДАЧА 10.3

Унификация стандартов и метрик социально-этической оценки ИИ и измерение социально-экономических последствий внедрения технологий ИИ

Для решения данной задачи требуется выработать сопоставимые метрики и прозрачные фреймворки для оценки недискриминации, объяснимости, конфиденциальности, безопасности и подотчетности. Существующие подходы фрагментированы, что осложняет сравнение систем между собой и снижает уровень доверия. Измерение эффектов на производительность, занятость, распределение доходов и концентрацию рыночной власти. Разработка сценариев трансформации рынка труда и механизмов справедливого распределения выгод.

Проектируемым результатом будет являться сформированная универсальная система оценки, позволяющая сопоставлять разные модели ИИ, повышать прозрачность и управляемость внедрения технологий, а также формировать основу для сертификации, валидированные ориентиры для государственной политики, включая регуляторные и фискальные инструменты, образовательные стратегии и социальные программы поддержки.

★ ЗАДАЧА 10.4

Исследование эффектов влияния технологий ИИ на общество

Новые поколения ИИ, включая ИИ-агентов, становятся всё более автономными, что усиливает риски непрозрачности и неконтролируемого поведения. Эти технологии способны изменить модели коммуникации, делегирования полномочий и функционирования социальных институтов. Ключевая задача заключается в том, чтобы определить ту составляющую человеческой природы, которую мы желаем сохранить в процессе технологического выбора. Решение задачи предполагает анализ психологических и социальных последствий делегирования решений и установление «права на отказ» при взаимодействии с системами ИИ.

Проектируемые результаты данного исследования, выражающиеся в изучении пределов приемлемости автономности ИИ для человека, позволят сформировать научную основу для разработки человеко-ориентированных интерфейсов и нормативно-правовых требований, обеспечивающих баланс между инновационным потенциалом технологий и защитой фундаментальных прав и безопасности граждан.

5. Важные выводы: Экспертное заключение

Учитывая трансграничный характер технологий ИИ, большинство экспертов и ученых сходятся во мнении о том, что поддержание международного диалога и разработка единого глобального соглашения по управлению ИИ являются сложнейшей и наиболее приоритетной задачей для блага общества и равномерного развития технологий ИИ.

Отмечается, что развитие технологий ИИ сопровождается закреплением базовых этических требований прозрачности, подотчетности, защиты данных и недискриминации. Однако подходы к их реализации остаются фрагментарными и различаются от страны к стране, что снижает сопоставимость практик и замедляет формирование единого поля доверия.

Социально-экономические последствия становятся всё более очевидными: автоматизация приводит к трансформации рынка труда и усиливает дискуссию о справедливом распределении выгод. На повестке стоят налогообложение автоматизации, новые формы социальной поддержки и развитие систем переквалификации.

Культурное измерение приобретает особое значение. С одной стороны, глобальные технологии тяготеют к стандартизации, с другой — сохраняется потребность в уважении культурного разнообразия, языковых и социальных норм. Уровень общественной приемлемости ИИ во многом определяется качеством просвещения граждан и их способностью критически воспринимать результаты работы систем.

Механизмы глобального управления ИИ должны основываться на открытом глобальном диалоге, в ходе которого регулирование ИИ рассматривается как залог устойчивого сосуществования. Создаваемая система регулирования должна носить социально-технический, а не сугубо технический характер, с признанием того, что ИИ является продуктом человеческого выбора, встроенным в существующие поля власти

Этика должна представлять собой процесс коллективного осмысления, а не просто свод жестких правил. Данный подход должен интегрировать принцип «морального плюрализма»

Проблема построения всеобъемлющей системы регулирования ИИ заключается не только в выборе подхода, но и в принятии гибкой нормативной рамки, способной эволюционировать вместе с техническим прогрессом

ЗАКЛЮЧЕНИЕ

Настоящее исследование было задумано как срез эпохи — масштабная попытка собрать в одном корпусе знания о траекториях ИИ, их ограничениях и окнах возможностей. Мы рассматривали ИИ не как набор разрозненных технологических трендов, а как взаимосвязанную систему: архитектуры и алгоритмы, вычисления и энергия, данные и право, фундаментальные и генеративные модели, безопасность и доверие, узкоспециализированные решения и мультиагентные системы, элементы рассуждения на пути к AGI, человеко-машинные интерфейсы, а также социально-экономические эффекты и регуляторика — 10 тем, складывающихся в единый ландшафт. Такой подход позволяет увидеть системные причинно-следственные связи: что именно сдерживает развитие, что его ускоряет и где пройдут границы масштабируемости в ближайшие годы.

Методологически работа опиралась на широкую международную кооперацию: 21 форсайт-сессию и интервью с более чем сотней ведущих исследователей из 36 стран, дополненные анализом открытых источников и отраслевых обзоров. Это сочетание академической оптики и практики внедрения позволило зафиксировать и зоны консенсуса (например, критичность энергопотребления и дефицита данных), и принципиальные расхождения (границы агентности, пути к объяснимости, архитектурные приоритеты), а главное — выявить узкие места инженерного цикла ИИ. В результате мы получили не статичную «фотографию», а динамическую карту, пригодную для регулярного обновления и сравнения по времени.

По итогам проведенного масштабного исследования в заключении отчета мы выделяем 3 опорных вывода:

Вывод 1

Траектория качества ИИ сегодня определяется не одной «волшебной кнопкой», а ко-оптимизацией «алгоритмы — ПО — железо» на фоне растущей стоимости вычислений и энергии.

Вывод 2

Наступает эпоха агентных и гибридных систем: переход от пассивных ассистентов к проактивным ИИ-агентам и к архитектурам, сочетающим статистическое обучение с явными знаниями и моделями мира.

Вывод 3

Устойчивость прогресса упирается в двойной дефицит — энергии и качественных данных; синтетические данные, активное и контекстное обучение, а также новые источники данных из динамических сред становятся не роскошью, а необходимостью.

Практическая ценность исследования — в том, что оно предлагает общий язык для ученых, инженеров, бизнеса и регуляторов. Мы избегали универсальных рецептов, но показали рамки решений: где уместны масштабные foundation-подходы, а где критична локальная адаптация и edge-ИИ; когда уместна централизация данных, а когда — федеративное обучение; как сопоставлять выигрыш в качестве и цену инференса; каким образом измерять доверие и объяснимость для реальных интерфейсов «человек — ИИ». Эта рамка помогает проектировать дорожные карты так, чтобы они оставались реалистичными и сопоставимыми между регионами и отраслями.

Мы рассматриваем данный выпуск как итерацию «живого» глобального наблюдателя за ИИ. Чтобы отслеживать сдвиги — от смены оптимизаторов и методов сжатия до всплесков в мультиагентных системах и новых классов ускорителей — необходима регулярность.

Международная кооперация при создании отчета является необходимой, т. к. энергетические и экологические ограничения, стандарты безопасности и приватности, оборот данных и моделей, совместимость инструментов и протоколов — всё это по определению трансгранично. Именно совместная работа лабораторий, компаний и регуляторов снижает транзакционные издержки прогресса и ускоряет трансфер знаний из наук об ИИ в практики здравоохранения, образования, промышленности и управления.

