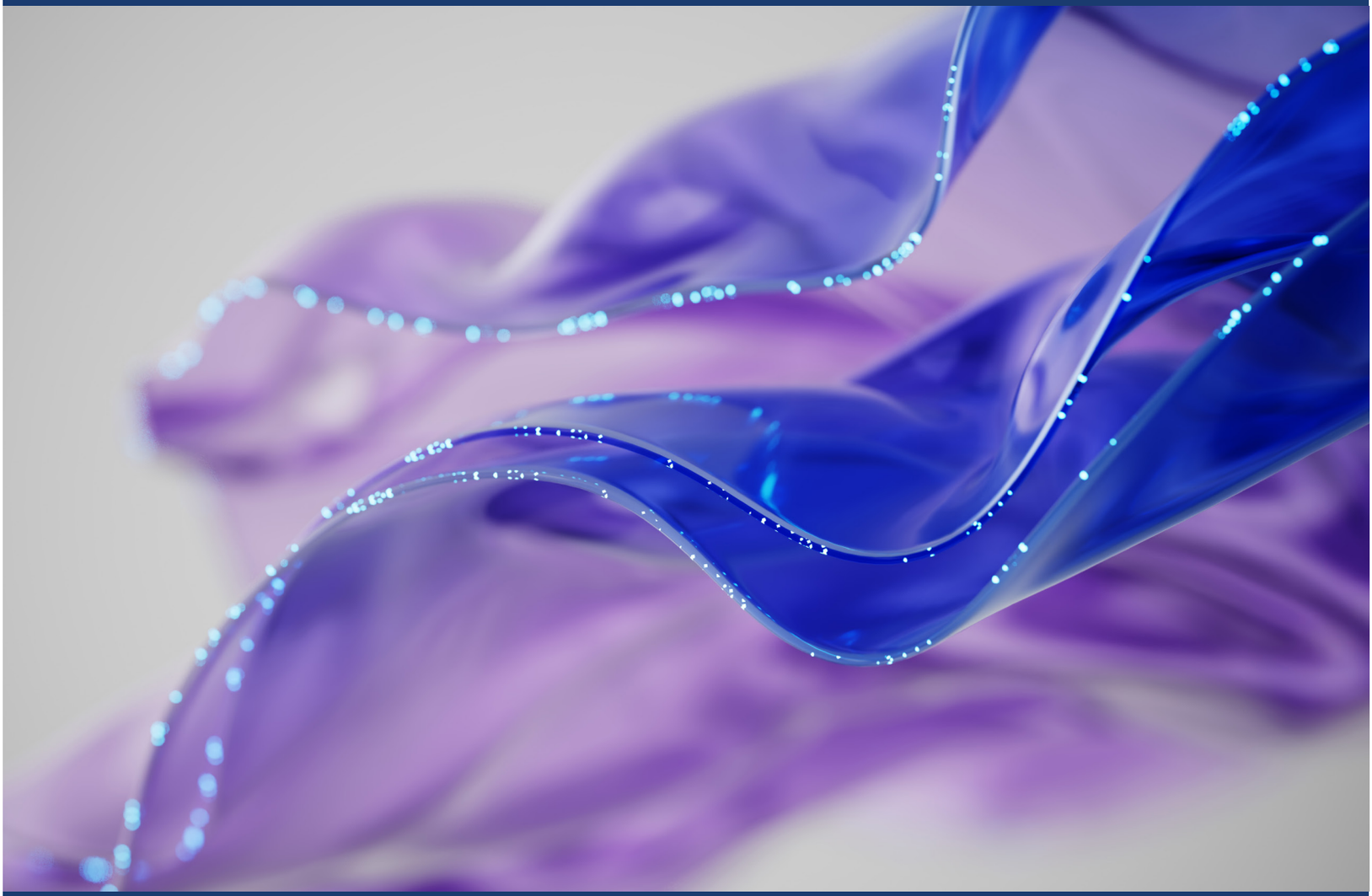
Ассоциация по развитию искусственного интеллекта

ПРЕЗИДЕНТСКАЯ КОМИССИЯ АССОЦИАЦИИ ПО РАЗВИТИЮ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА (ASSOCIATION FOR THE ADVANCEMENT OF ARTIFICIAL INTELLIGENCE, AAAI)

Будущее исследований в области ИИ



Дата публикации: март 2025 г.