Мы благодарим всех участников за профессиональный и честный разговор. Этот отчет подводит итоги проделанной работы, но не ставит точку. Мы приглашаем партнеров из разных стран присоединиться к следующему раунду — чтобы вместе поддерживать общее поле знаний, снимать барьеры, тестировать гипотезы и делать развитие ИИ ответственным, экономичным и по-настоящему глобальным.

ПРИЛОЖЕНИЕ

НАПРАВЛЕНИЕ 1

Архитектуры, алгоритмы машинного обучения, оптимизация и математика

Поднаправление		Исследовательская задача	2025	2026	2027	2028	2029	2030	2031	2032	2033	2034	2035+
1.1. Разработка новых алгоритмов машинного обучения	1.1.1	Интеграция машинного обучения с экспертными системами (гибридный ИИ)											
	1.1.2	Развитие методов самообучения (self-supervised learning)											
	1.1.3	Создание универсальных методов обработки разнородных структур данных											
	1.1.4	Разработка алгоритмов для специализированных аппаратных комплексов и вычислительных устройств											
	1.1.5	Разработка методов контекстного и вербального обучения (вербальное обучение с подкреплением; индуктивное обучение; обучение мультиагентных систем, обратное распространение)											
	1.1.6	Разработка алгоритмов для многомерных сценариев с низкой выборкой (low-sample)											
	1.1.7	Разработка гибридных (классических и квантовых) алгоритмов, квантовые нейронные сети											
	1.1.8	Создание новых алгоритмов для предметно-ориентированного измерения качества моделей машинного обучения											
	1.1.9	Разработка самоэволюционирующихся алгоритмов ИИ											
1.2. ИИ-архитектуры	1.2.1	Создание адаптивных методов для архитектур глубоких сетей											
	1.2.2	Создание архитектур нейронных сетей, вдохновленных нейробиологией и психологией, включая спайковые нейронные сети											
	1.2.3	Развитие методов AutoML: метаобучение, автоматическое проектирование признаков, оптимизация гиперпараметров, автоматическое сжатие моделей и т. п.											
1.3. Ускорение вычислений	1.3.1	Развитие методов сжатия нейронных сетей: квантизация, teacher-student, network prunin и т. п.											
	1.3.2	Оптимизация вычислений для известных архитектур нейронных сетей (на этапах обучения и инференса). Структура и физика нейронных сетей (в т. ч. тензорных и матричных)											
	1.3.3	Разработка программных инструментов для ускорения вычислений											
1.4. Распределенное и федеративное обучение	1.4.1	Развитие методов оптимизации для распределенного и федеративного обучения больших моделей ИИ: снижение «накладных расходов» обмена данными, улучшение методов синхронизации распределенной модели и др.											
	1.4.2	Разработка методов распределенного децентрализованного обучения											
	1.4.3	Разработка архитектур (в т. ч. смешанных) для федеративного обучения											
	1.4.4	Предотвращение несанкционированного доступа к данным во время обработки и хранения при федеративном обучении											
1.5. Математические основы ИИ	1.5.1	Исследование математических основ понижения сложности моделей машинного обучения											

Год ожидаемого прорыва — год, в который прогнозируется значительное научное достижение, способное кардинально приблизить решение поставленной задачи и оказать существенное влияние на развитие технологий и мир в целом.

Год исчерпания задачи — год, в который задача может считаться решенной, и когда не ожидается новых фундаментальных или прикладных открытий в данной области.

Поднаправление		Исследовательская задача	2025	2026	2027	2028	2029	2030	2031	2032	2033	2034	2035+
	1.5.2	Разработка численных методов математической оптимизации для ИИ (включая теорию сложности)											
	1.5.3	Разработка численных методов линейной алгебры для ИИ											
	1.5.4	Исследование основ стохастических методов для ИИ											
	1.5.5	Развитие и расширение теории информации для ИИ											
	1.5.6	Исследование теории аппроксимации (объяснение поведения ИИ)											
	1.5.7	Создание состязательной устойчивости (применительно к он-лайн-обучению)											
	1.5.8	Исследование теоретических основ обучения с подкреплением, стохастического оптимального управления											
	1.5.9	Разработка математически обоснованных моделей глубокого обуче-ния меньшего размера											
	1.5.10	Исследование ландшафта целевых функций и разработка более эффективных целевых функций и путей обучения											
	1.5.11	Исследование теоретических основ генеративных моделей											
	1.5.12	Исследование теоретических основ моделей трансформеров											
	1.5.13	Формирование математических основ для понимания мульти-агентных моделей											
	1.5.14	Исследование теоретических основ работы с неопределенностью											

НАПРАВЛЕНИЕ 2

Вычисления для ИИ

Поднаправление		Исследовательская задача	2025	2026	2027	2028	2029	2030	2031	2032	2033	2034	2035+
2.1. Разработка специализированных вычислителей для ИИ (квантовых, фотонных, нейроморфных и др.)	2.1.1	Разработка архитектур специализированных аппаратных и вычислительных устройств, оптимизированных под архитектуры нейронных сетей											
	2.1.2	Создание процессоров на фотонных принципах и алгоритмов к ним для целей ИИ											
	2.1.3	Создание процессоров на нейроморфных принципах и их алгорит-мов для целей ИИ; сенсоры, окружение и исполнительные устройства для нейроморфных процессоров											
	2.1.4	Исследование и создание процессоров на квантовых принципах и алгоритмов к ним для целей ИИ											
	2.1.5	Создание системного ПО, повышающего эффективность работы с оборудованием											
	2.1.6	Разработка методов и моделей для повышения эффективности процесса обучения											
	2.1.7	Разработка методов адаптации чипов											
2.2. Разработка аппаратно-программных комплексов для ИИ	2.2.1	Создание высокоскоростной сети обмена данными между микро-процессорами											
	2.2.2	Создание новых, более доверенных библиотек (с использованием ИИ-генерации кода и ИИ-проверки существующего кода по критерию «силы доверия»)											

Год ожидаемого прорыва — год, в который прогнозируется значительное научное достижение, способное кардинально приблизить решение поставленной задачи и оказать существенное влияние на развитие технологий и мир в целом.

Год исчерпания задачи — год, в который задача может считаться решенной, и когда не ожидается новых фундаментальных или прикладных открытий в данной области.

Поднаправление		Исследовательская задача	2025	2026	2027	2028	2029	2030	2031	2032	2033	2034	2035+
2.3. Фреймворки машинного обучения и ИИ	2.3.1	Оптимизация вычислений, в т. ч. на гетерогенном оборудовании											
	2.3.2	Создание фреймворков для обучения и инференса											
	2.3.3	Создание фреймворков для символьного и гибридного ИИ (например, SymbolicAI Framework, IBM Neuro-Symbolic AI Toolkit (NSTK), SmythOS, классические инструменты символического ИИ (Prolog, экспертные системы, семантические сети/онтологии), нейросимвольные гибриды)											
	2.3.4	Создание фреймворков для агентных схем и приложений (включая воплощенных агентов) (например, AutoGen, CrewAI, LangChain, LlamaIndex, Semantic Kernel и т. д.)											
	2.3.5	Создание фреймворков для промпт-инжиниринга (например, RTF, COSTAR, IEEI, APE, PECRA, OSCAR и TAG)											