Ассоциация по развитию искусственного интеллекта

ПРЕЗИДЕНТСКАЯ КОМИССИЯ АССОЦИАЦИИ ПО РАЗВИТИЮ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА (ASSOCIATION FOR THE ADVANCEMENT OF ARTIFICIAL INTELLIGENCE, AAAI)

Будущее исследований в области ИИ

Дата публикации: март 2025 г.

Содержание

Введение

По мере стремительного развития искусственного интеллекта исследования в этой области меняются так же быстро и комплексно — трансформация затрагивает темы, методы, исследовательское сообщество и рабочую обстановку. Такие темы, как «Рассуждения ИИ» и «Агентный ИИ» уже несколько десятилетий являются предметом исследований. В настоящее время они привлекают еще больше внимания, учитывая текущие достижения и ограничения искусственного интеллекта. Вопросы, касающиеся этики и безопасности ИИ, устойчивого ИИ и возможностей применения ИИ для общественного блага, стали главными темами на всех крупных конференциях. Более того, исследования в области ИИ-алгоритмов и программных систем становятся все более зависимыми от мощного оборудования, особенно графических процессоров. В результате исследователи стали чаще, чем за последние 30 лет, прибегать к совместной разработке ИИ-архитектур. В связи с этими переменами многие исследователи в области ИИ работают в корпоративной среде, где необходимое оборудование и другие ресурсы более доступны, чем в научных учреждениях, что ставит под сомнение важность академических исследований искусственного интеллекта, удержания студентов и привлечения научных специалистов.

Повсеместное распространение ИИ в повседневной жизни и его влияние на людей, общество и окружающую среду делает искусственный интеллект социально-технической темой для научных работ, поэтому исследователям ИИ становится важнее взаимодействовать с экспертами из других сфер — психологами, социологами, философами и экономистами. На первый план выходят неожиданные реакции ИИ, а не те свойства, которые были заложены при разработке и одобрении этих систем. В связи с этим важность принципиальной эмпирической оценки становится как никогда актуальной. Таким образом, возникает необходимость в продуманных контрольных показателях, методиках тестирования и надежных процедурах для получения достоверных выводов из результатов вычислительных экспериментов. Стремительно растущее количество публикаций по теме ИИ и темпы инноваций в этой сфере испытывают на прочность устойчивость системы экспертной оценки, поскольку во многих областях ИИ-исследований все чаще встречается немедленный выпуск статей без рецензирования. Достижения в сфере ИИ становятся все более популярной темой для традиционных и онлайн-СМИ, которые нередко выпускают противоречивые заявления, что запутывает читателей и стирает грань между восприятием возможностей ИИ и реальностью. Все это происходит в геополитической среде, в которой компании и страны конкурируют между собой и стремятся возглавить гонку за лидерство в сфере ИИ. Такое соперничество может повлиять на доступность результатов и инфраструктуры исследований, а также на меры по глобальному регулированию, что лишний раз подчеркивает необходимость в международном взаимодействии по исследованиям и инновациям, связанным с ИИ.

В такой сложной, разнонаправленной и динамичной обстановке важно уметь четко и структурированно задать траекторию ИИ-исследований. Подобные усилия помогают определить текущие тенденции и проблемы, которые предстоит преодолеть, чтобы сделать ИИ более эффективным и надежным для безопасного использования в быту и, что еще важнее, для решения задач первостепенной важности.