НАПРАВЛЕНИЕ 3

Данные для ИИ

Поднаправление		Исследовательская задача	2025	2026	2027	2028	2029	2030	2031	2032	2033	2034	2035+
3.1. Разработка бенчмарков для ИИ	3.1.1	Создание и поддержание бенчмарков для сравнения и оценки производительности нейронных сетей (CNN, Transformers, LLM и т. д.)											
	3.1.2	Создание метрик оценки ИИ											
	3.1.3	Онлайн-оценка методов ИИ (оценка в реальном мире)											
	3.1.4	Оценка эффективности обеспечения безопасности											
3.2. Формирование, преобразование и сопровождение данных	3.2.1	Разработка методов создания синтетических наборов данных											
	3.2.2	Активное обучение и краудсорсинг с привлечением экспертных сообществ, сбор субъективных данных от широкого круга людей											
	3.2.3	Исследования методов упорядочивания данных (в т. ч. curriculum learning)											
	3.2.4	Системы моделирования для генерации данных, обучения и проверки моделей в режиме онлайн											
	3.2.5	Разработка методов оценки качества данных, фильтрации, курирования и сортировки											
	3.2.6	Разработка методов аугментации данных											
	3.2.7	Создание открытых наборов данных для обучения крупномасштабных практических моделей											
	3.2.8	Разработка методов проверки и алгоритмов коррекции смещений и предвзятостей в данных для их уменьшения											
3.3. Обеспечение конфиденциальности и защиты данных	3.3.1	Разработка методов «дифференциальной приватности»: внесение зашумлений и искажений в данные											
	3.3.2	Разработка методов анонимизации данных: обеспечение невозможности восстановления конфиденциальных данных											
	3.3.3	Разработка методов обеспечения конфиденциальности данных при федеративном обучении											
	3.3.4	Маркировка личных и защищенных авторским правом данных водяными знаками и методы обнаружения использования таких данных в процессе обучения											
	3.3.5	Маркировка сгенерированного контента водяными знаками и методы обнаружения такого контента											

Год ожидаемого прорыва — год, в который прогнозируется значительное научное достижение, способное кардинально приблизить решение поставленной задачи и оказать существенное влияние на развитие технологий и мир в целом.

Год исчерпания задачи — год, в который задача может считаться решенной, и когда не ожидается новых фундаментальных или прикладных открытий в данной области.

НАПРАВЛЕНИЕ 4

Фундаментальные генеративные модели

Поднаправление		Исследовательская задача	2025	2026	2027	2028	2029	2030	2031	2032	2033	2034	2035+
4.1. Фундаментальные генеративные модели для символьных данных	4.1.1	Исследование и разработка методов обучения и настройки фундаментальных генеративных моделей											
	4.1.2	Исследование и разработка методов создания генеративных моделей (включая RL в различных приложениях)											
	4.1.3	Создание вычислительно эффективных архитектур фундаментальных генеративных моделей											
	4.1.4	Разработка высококачественной модели данных на основе генеративных моделей обучения											
	4.1.5	Исследование влияния качества данных на процессы обучения фундаментальных генеративных моделей											
	4.1.6	Разработка методов снижения влияния галлюцинаций и количественной оценки неопределенности в фундаментальных генеративных моделях											
	4.1.7	Исследование современных архитектур (включая архитектуры трансформерного типа) для различных задач последовательной обработки данных											
	4.1.8	Создание правдоподобных (plausible) методов генерации											
	4.1.9	Технологии имплементации рассуждений в фундаментальных генеративных моделях (с учетом различных предметных областей)											
	4.1.10	Разработка методов обучения представлениям (representation learning)											
4.2. Фундаментальные генеративные модели для несимвольных данных	4.2.1	Разработка базовых генеративных моделей для обработки изображений и видеоданных											
	4.2.2	Разработка базовых генеративных моделей для обработки временных рядов (сенсоры в робототехнике, лидары, датчики Интернета вещей)											
	4.2.3	Исследование, разработка и использование обучаемых представлений несимвольных данных											
	4.2.4	Развитие исследований и разработок фундаментальных генеративных моделей для 3D/пространственных данных											
	4.2.5	Разработка фундаментальных генеративных моделей для решения задач в области биологии, фармакологии, метеорологии и других научных областей											
	4.2.6	Разработка эффективных архитектур и методов для понимания и обработки видео, в т. ч. приложений VLM											
4.3. Мультимодальные фундаментальные генеративные модели	4.3.1	Разработка механизмов расширения фундаментальных языковых моделей способностью обрабатывать несимвольные модальности											
	4.3.2	Разработка методов эффективного кодирования данных несимвольных модальностей											
	4.3.3	Исследование новых методов смешивания энкодеров различных модальностей											
	4.3.4	Разработка и исследование мультимодальных генеративных моделей, адаптированных для узкоспециализированных доменных задач											
4.4. Трансфер знаний с адаптацией базовой генеративной модели	4.4.1	Разработка методов файнтюнинга фундаментальных генеративных моделей (например, LoRA, P-tuning)											
	4.4.2	Разработка методов дистилляции моделей для узкоспециализированных задач											
	4.4.3	Разработка методов персонализированной генерации в различных модальностях											

Год ожидаемого прорыва — год, в который прогнозируется значительное научное достижение, способное кардинально приблизить решение поставленной задачи и оказать существенное влияние на развитие технологий и мир в целом.

Год исчерпания задачи — год, в который задача может считаться решенной, и когда не ожидается новых фундаментальных или прикладных открытий в данной области.

Поднаправление		Исследовательская задача	2025	2026	2027	2028	2029	2030	2031	2032	2033	2034	2035+
4.5. Аугментация фундаментальных генеративных моделей	4.5.1	Исследование и создание кратковременной и долговременной памяти, а также RAG и графовых RAG											
	4.5.2	Исследование и создание методов контекстного обучения (in-context learning) и промпт-инжиниринга											
	4.5.3	Использование инструментов (например, API Calls) в базовых генеративных моделях											

НАПРАВЛЕНИЕ 5

Безопасность, доверие и объяснимость

Поднаправление		Исследовательская задача	2025	2026	2027	2028	2029	2030	2031	2032	2033	2034	2035+
5.1. Alignment	5.1.1	Разработка методов снижения рисков, связанных с неверными или вредоносными данными (неточные данные, данные с ограниченным доступом, включая информацию, распространение которой запрещено в соответствии с законом, данные, не соответствующие ценностям общества)											
	5.1.2	Разработка методов определения «следов» (footprints) алгоритмов											
	5.1.3	Создание методов для формирования ценностных установок в генеративных (языковых, мультимодальных) моделях широкого применения (ценности общества, культуры, отдельных групп), включая выравнивание множественных предпочтений											
	5.1.4	Разработка подходов к стандартизации разработки бенчмарков для измерения соответствия целям выравнивания, создание бенчмарков выравнивания											
	5.1.5	Выравнивание моделей с рассуждениями											
	5.1.6	Мультимодальное выравнивание											
	5.1.7	Машинное разобучение (unlearning)											
	5.1.8	Разработка методов исправления ошибок ИИ-моделей											
5.2. Объяснимость ИИ	5.2.1	Формулировка общих подходов для обеспечения объяснимости, повышения доверия работы ИИ — Explainable AI (XAI)											
	5.2.2	Исследование и создание методов post-factum объяснения (интерпретации) непрозрачных методов ИИ на основе нейронных сетей											
	5.2.3	Создание интерпретируемых и самопоясняемых моделей (Self-explainable AI)											
	5.2.4	Создание систем аргументации и концептуального обучения (Concept & Argumentation-based AI)											
	5.2.5	Исследование и создание методов прозрачного ИИ на основе формальных систем, логики и баз знаний											
	5.2.6	Разработка требований и протоколов тестирования систем AI на объяснимость											
	5.2.7	Создание человекоориентированных аспектов XAI (Human-centered XAI), в т. ч. учитывающих отраслевую специфику											
	5.2.8	Обеспечение объяснимости больших языковых и мультимодальных моделей (LLM-/VLM-XAI)											
5.3. Обеспечение безопасной разработки и эксплуатации ИИ	5.3.1	Создание методов и инфраструктуры для обеспечения безопасной разработки систем с ИИ (MLSecOps)											
	5.3.2	Разработка методов и программных инструментов для защиты от атак на датасеты (в т. ч. от отравления и закладок)											

Год ожидаемого прорыва — год, в который прогнозируется значительное научное достижение, способное кардинально приблизить решение поставленной задачи и оказать существенное влияние на развитие технологий и мир в целом.

Год исчерпания задачи — год, в который задача может считаться решенной, и когда не ожидается новых фундаментальных или прикладных открытий в данной области.

Поднаправление		Исследовательская задача	2025	2026	2027	2028	2029	2030	2031	2032	2033	2034	2035+
	5.3.3	Создание методов поиска уязвимостей в коде сторонних библиотек, применяемых при создании систем с ИИ											
	5.3.4	Разработка методов и программных инструментов защиты от атак на ИИ-модели во время обучения и исполнения											
	5.3.5	Создание бенчмарков для оценки безопасности и обеспечения доверия											
	5.3.6	Исследование фундаментальных границ для гарантий безопасности, включая математические аспекты с учетом эмпирических/эвристических методов											
	5.3.7	Развитие методов безопасной разработки кодовой базы, созданной с помощью инструментов ИИ											
	5.3.8	Применение методов федеративного обучения для снижения рисков работы с чувствительными данными											
	5.3.9	Обеспечение безопасности приложений LLM											
5.4. Обеспечение защиты от результатов использования ИИ с целью взлома	5.4.1	Разработка методов поиска (с использованием ИИ) уязвимостей в ПО, кибербезопасность, социальная инженерия, подделка информации и т. п.											
	5.4.2	Разработка методов обнаружения и защиты от дипфейков, включая водяные знаки											
	5.4.3	Разработка методов обнаружения атак на киберфизические системы, включая автономное вождение											
	5.4.4	Создание бенчмарков для методов обнаружения дипфейков											
	5.4.5	Выявление и формализация лучших практик по борьбе с уязвимостями агентов ИИ											

НАПРАВЛЕНИЕ 6

ИИ для узких задач (Narrow AI)

Поднаправление		Исследовательская задача	2025	2026	2027	2028	2029	2030	2031	2032	2033	2034	2035+
6.1. Компьютерное зрение (CV)	6.1.1	Исследование и разработка эффективных методов обучения и исполнения сверточных, трансформерных и гибридных архитектур для задач CV (включая AutoML)											
	6.1.2	Создание фундаментальных моделей для различных задач CV (типа SAM, Dyno v2, CLIP, генеративных VLM) (в т. ч. решение задач классификации, обнаружения, сегментации с открытым списком классов)											
	6.1.3	Развитие методов компьютерного зрения для симуляции сценариев реального мира (в т. ч. воплощенный ИИ)											
	6.1.4	Разработка методов fine tuning для специфических задач компьютерного зрения											
	6.1.5	Создание эффективных пространственных представлений (мультимодальные, мультисенсорные, NeRF, Gaussian Splatting и т. д.) для решения задач компьютерного зрения											
6.2. Обработка естественного языка (NLP)	6.2.1	Исследование и разработка эффективных методов обучения и выполнения для архитектур обработки естественного языка (включая AutoML)											
	6.2.2	Создание NLP для различных языков программирования											

Год ожидаемого прорыва — год, в который прогнозируется значительное научное достижение, способное кардинально приблизить решение поставленной задачи и оказать существенное влияние на развитие технологий и мир в целом.

Год исчерпания задачи — год, в который задача может считаться решенной, и когда не ожидается новых фундаментальных или прикладных открытий в данной области.

Поднаправление	Исследовательская задача		2025	2026	2027	2028	2029	2030	2031	2032	2033	2034	2035+
6.3. Другие узкие ИИ-технологии (S2T, RecSys, TSA и т. д.)	6.3.1	Исследование и разработка эффективных методов обучения и выполнения для архитектур RecSys, S2T и TSA (включая автоматическое обучение)											

НАПРАВЛЕНИЕ 7

Управление, принятие решений и агентные/мультиагентные системы

Поднаправление	Исследовательская задача		2025	2026	2027	2028	2029	2030	2031	2032	2033	2034	2035+
7.1. Разработка алгоритмов обучения с подкреплением	7.1.1	Разработка классического обучения с подкреплением (RL): Distributional RL, Risk- Constrained RL, Entropy-regularized RL, Safe RL и др.											
	7.1.2	Разработка предобучения с открытым списком сред и задач											
	7.1.3	Развитие политик RL в сложных динамически изменяющихся средах: Meta-Learning, Hierarchical RL, in-context RL и др.											
	7.1.4	Разработка Offline-to-Online RL и Real-time learning											
	7.1.5	Разработка RL с использованием генеративных моделей и агентов (Agentic RL)											
	7.1.6	Развитие методов переноса обучения между реальными и виртуальными средами											
	7.1.7	Развитие самосовершенствующихся алгоритмов (Self-Evolving Algorithms)											
	7.1.8	Развитие алгоритмов обратного обучения с подкреплением (Inverse RL)											
	7.1.9	Разработка алгоритмов для высокоразмерных сценариев с малым количеством примеров											
	7.1.10	Технологии unsupervised Reinforcement Learning											
7.2. Агентные системы	7.2.1	Создание универсальных моделей действий физического агента: физические манипуляции, многозадачность											
	7.2.2	Создание универсальных многомодальных моделей, объединяющих работу с текстом и другими модальностями: Vision- Language-Action (VLA)											
	7.2.3	Создание эффективных методов получения знаний агентами посредством взаимодействия с окружающей средой для достижения целей											
	7.2.4	Создание фундаментальных агентов для виртуального и физического мира: многозадачность, самообучение, работа в открытой среде, с открытым списком инструментов											
	7.2.5	Разработка методов обучения на неразмеченных поведенческих данных (Learning Without Actions)											
	7.2.6	Разработка методов автоматизированного конструирования агентов											
7.3. Мультиагентные системы	7.3.1	Разработка мультиагентных архитектур на основе базовых моделей											
	7.3.2	Развитие возможностей обмена знаниями для агентов: обучение у других агентов и у людей											
	7.3.3	Исследование и создание системы агентов: «мастер — система» для создания узких агентов, управление ресурсами, устойчивость, эволюционный отбор и др.											

Год ожидаемого прорыва — год, в который прогнозируется значительное научное достижение, способное кардинально приблизить решение поставленной задачи и оказать существенное влияние на развитие технологий и мир в целом.

Год исчерпания задачи — год, в который задача может считаться решенной, и когда не ожидается новых фундаментальных или прикладных открытий в данной области.