Настоящее исследование охватывает 17 тем, связанных с изучением ИИ, которые помогут ответить на все упомянутые выше вопросы. Каждая глава исследования посвящена одной из этих тем и освещает ее историю, текущие тенденции и актуальные проблемы.

Для проведения этого исследования мы привлекли группу опытных исследователей ИИ, которые великодушно приняли наше приглашение и посвятили этой работе много времени. С лета 2024 года до весны 2025 года мы усердно работали, чтобы структурировать исследование, определить главные темы, обсудить содержание, предоставить обратную связь и дополнить те или иные главы.

Кроме того, в написании некоторых глав участвовали дополнительные соавторы, которые поделились своим опытом по конкретной теме. Главным образом работа велась в сети и включала в себя ежемесячные онлайн-конференции с приглашением всех членов группы и дополнительные звонки командам экспертов, работающих над каждой главой. Кроме того, в январе 2025 года была проведена однодневная очная встреча участников.

Чтобы учесть мнение всего AAAI-сообщества, мы запустили подробный опрос по темам исследования, в котором приняли участие 475 респондентов, около 20 % из которых — студенты. Основной группой респондентов стали исследователи (67 %), а на третьем месте по количеству опрошенных стали представители компаний (19 %). Если говорить о географическом охвате, наиболее представленными регионами стали Северная Америка (53 %), Азия (20 %) и Европа (19 %). Хотя большинство респондентов назвало ИИ одним из главных направлений своей работы, были упомянуты и другие сферы, например нейробиология, медицина, биология, социология, философия, политология и экономика. Участие респондентов из разных сфер отразилось на заинтересованности в мультидисциплинарном исследовании, которую выразили 95 % опрошенных.

Каждая глава отчета включает в себя краткую сводку ответов на вопросы, связанные с соответствующей темой.

Работа в рамках этого исследования стала возможной и велась при поддержке Мередит Эллисон (Meredith Ellison), исполнительного директора AAAI, и сотрудников AAAI, которые подготовили и провели опрос.

Надеюсь, этот отчет окажется полезным для всего сообщества по исследованию ИИ. Стоит отметить, что отчет составлен доступным языком, что делает его понятным для самых разных аудиторий: экспертов из других областей, разработчиков регулирующих политик, сотрудников финансирующих организаций и широкой общественности. Нам всем необходимо объединить усилия для ответственного развития ИИ, чтобы технологический прогресс способствовал улучшению жизни людей и соответствовал нашим ценностям.

Франческа Росси (Francesca Rossi)
Президент AAAI, 2022–2025 гг.

Результаты работы комиссии – это мнения ее членов, которые не отражают точку зрения институтов или компаний, где они работают.

Члены комиссии и дополнительные соавторы

Члены комиссии

Франческа Росси (Francesca Rossi), IBM Research

Кристин Бессьер (Christian Bessiere), Университет Монпелье

Джойдип Бисвас (Joydeep Biswas), Техасский университет в Остине

Родни Брукс (Rodney Brooks), Массачусетский технологический институт

Винсент Конитцер (Vincent Conitzer), Университет Карнеги – Меллона

Томас Г. Дитрих (Thomas G. Dietterich), Государственный университет штата Орегон

Вирджиния Дигнум (Virginia Dignum), Университет Умео

Орен Эциони (Oren Etzioni), Вашингтонский университет

Кеннет Д. Форбус (Kenneth D. Forbus), Северо-Западный университет

Юджин Фройдер (Eugene Freuder), Ирландский национальный университет, Корк

Иоланда Гил (Yolanda Gil), Университет Южной Калифорнии

Хольгер Хус (Holger Hoos), Рейнско-Вестфальский технический университет Ахена, Германия, и Лейденский университет, Нидерланды