Поднаправление	Исследовательская задача		2025	2026	2027	2028	2029	2030	2031	2032	2033	2034	2035+
	7.3.4	Исследование и создание виртуальных и физических мультиагентных систем, использующих различные сценарии взаимодействия между агентами для принятия групповых решений: конкуренция, координация, кооперация, рой											
	7.3.5	Изучения явления фазовых переходов в мультиагентных системах с увеличением числа агентов											
	7.3.6	Автоматизированное конструирование и оптимизация топологии мультиагентных систем											
	7.3.7	Разработка эффективных сред для мультиагентных систем (текстовых, мультимодальных)											
	7.3.8	Создание иерархических мультиагентных систем (сети сетей и системы систем)											
	7.3.9	Создание методов научных и инженерных приложений мультиагентных систем											

НАПРАВЛЕНИЕ 8

Элементы AGI

Поднаправление	Исследовательская задача		2025	2026	2027	2028	2029	2030	2031	2032	2033	2034	2035+
8.1. Рассуждения и рефлексия	8.1.1	Разработка общих моделей рассуждения (reasoning) и схем принятия решений											
	8.1.2	Формирование и использование модели мира и модели себя											
	8.1.3	Разработка методов и подходов к моделированию самосознания, рефлексии и самокритики											
	8.1.4	Разработка подходов к целеполаганию и планированию											
	8.1.5	Распознавание, анализ и управление вводящим в заблуждение поведением											
8.2. Lifelong learning	8.2.1	Разработка методов активного и пассивного обучения											
	8.2.2	Разработка методов онлайн- и оффлайн-обучения											
	8.2.3	Разработка методов обучения, включающих непосредственное взаимодействие с окружением или использование предварительно обработанных данных											
	8.2.4	Анализ работы моделей при нестабильности данных (concept feature drift)											
8.3. Гибридный ИИ	8.3.1	Разработка универсальных механизмов извлечения формализованных знаний и генерации новых знаний											
	8.3.2	Совместное использование инструментов машинного обучения, символьного ИИ и компьютерного моделирования											
	8.3.3	Интуиция и эмоции моделей ИИ											
	8.3.4	Разработка VLA (Vision-Language-Action) моделей											
	8.3.5	Оценка AGI моделей											

Год ожидаемого прорыва — год, в который прогнозируется значительное научное достижение, способное кардинально приблизить решение поставленной задачи и оказать существенное влияние на развитие технологий и мир в целом.

Год исчерпания задачи — год, в который задача может считаться решенной, и когда не ожидается новых фундаментальных или прикладных открытий в данной области.

Поднаправление		Исследовательская задача	2025	2026	2027	2028	2029	2030	2031	2032	2033	2034	2035+
8.4. Embodiment (вещественный ИИ)	8.4.1	Создание фундаментальных моделей для управления роботизированной системой (Cross-Embodiment and Cross-Reality Generalization)											
	8.4.2	Разработка систем ИИ для автономных роботов											
	8.4.3	Разработка технологий мультимодального и мультисенсорного ввода											
8.5. Моделирование мозга и психики	8.5.1	Моделирование работы мозга и нервной системы											
	8.5.2	Моделирование процессов работы человеческой психики											
	8.5.3	Создание двойников личности											
	8.5.4	Исследование и создание механизмов запоминания и забывания											

НАПРАВЛЕНИЕ 9

Взаимодействие человека и машины

Поднаправление		Исследовательская задача	2025	2026	2027	2028	2029	2030	2031	2032	2033	2034	2035+
9.1. Технические средства прямого взаимодействия с нервной системой человека	9.1.1	Исследование и разработка двунаправленных интерфейсов «мозг — компьютер»											
	9.1.2	Разработка нового оборудования и материалов для инвазивных ИМК (высокоплотные электроды, биосовместимость)											
	9.1.3	Разработка нового оборудования для неинвазивных ИМК (носимые гаджеты, ЭЭГ ультравысокой плотности, средства регистрации активности автономной нервной системы)											
	9.1.4	Разработка технологий и методов сжатия и передачи сигналов нейрональной активности											
	9.1.5	Разработка новых способов формирования естественного контакта нервной ткани с кибернетическим устройством (синаптические нейроинтерфейсы, наночастицы)											
	9.1.6	Ко-дизайн аппаратно-программных модулей для многоканальной регистрации и стимуляции активности мозга											
	9.1.7	Создание специализированных микросхем и ПО для обработки бионейросигналов											
	9.1.8	Создание легковесных архитектур и алгоритмов для real-time мультимодального слияния на устройствах											
9.2. Технические средства традиционного человеко-машинного взаимодействия	9.2.1	Создание технологий генеративного, адаптивного и персонализированного воздействия на человека											
	9.2.2	Создание средств для формирования коллективов людей и ИИ-агентов											
	9.2.3	Создание мультимодальных иммерсивных сред для повышения эффективности взимодействия											
	9.2.4	Создание мобильных устройств для предъявления видео, вибро-тактильной и ольфакторной обратной связи											
	9.2.5	Создание фундаментальных моделей для учета социального контекста в HMI											

Год ожидаемого прорыва — год, в который прогнозируется значительное научное достижение, способное кардинально приблизить решение поставленной задачи и оказать существенное влияние на развитие технологий и мир в целом.

Год исчерпания задачи — год, в который задача может считаться решенной, и когда не ожидается новых фундаментальных или прикладных открытий в данной области.

Поднаправление		Исследовательская задача	2025	2026	2027	2028	2029	2030	2031	2032	2033	2034	2035+
9.3. Методы и алгоритмы взаимодействия с человеком	9.3.1	Исследования «кода мозга» и создание фундаментальных моделей данных функционального картирования мозга											
	9.3.2	Разработка алгоритмов эфективного функцинирования коллективов людей и ИИ-агентов (Human-Machine Teaming)											
	9.3.3	Разработка алгоритмов декодирования активности мозга в речевых, моторных и зрительных нейропротезах											
	9.3.4	Разработка алгоритмов декодирования нейромиографической активности											
	9.3.5	Когнитивное нейропротезирование (восстановление памяти)											
	9.3.6	Исследование эффективных и адаптивных человеко-машинных интерфейсов (ЧМИ) в различных модальностях											
	9.3.7	Создание интуитивных агентов, понимающих запросы и ожидания пользователя, его эмоциональное состояние и т. п.											
	9.3.8	Мультимодальные машинные интерфейсы. Иммерсивное взаимодействие в смешанной реальности											
	9.3.9	Расширение возможностей человека посредством взаимодействия с ИИ с использованием интерфейсов «мозг-компьютер»											
	9.3.10	Установление метрологической базы для оценки HMI											

НАПРАВЛЕНИЕ 10

Общество в эпоху ИИ

Поднаправление		Исследовательская задача	2025	2026	2027	2028	2029	2030	2031	2032	2033	2034	2035+
10.1. Механизмы глобального управления ИИ, включая регулирование ИИ	10.1.1	Формирование подходов к глобальной системе управления ИИ											
	10.1.2	Формирование национальных систем регулирования технологий ИИ											
10.2. Этика ИИ	10.2.1	Разработка и институционализация методик оценки этических последствий и воздействия на права человека											
10.3. Изучение эффектов влияния технологий ИИ на общество	10.3.1	Исследование эффектов влияния технологий ИИ на общество											

Год ожидаемого прорыва — год, в который прогнозируется значительное научное достижение, способное кардинально приблизить решение поставленной задачи и оказать существенное влияние на развитие технологий и мир в целом.