Эрик Хорвиц (Eric Horvitz), Microsoft

Суббарао Камбхампати (Subbarao Kambhampati), Университет штата Аризона

Генри Каутц (Henry Kautz), Университет Вирджинии

Джийе Ким (Jihie Kim), Университет Донгук

Хироаки Китано (Hiroaki Kitano), Sony Research

Алан Макуорт (Alan Mackworth), Университет Британской Колумбии

Карен Майерс (Karen Myers), SRI International

Люк Де Рэдт (Luc De Raedt), Лёвенский католический университет и Университет Эребру

Стюарт Рассел (Stuart Russell), Калифорнийский университет, Беркли

Барт Сельман (Bart Selman), Корнеллский университет

Питер Стоун (Peter Stone), Техасский университет в Остине и Sony AI

Миллинд Тамбе (Millind Tambe), Гарвардский университет

Майкл Вулдридж (Michael Wooldridge), Оксфордский университет

Дополнительные соавторы

Адитья Акелла (Aditya Akella), Техасский университет в Остине
Глава: аппаратное обеспечение и ИИ

Йошуа Бенджио (Yoshua Bengio), MILA
Глава: сильный искусственный интеллект (AGI)

Абеба Бирхан (Abeba Birhane), Тринити-колледж, Дублин
Глава: сторонние исследования ИИ

Билл Далли (Bill Dally), NVIDIA
Глава: аппаратное обеспечение и ИИ

Фей Фанг (Fei Fang), Университет Карнеги — Меллона
Глава: ИИ для общественного блага

Джонатан Грэтч (Jonathan Gratch), Университет Южной Калифорнии
Глава: ИИ и когнитивистика

Норм Джуппи (Norm Jouppi), Google
Глава: аппаратное обеспечение и ИИ

Джон Е. Лэрд (John E. Laird), Мичиганский университет
Глава: ИИ и когнитивистика

Эми Луэрс (Amy Luers), Microsoft
Глава: ИИ и устойчивое развитие

Питер Норвиг (Peter Norvig), Google
Глава: сильный искусственный интеллект (AGI)

Бесмира Нуши (Besmira Nushi), Microsoft Research
Глава: сильный искусственный интеллект (AGI)

Балараман Равиндран (Balaraman Ravindran), Индийский технологический институт Мадраса
Глава: ИИ для общественного блага

Йоав Шоам (Yoav Shoham), Стэнфордский университет
Глава: ИИ-агенты

Карлес Сьерра (Carles Sierra), Национальный научно-исследовательский совет Испании
Глава: ИИ-агенты

Прадип Варакантам (Pradeep Varakantham), Сингапурский университет менеджмента
Глава: ИИ для общественного блага

Механизмы рассуждения в ИИ

Способность рассуждать – выдающееся свойство человеческого интеллекта, и системы на базе ИИ тоже должны обладать этим умением.

Основные выводы

- Рассуждение всегда считалось ключевой характеристикой человеческого интеллекта. Этот процесс позволяет получить новую информацию из имеющихся базовых знаний; корректность новой информации зависит от правильных рассуждений – без них сведения могут быть не более, чем правдоподобными.
- Исследования в области ИИ привели к появлению целого ряда технологий автоматизированных рассуждений. Эти технологии легли в основу ИИ-алгоритмов и систем, включая решатели задач выполнимости булевых формул (Boolean Satisfiability Problem, SAT), решатели задач выполнимости формул в теориях (Satisfiability Modulo Theories, SMT) и решатели задач удовлетворения ограничений, а также вероятностных графических моделей, которые играют ключевую роль в реализации критически важных сценариев.
- Хотя большие предварительно обученные системы (такие как LLM) далеко продвинулись в своих способностях к рассуждениям, для их гарантированной правильности и глубины требуется больше исследований; такие гарантии особенно важны для автономно работающих ИИ-агентов.