Год исчерпания задачи — год, в который задача может считаться решенной, и когда не ожидается новых фундаментальных или прикладных открытий в данной области.

АВТОРЫ ИТОГОВОГО ОТЧЕТА

Главная редакционная коллегия

Абрахам Аджит
Профессор, Вице-канцлер, Университет Сай
Индия

Аветисян Арутюн
Академик РАН, директор Института системного программирования им. В.П. Иванникова Российской академии наук
Россия

Бачанин-Джакула Небойша
Профессор, проректор по научной работе, руководитель программы прикладного ИИ, Университет Сингидунум
Сербия

Баэса-Йатес Рикардо
Ординарный профессор, Университет Помпеу Фабра
Испания

Бурнаев Евгений
Доктор физико-математических наук, директор Центра ИИ Сколковского института науки и технологий (Сколтех)
Россия

Бухановский Александр
Директор мегафакультета трансляционных информационных технологий, научный руководитель исследовательского центра в сфере ИИ «Сильный ИИ в промышленности»
Россия

Визильтер Юрий
Директор по направлению «ИИ и техническое зрение» государственного научно-исследовательского института авиационных систем, научный директор Института ИИ МФТИ, доктор физико-математических наук, профессор РАН
Россия

Гутер Кристоф
Кандидат наук, Генеральный директор компании g.tec medical engineering GmbH
Австрия

Гасников Александр
Ректор Университета Иннополис
Россия

Дарвиш Ашраф
Декан факультета вычислительной техники и ИИ, Университет Обура науки и технологий
Египет

Кузнецов Андрей
Директор лаборатории FusionBrain, AIRI; Исполнительный директор по исследованию данных, ПАО Сбербанк
Россия

Лян Чжэн
Заместитель декана Института международного управления ИИ, Университет Цинхуа
Китай

Незнамов Андрей
Кандидат юридических наук, Управляющий директор Центра человекоцентричного ИИ, ПАО Сбербанк; Генеральный секретарь, Международный Альянс в сфере ИИ
Россия

Оселедец Иван
Доктор физико-математических наук, профессор РАН, генеральный директор Института AIRI
Россия

Осадчий Алексей
Директор центра биоэлектрических интерфейсов, НИУ ВШЭ
Россия

Ан Хуэй Фан
Профессор; руководитель лаборатории интеллектуальной обработки сигналов и изображений, Центр ИИ, Сколтех
Вьетнам

Роша Андерсон
Профессор и руководитель лаборатории ИИ, Институт вычислительной техники, Университет Кампинаса (Unicamp)
Бразилия

Тянь Е
Профессор, Университет Аньхой
Китай

Хаоу Хаовэнь
Доцент Гуандунской лаборатории ИИ и цифровой экономики, Университет Шэньчжэня
Китай

Юдин Дмитрий
Кандидат технических наук, ведущий научный сотрудник Лаборатории когнитивных систем ИИ AIRI, заведующий Лабораторией интеллектуального транспорта Центра когнитивного моделирования Института ИИ МФТИ
Россия

Ахатов Акмаль

Доктор технических наук, профессор, проректор по международному сотрудничеству, Самаркандский государственный университет имени Шарофа Рашидова

Узбекистан

Альмеида Круз Юдивиан

Дата-журналист, исследователь ИИ,
профессор факультета математики и информатики, Гаванский университет

Куба

Атхавале Виджай Анант

Проректор, профессор кафедры информатики и инженерии, Технологический институт Уолчанд

Индия

Баскин Хаим

Доцент Школы электротехники и вычислительной техники, Университет имени Бен-Гуриона; руководитель лаборатории INSIGHT, сотрудник исследовательского центра Data Science

Израиль

Белло Перес Рафаэль

Профессор, директор Научно-исследовательского центра информатики, руководитель группы исследований ИИ,
Центральный университет Лас-Вильяс

Куба

Боровик Рустам

Руководитель направления ИИ-трансформации Центра человекоцентричного ИИ, ПАО Сбербанк

Россия

Буэдо Идальго Денис

Магистр наук, директор по компьютеризации Министерства высшего образования

Куба

Буденный Семен

Управляющий директор Управления развития перспективных технологий ИИ, ПАО Сбербанк

Россия

Бушков Николай

Инжиниринг продуктивности R&D Архитектура, R&D Центр, T-Технологии

Россия

Вирири Серестина

Профессор (информатика). Руководитель группы исследований компьютерного зрения и машинного обучения, Университет КваЗулу-Наталь (UKZN)

ЮАР

Витулёва Елизавета

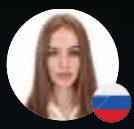
Алматинский университет энергетики и связи имени Гумарбека Даукеева; Казахский национальный университет имени аль-Фараби

Казахстан

Гарсия Боррото Милтон

Старший научный сотрудник Центра сложных систем физического факультета, Гаванский университет

Куба



Гвоздырева Арина

Эксперт Центра человекоцентричного ИИ, ПАО Сбербанк

Россия

Гончарова Елизавета

Кандидат технических наук, руководитель группы исследований мультимодальности, лаборатория FusionBrain, Институт ИИ AIRI

Россия

Гульни Аакаш

Специалист по государственной политике, Фонд «Цифровая Индия»

Индия

Гэззаз Азидин

Руководитель исследовательской группы SISAR, Университет Кади Айяд

Марокко

Димитров Денис

Управляющий директор по исследованию данных — начальник управления базовых моделей Kandinsky, ПАО Сбербанк

Россия

Динькуэй Ван

Профессор, директор Исследовательского центра ИИ для инженерии, Университет Циндао

Китай

Ершов Егор

Кандидат технических наук, руководитель группы исследований вычислительной цветной фотографии, Институт ИИ AIRI

Россия

Жавад Моханад Али Мохаммед

Генеральная компания услуг воздушной навигации

Ирак

Илюшин Евгений

Ассистент кафедры информационной безопасности, факультет вычислительной математики и кибернетики, МГУ имени М.В. Ломоносова, Директор по ИИ, «Окко»

Россия

Ириате Леонель

Старший научный сотрудник и руководитель исследовательской группы по ИИ, DATYS

Куба

Кабальеро Мота Ялие

Профессор и руководитель исследовательской группы по ИИ в Университете Камагуэя; Член Международной академии наук

Куба

Карпов Алексей

Руководитель лаборатории речевых и многомодальных интерфейсов, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук

Россия

Кетеча Кетан
Декан инженерного факультета, Университет «Симбиоз Интернешнл»
Индия

Конушин Антон
Кандидат физико-математических наук, руководитель группы исследований Spatial AI, Институт ИИ AIRI
Россия

Колюбин Сергей
Доктор технических наук, профессор, руководитель исследовательской лаборатории BE2R, руководитель магистерской программы по робототехнике и ИИ, заместитель директора Школы компьютерных технологий и управления, ИТМО
Россия

Коста Жоао Пита
Руководитель исследований ИИ в Международном исследовательском центре по ИИ под эгидой ЮНЕСКО — IRCAI
ЮНЕСКО

Крайнов Александр
Директор по развитию технологий ИИ, «Яндекс»
Россия

Лапина Мария
Доцент кафедры вычислительной математики и кибернетики, факультет математики и компьютерных наук, Северо-Кавказский федеральный университет
Россия

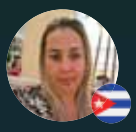
Линь Синь
Профессор, Восточно-Китайский педагогический университет
Китай