ПРЕДСЕДАТЕЛЬСТВУЮЩИЕ

Кристин Бессьер (Christian Bessiere), Университет Монпелье

Хольгер Хус (Holger Hoos), Рейнско-Вестфальский технический университет Ахена, Германия, и Лейденский университет, Нидерланды

Суббарао Камбхампати (Subbarao Kambhampati), Университет штата Аризона

Контекст и история

Рассуждения – это важная составляющая человеческого интеллекта. Со времен зарождения человечества абдуктивные рассуждения помогали прогнозировать опасность, а индуктивные позволяли изучать существующие в мире закономерности. Методики дедуктивных рассуждений появились в Древней Греции, помогая людям делать достоверные выводы, логически проистекающие из фактов. Развитие методик рассуждения с независимыми от опыта гарантиями стало ключевым фактором в появлении современной науки, математики и инженерии.
Более того, по мнению таких философов, как Чарльз Сандерс Пирс, взаимозависимость между абдукцией, дедукцией и индукцией формирует основу научной методологии, а значит, и всей современной науки. Еще в XIII веке философ Раймунд Луллий предпринимал попытки механизировать логические рассуждения, которые можно отследить в концепции вычислений. Вероятностные рассуждения и заключения в значительной степени повлияли на методики рассуждений, которые нередко полагаются на известную теорему Томаса Байеса об обратной вероятности, лежащей в основе многих подходов к машинному обучению и статистике.
Наконец, оценка корректного (обоснованного) рассуждения является ключевой составляющей большинства методик количественного анализа когнитивных способностей человека.

Неудивительно, что рассуждения играют важную роль в ИИ-системах. Действительно, уже в самых ранних исследованиях ИИ – начиная с Logic Theorist и далее [1] – уделялось большое внимание рассуждениям [2]. С 1960-х годов в развитии ИИ важное место занимали вероятностные рассуждения и модели, особенно для медицинской диагностики [3]. С тех пор в различных исследованиях ИИ было рассмотрено множество задач, связанных с рассуждениями, – от планирования и темпоральных рассуждений до диагностики и объяснения. В то время как в ранних работах об ИИ внимание уделялось правдоподобным рассуждениям (по прецедентам, по аналогии, качественным) и обоснованным формальным рассуждениям с гарантиями (логическим, вероятностным, на основе ограничений), со временем фокус сместился в сторону рассуждений с формальными гарантиями. Для этого есть веские причины. При создании систем ИИ и технологий, которые могут компенсировать ограничения и слабости человека, важно учитывать этот аспект. Ведь людям порой бывает трудно аргументировать свои выводы. В результате появились практичные сценарии применения ИИ-систем, таких как например SAT, SMT и решатели задач удовлетворения ограничений, которые выполняют проверку свойств корректности аппаратного и программного обеспечения компьютера, безопасность протоколов связи, создание новых белков и надежность нейронных сетей при противодействии атакам злоумышленников. Кроме того, появились вероятностные графические модели [4, 5] – мощные инструменты для моделирования и получения выводов, которые могут использоваться в разных сценариях рассуждений в медицине, робототехнике и других областях.

Текущая ситуация и тенденции

Развитие интернета и сопутствующих технологий, позволяющих отслеживать цифровой след человека, а также стремительный рост вычислительных мощностей создали возможности для использования передовых подходов к изучению данных. Среди них можно отметить крупные предварительно обученные модели, например LLM, которые отлично себя показали в правдоподобных рассуждениях. В отличие от более ранних исследований о рассуждениях в ИИ, LLM фокусируются на алгоритмах достоверных рассуждений, которые автоматически формируются в результате крупномасштабного обучения на петабайтах информации. Хотя результаты уже можно назвать впечатляющими, рассуждения в этом контексте считаются вполне «правдоподобными», однако не дают гарантий.

Тем временем, методики обоснованных формальных рассуждений остаются ключевым условием для критических важных и полезных вариантов применения передовых ИИ-технологий для проверки аппаратного и программного обеспечения компьютера, а также для решения реальных задач планирования и распределения ресурсов. Также их все чаще называют основой для формальной проверки методик машинного обучения, таких как нейронные сети, например в контексте локальной надежности перед лицом атак злоумышленников [6]. В этих направлениях ведутся серьезные исследования, которые нацелены на совершенствование различных типов алгоритмов рассуждений (в частности, в отношении их вычислительной сложности), используют обучение для обоснованных формальных рассуждений и совмещают методики рассуждения и обучения [7, 8].

Исследовательские задачи

Применение тщательных априорных или ретроспективных гарантий в алгоритмах правдоподобных рассуждений, оптимизированных предварительно обученными моделями, стало активной и многообещающей областью исследований – особенно там, где ИИ-системы должны работать автономно в доменах, требующих высокого уровня безопасности. Чтобы решить эти проблемы, изучаются так называемые «большие модели рассуждений», а также нейро-символические подходы.

Более того, хотя формальное рассуждение с гарантиями правильности сейчас распространено гораздо меньше, чем использование методик генеративного ИИ для правдоподобных рассуждений, в этой области по-прежнему

есть серьезные проблемы. В таком контексте для достижения экономически и общественно значимых результатов необходимо совместить методики машинного обучения с формальными рассуждениями, особенно в области безопасности и прозрачности ИИ.

Перед нами стоит множество вопросов и задач, начиная с философских:

- что такое «рассуждение»?

И заканчивая практическими:

- можно ли доверять «рассуждению» LLM?

Наряду со следующими.

- Как будет развиваться и какую роль будет играть символическое рассуждение?
- В какой степени LLM или другие генеративные модели способны воспроизвести или заменить символическое рассуждение?
- В какой степени символическое рассуждение будет необходимым или достаточным для преодоления текущих ограничений LLM?
- Насколько хорошо можно объяснить и понять рассуждения ИИ, особенно на базе LLM?
- Каким образом компьютеры могли бы лучше понимать и моделировать человеческие рассуждения?
- Какую роль играет совместное рассуждение людей и компьютеров?
- Как лучше всего применить LLM и символическое рассуждение в «нейросимволическом рассуждении»?
- Нужно ли совершить новые открытия, помимо LLM и традиционных методов символического мышления, чтобы достичь уровня рассуждений, соответствующего сильному ИИ?
- Какие формы рассуждений способны помочь человеку при решении различных задач, например в медицине, науке, машиностроении и юриспруденции?

1. Ньюэлл, А. и Саймон, Х. (Newell, A. & Simon, H.) (1956 г.). Машина логической теории: система обработки комплексной информации. (The logic theory machine: A complex information processing system). IRE Transactions on Information Theory 2: 61-79.
2. Брахман, Р. и Левеск, Х. (Brachman, R. and Levesque, H.) (2004 г.) Knowledge Representation and Reasoning (1-е издание). Morgan Kaufman.
3. Рассел, С. Дж. и Норвиг, П. (Russell, S. J., & Norvig, P.) (2016 г.). Искусственный интеллект: современный подход (Artificial intelligence: a modern approach) (4-е издание). Pearson.
4. Перл, Дж. (Pearl, J.) (1988 г.). Вероятностные рассуждения в интеллектуальных системах: сети правдоподобных заключений (Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference). Morgan Kaufman.
5. Коллер, Д. и Фридманн, Н. (Koller, D. and Friedmann, N.) (2009 г.). Вероятностные графические модели (Probabilistic Graphical Models). The MIT Press.
6. Кёниг, М. (König, M.) и др. (2024 г.) Критическая оценка передовых достижений в области нейросетевой проверки (Critically Assessing the State of the Art in Neural Network Verification). Journal of Machine Learning Research 25(12): 1-53
7. Гуо, Д. и др. (Guo, D. et. al.) (2025 г.) DeepSeek-R: Стимулирование возможностей рассуждений в LLM посредством обучения с подкреплением (DeepSeek-R: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning). https://arxiv.org/abs/2501.12948
8. Камбхампати, С (Kambhampati, S.) (2024 г.). Способны ли большие языковые модели рассуждать и планировать? (Can Large Language Models Reason and Plan?) Annals of New York Academy of Sciences. Март 2024 г.