Маматов Нарзилло
Доктор технических наук, профессор, заведующий кафедрой «Цифровые технологии и ИИ», НИУ «Ташкентский институт инженеров ирригации и механизации сельского хозяйства»
Узбекистан

Мансурова Мадина
Кандидат физико-математических наук, профессор, заведующая кафедрой ИИ и Big Data, Казахский национальный университет имени аль-Фараби
Казахстан

Марков Сергей
Директор по развитию технологий ИИ, ПАО Сбербанк
Россия

Массимо Мечелла
Профессор, Университет Сапиенца в Риме, Департамент компьютерной, автоматизированной и управленческой инженерии Антонио Руберти (DIAG)
Италия



Мендес Васкес Хейди
Директор и старший научный сотрудник Центра передовых технологических приложений, распознавания образов и интеллектуального анализа данных (CENATAV)
Куба

Милованович Владимир
Профессор электротехники и информатики, Университет Крагуеваца
Сербия

Монтесино Перурена Райдель
Доктор технических наук, профессор. Ректор, Университет информатики
Куба

Моралес Гонсалес Аннет
Старший научный сотрудник Центра передовых технологических приложений, распознавания образов и интеллектуального анализа данных (CENATAV)
Куба

Мохаммад Абдул Кадир
Доцент кафедры компьютерной инженерии, Алигархский мусульманский университет
Индия

Мухамедиева Дилноз
Доктор технических наук, , профессор, НИУ «Ташкентский институт инженеров ирригации и механизации сельского хозяйства»
Узбекистан

Назаров Файзулло
Кандидат наук, декан факультета ИИ и цифровых технологий, Самаркандский государственный университет имени Шарофа Рашидова
Узбекистан

Наумов Алексей
Директор Института ИИ и цифровых наук, НИУ ВШЭ
Россия

Николенко Сергей
Старший научный сотрудник, ПДМИ РАН; руководитель образовательной программы A1360 «Математика машинного обучения» факультета математики и компьютерных наук, Санкт-Петербургский государственный университет
Россия

Нилакандан Субрамани
Профессор, Лаборатория интеллектуальных нейрокогнитивных знаний и анализа данных — SNKDIR, Инженерный колледж R.M.K.
Индия

Ойха Апараджита
Профессор компьютерных наук и инженерии, главный специалист, Академия электроники и ИКТ, Индийский институт информационных технологий, дизайна и производства (PDPM)
Индия

Олаванде Дарамола
Профессор, кафедра информатики, Университет Претории
ЮАР

Панов Александр
Доктор физико-математических наук, руководитель лаборатории когнитивных систем ИИ, Институт ИИ AIRI
Россия

Патель Самприт

Доцент, кафедра информатики и инженерии, Технологический институт Уолчанд

Индия

Пиньеро Перес Педро Йобанис

Субдиректор Центра Исследований, Развития и Инновации в ИИ (CIDIIA) Центрального Университета Востока

Доминиканская Республика

Плас Морффис Алехандро

Информатик, исследователь ИИ на факультете математики и информатики, Гаванский университет

Куба

Погосян Эдуард

Профессор, кандидат технических наук, руководитель направления когнитивных алгоритмов и моделей, Институт проблем информатики и автоматизации НАН

Армения

Прити Ханна

Профессор компьютерных наук и инженерии, декан по исследованиям, спонсируемым и консультационным проектам, Индийским институт информационных технологий, дизайна и производства (PDPM)

Индия

Пьера Фуэнтес Алан

Магистр наук, ассистент профессора Университета информационных наук; руководитель проектов по ИИ в Научно-технологическом парке Гаваны

Куба

Пэн Чэнь

Профессор, заведующий кафедрой интеллектуальных наук и технологий, Аньхойский университет

Китай

Рана Файяз Ахмад

Директор Центра ИИ-технологий (AITeC) Национального Центра Физики

Пакистан

Рахимето Самуэль

Магистр наук, директор отдела интерпретируемых исследований и NLP, Эфиопский институт ИИ EAII

Эфиопия

Рашидов Акбар

Кандидат технических наук, доцент кафедры ИИ и информационных систем, Самаркандский государственный университет имени Шарофа Рашидова

Узбекистан

Родригес Фигерэдо Гектор

Магистр наук, вице-президент Научно-технологического парка Гаваны; представитель Кубы в Альянсе в сфере ИИ (AI Alliance Network)

Куба

Рогов Олег

Кандидат физико-математических наук, руководитель группы исследований надежных и безопасных интеллектуальных систем, Институт ИИ AIRI; руководитель SAIL Lab, AIRI-МТУСИ

Россия

Рустамов Самир

Кандидат компьютерных наук, доцент, ADA University

Азербайджан

Савченко Андрей

Профессор, доктор технических наук, научный руководитель, Лаборатория ИИ ПАО Сбербанк

Россия

Саджид Мохаммад

Доцент кафедры компьютерных наук, Алигархский мусульманский университет

Индия

Самсонов Сергей

Заведующий Международной лабораторией стохастических алгоритмов и анализа многомерных данных, НИУ ВШЭ

Россия

Сафири Сри

Руководитель групповой корпоративной трансформации, «Телеком Индонезия»

Индонезия

Сеперо Найма

Профессор факультета компьютерной инженерии; Руководитель исследовательской группы по ИИ, Технологический университет Гаваны (CUJAE)

Куба

Скрынник Алексей

Кандидат физико-математических наук, руководитель группы «RL агенты» лаборатории когнитивных систем ИИ, Институт ИИ AIRI

Россия

Судьяна Доди

Профессор, кафедра электротехники, инженерный факультет, Университет Индонезии

Индонезия

Сулейменов Ибрагим

Доктор технических наук, профессор кафедры «Smart технологии в инженерии», Международный инженерно-технический университет

Казахстан

Сумбваньямбе Мбую

Заведующий кафедрой компьютерных наук, Южно-Африканский университет

ЮАР

Торралбас Эспелета Рафаэль Луис

Магистр наук, президент Научного и технологического парка Гаваны; представитель Кубы в Альянсе в сфере ИИ (AI Alliance Network)

Куба

Тутубалина Елена

Доктор компьютерных наук, руководитель группы «Прикладное NLP» Института ИИ AIRI; старший научный сотрудник Института системного программирования имени В.П. Иванникова Российской академии наук

Россия

Турдаков Денис
Руководитель Исследовательского центра доверенного ИИ, Институт системного программирования имени В.П. Иванникова Российской академии наук
Россия

Уткин Лев
Доктор технических наук, профессор Высшей школы технологий ИИ, Санкт-Петербургский политехнический университет Петра Великого
Россия

Феблес Эстрада Айлин
Профессор; Заместитель министра Министерства связи
Куба

Фазилов Шавкат
Доктор технических наук, профессор, заведующий лабораторией «ИИ и машинное обучение», Институт цифровых технологий и ИИ
Узбекистан

Феррер Мингес Гонсало
Доцент, руководитель лаборатории мобильной робототехники, Центр ИИ, Сколтех
Испания / Россия

Фэйвэй Цинь
Профессор факультета компьютерных наук и технологий, Ханчжоуский университет Дяньцзы
Китай

Хамдамов Рустам
Доктор технических наук, профессор, заведующий лабораторией «Умные системы и Интернет вещей», НИИ развития цифровых технологий и ИИ при Министерстве цифровых технологий Республики Узбекистан
Узбекистан