Мнение сообщества

Участники AAAI-сообщества полностью согласны с тем, что рассуждения в ИИ-системах играют важную роль. В рамках опроса нашего сообщества немногим более 55 % респондентов решили ответить на вопросы по теме рассуждений. Из них 79 % указали, что тема рассуждений актуальна для их исследований (44,7 % назвали ее «очень актуальной»). Из свойств, позволяющих признать процесс рассуждением, 77,5 % опрошенных выбрали «Знание может быть применено», 72,5 % «Могут быть предоставлены объяснения», 56,9 % – «Требуется несколько шагов для получения вывода».
Что интересно, 37,4 % респондентов выбрали вариант «Гарантированная корректность результатов заключений» и лишь 23,7 % назвали «Используется формальная система и решатель задач», что отражает внимание последних лет к неформальным, правдоподобным рассуждениями, вероятно, в контексте методик генеративного ИИ. Это означает, что для демонстрации важности и эффективности формальных, обоснованных рассуждений могут понадобиться усилия. Наконец, 44,7 % респондентов согласились, что «Рассуждения включают в себя процесс поиска».
Многие участники опроса согласились с тем, что использовать человеческий подход к рассуждениям для исследования рассуждений на базе ИИ – это полезно (41,6 %) или даже необходимо (47 %); аналогично, 49,6 % респондентов назвали упор на узконаправленные отраслевые алгоритмы рассуждений полезным, а 42,8 % – необходимым. Это наглядно отражает важное значение, которое придают рассуждениям авторы исследований.
Сообщество также видит потенциал объединения логических и вероятностных моделей рассуждения, которые были разработаны в сфере ИИ до больших предварительно обученных моделей. Это четко видно по тому, что 76,9 % участников опроса назвали интеграцию подходов обучения и рассуждения очень важной (6 или 7 по 7-балльной шкале); что интересно, доля респондентов, считающих возможность объяснения и проверки очень важной, также была достаточно высокой (71,7 %).
Наконец, 61,8 % респондентов считают, что минимальная доля символических ИИ-технологий, необходимая для достижения человеческого уровня рассуждения, составляет не менее 50 % (24,8 % назвали 75 % или больше, а 38,2 % опрошенных назвали минимальную долю 25 % или меньше). Неясным остается то, насколько исследователи и специалисты в области ИИ понимают, что для выдающихся и успешных реализаций методик формальных рассуждений ИИ на благо научных и математических достижений и технических применений, а также для обеспечения безопасности в области ИИ требуются явно сверхчеловеческие уровни рассуждения.

Фактическая точность и надежность

ПРЕДСЕДАТЕЛЬСТВУЮЩИЙ

Генри Каутц (Henry Kautz),
Университет Вирджинии

На сегодняшний день ключевая задача исследований в области искусственного интеллекта – повышение фактической точности и надежности ИИ-систем. Хотя в этой области наблюдается значительный прогресс, многие ученые сомневаются, что эти проблемы будут решены в ближайшем будущем.

Основные выводы

- ИИ-система считается точной, если не выдает ложные утверждения. Повышение фактической точности ИИ-систем на базе нейросетевых больших языковых моделей, возможно, является главным направлением современных исследований в сфере ИИ.
- Надежность рассуждений можно оценить по таким критериям, как доступность для человеческого понимания, устойчивость к ошибкам и применение человеческих ценностей. ИИ-системы, не обладающие достаточной надежностью, не допускаются к развертыванию в критически важных сценариях.
- Для повышения фактической точности и надежности ИИ-систем используются такие подходы, как доработка, генерация ответа, дополненная результатами поиска, проверка результатов, выданных машиной, а также замена сложных моделей на простые и более понятные.