Лавров Игорь
Доктор медицинских наук, руководитель группы нейромодуляции спинного мозга, клиника Майо, Рочестер
Россия

Асланян Левон
Доктор физико-математических наук, профессор, чл.-корр. НАН РА, заведующий отделом «Дискретная математика» Института проблем информатики и автоматизации Национальной Академии наук Республики Армения
Армения

Хаммадов Мунис
Руководитель исследовательской лаборатории ИИ, доцент кафедры ИИ, Самаркандский государственный университет
Узбекистан

Хернандес Эредия Янио
Профессор, руководитель исследовательской группы по ИИ, Университет информационных наук (UCI); Президент KAINOS, S.A.
Куба

Хуссейн Эссам Халим
Профессор ИИ, декан факультета вычислительной техники и информатики, Университет Минии
Египет

Чаншэн Чэнь
Заместитель декана Института международного управления ИИ, Университет Цинхуа
Китай

Чекалина Виктория
Кандидат технических наук, старший научный сотрудник, группа исследований мультимодальности, лаборатория FusionBrain, Институт ИИ AIRI
Россия

Шпильман Алексей
Управляющий директор Центра «ИИ для науки», ПАО Сбербанк
Россия

Эстевес Рамс Эрнесто
Старший научный сотрудник, профессор физического факультета Гаванского университета; заслуженный академик Кубинской академии наук
Куба

Ален А. Гарофало Эрнандес
Доктор философии, Директор по продуктам, Авангенио СРЛ
Куба

Риза Хаммам
Президент KORIKA (Ассоциация ИИ Индонезии); представитель Индонезии в Альянсе в сфере ИИ (AI Alliance Network)
Индонезия

Выражаем искреннюю благодарность Горбаню Александру Николаевичу за участие в данном проекте. Александр Николаевич был великим ученым, чьи труды внесли значительный вклад в развитие современной науки.

Его исследования охватывали широкий спектр областей — от статистической физики и неравновесной термодинамики до машинного обучения и математической биологии.

Вечная светлая память.

Горбань Александр Николаевич
Руководитель лаборатории ИИ, анализа данных и моделирования Центрального университета и Института ИИ AIRI
Россия

КОМАНДА ИТОГОВОГО ОТЧЕТА

Незнамов Андрей

Управляющий директор Центра человекоцентричного ИИ, кандидат юридических наук, ПАО Сбербанк; Генеральный секретарь, Международный Альянс в сфере ИИ

Артюгин Олег

Исполнительный директор, дирекция реализации и популяризации ИИ-инициатив, ПАО Сбербанк

Боровик Рустам

Руководитель направления ИИ-трансформации Центра человекоцентричного ИИ,
ПАО Сбербанк

Лисов Арсений

Руководитель направления Проектного офиса «Стратегическое агентство поддержки и формирования ИИ-разработок»

Гвоздырева Арина

Эксперт Центра человекоцентричного ИИ,
ПАО Сбербанк

Чаче Эльвира

Исполнительный директор Центра человекоцентричного ИИ,
ПАО Сбербанк

Чурилова Дарья

Руководитель направления Центра человекоцентричного ИИ,
ПАО Сбербанк

Фокина Софья

Руководитель направления Центра человекоцентричного ИИ,
ПАО Сбербанк

Земцова Юлия

Ведущий исследователь данных, дирекция реализации и популяризации ИИ-инициатив, ПАО Сбербанк

БЛАГОДАРНОСТИ

Коллектив редакторов и организаторов выражает благодарность:

Правительству Российской Федерации и Министерству экономического развития Российской Федерации за поддержку в реализации данного проекта и неоценимый вклад в развитии науки ИИ в России в лице:

Чернышенко Дмитрия
Заместителя Председателя Правительства Российской Федерации

Колесникова Максима
Первого заместителя Министра экономического развития Российской Федерации

Сотниковой Оксаны
Директора Департамента стратегического развития и инноваций Министерства экономического развития Российской Федерации

Стратегическому агентству поддержки и формирования ИИ-разработок (САПФИР) за организацию и поддержку данного проекта в лице:

Союзновой Татьяны
Директора Проектного офиса «Стратегическое агентство поддержки и формирования ИИ-разработок»

Землянухина Кирилла
Исполнительного директора Проектного офиса «Стратегическое агентство поддержки и формирования ИИ-разработок»

Ассоциации «Альянс в сфере искусственного интеллекта» за поддержку данного проекта в лице:

Воробьевой Валерии
Генерального директора Ассоциации «Альянс в сфере искусственного интеллекта»

ПАО Сбербанк за организацию и поддержку данного проекта в лице:

Ведяхина Александра
Первого заместителя Председателя Правления ПАО Сбербанк

Белевцева Андрея
Старшего вице-президента — руководителя блока «Технологическое развитие» ПАО Сбербанк

Еременко Максима
Вице-президента-директора департамента развития ИИ и машинного обучения ПАО Сбербанк

Авербаха Владимира
Старшего управляющего директора дирекции реализации и популяризации ИИ-инициатив ПАО Сбербанк

БЛАГОДАРНОСТИ

Коллектив редакторов и организаторов выражает благодарность:

Институту исскусственного интеллекта AIRI
за привлечение ученых к работе над итоговым отчетом
в лице:

Бройтман Александры
Директора по маркетингу и коммуникациям Института ИИ AIRI

Певной Елизаветы
PR-менеджера Института ИИ AIRI

Команде генеративной сети Kandinsky
ПАО Сбербанк за создание иллюстрационных
материалов:

Никулиной Татьяне
Руководителю направления Управления базовых моделей Kandinsky,
ПАО Сбербанк

За создание дизайна итогового отчета:

Чмир Ольге
Эксперту Центра перспективных ИИ-разработок в индустриях

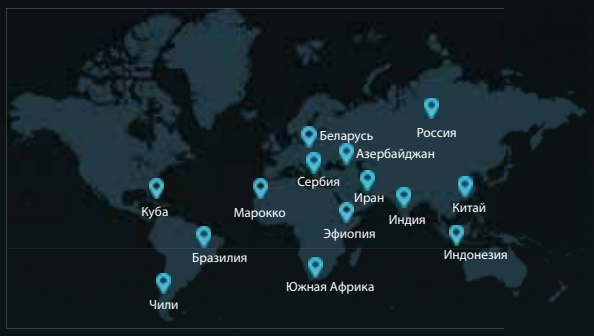
ИНФОРМАЦИЯ ОБ ОРГАНИЗАТОРЕ: МЕЖДУНАРОДНЫЙ АЛЬЯНС В СФЕРЕ ИИ

AI ALLIANCE NETWORK

aianet.org

AI Alliance Network

международное сообщество, объединяющее ассоциации в области ИИ



20+ стран

25+ ассоциаций в области ИИ

7000+ аффилированных парнеров





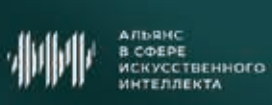







Министерство
экономического развития
Российской Федерации

САПФИР

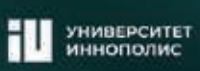


Сколтех Центр искусственного интеллекта

МФТИ

ИТМО
ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР
«ОНЛАЙН ИИ В ПРОМЫШЛЕННОСТИ»

ИСП РАН



Московский
государственный
университет
имени М.В.Ломоносова

Центр ИИ
СПбГУ