Контекст и история

Фактически точная ИИ-система не выдает ошибочную информацию и не создает ложную. До эры генеративного ИИ проблемы с фактической точностью возникали, когда системы обучались на некачественных данных, что можно описать фразой «что посеешь, то и пожнешь». Работа над повышением качества данных ведется в сфере ИИ достаточно давно [1].

Системы на базе генеративного ИИ и, в частности, большие языковые модели, используют реконструктивную память – это означает, что они восстанавливают воспоминания по необходимости на основе распределенных битов информации, а не извлекают их из фиксированного хранилища. Ситуация изменилась с появлением ранних генеративных LLM, которые могли генерировать связные, но полностью вымышленные истории [2]. Фактическую точность LLM по определенной теме можно было улучшить путем доработки модели на основе узконаправленных данных [3].

Надежность – это более широкое понятие, чем фактическая точность, поскольку она отличается доступностью для человеческого понимания, устойчивостью к ошибкам и применением человеческих ценностей. Традиционно для улучшения понятности ИИ-систем сложные модели типа «черный ящик» заменяются на простые и доступные для человека, такие как наивный байесовский классификатор [4] или обобщенная линейная регрессия [5]. Проводятся исследования устойчивости машинного обучения, направленные на понимание того, как изменяются выходные данные модели при небольших изменениях в обучающих данных. Например, контрастное обучение – это метод обучения глубоких нейросетей с повышенной устойчивостью [6]. Подробнее устойчивость генеративного ИИ обсуждается в разделе «Рассуждения» этого отчета. Вопрос уважения человеческих ценностейсистемами ИИ упоминается во многих разделах этого отчета и здесь обсуждаться не будет.

Текущая ситуация и тенденции

Как упоминалось, дообучение остается главным подходом, с помощью которого ученные и инженеры повышают фактическую точность генеративного ИИ. Помимо дообучения на документах, относящихся к определенной предметной области, современное дообучение включает в себя обучение с подкреплением при помощи обратной связи от тысяч людей. Расходы на привлечение такого числа оценщиков являются основным препятствием для масштабирования ИИ-систем, поэтому активно исследуются методы, позволяющие сократить объем необходимой обратной связи от людей [7].

Второй по значимости метод повышения фактической точности генеративного ИИ – генерация с дополненной выборкой (Retrieval Augmented Generation, RAG). В ответ на вопрос система собирает несколько подходящих документов с помощью традиционных алгоритмов поиска информации. Затем ИИ получает промпт с просьбой сгенерировать ответ на основе комбинирования и обобщения найденных документов. RAG действительно помогает повысить фактическую точность, но результат зависит от качества найденных данных. Например, если в целевой набор документов входит весь интернет, система может включить в ответ некорректную информацию и даже анекдоты.

Смежный подход – попросить генеративный ИИ использовать инструменты для проверки фактов. Это могут быть калькуляторы, базы данных с фактами, такие как индексы цитирования, а также системы формального планирования и рассуждений [9]. Новый подход к повышению фактической точности заключается в предоставлении системе набора правил, которые задают ограничения для пространства ответов. Полученные результаты сравниваются с этими правилами, и несоответствующие ответы отсеиваются [10]. Amazon Web Services уже поддерживает этот подход с помощью «автоматических проверок рассуждений» [11].

Третий метод повышения фактической точности генеративного ИИ – цепочка рассуждений (Chain of Thought, CoT), в которой серия промптов разбивает вопрос на менее крупные единицы [12]. CoT часто включает в себя этапы, на которых модель должна пересмотреть свои предварительные выводы и проверить их на наличие галлюцинаций. Подробнее CoT рассматривается в разделе «Рассуждения» в этом отчете.

Влияние качества данных на фактическую точность уже упоминалось. Дообучение можно проводить без участия людей на специально созданном наборе данных гарантированно высокого качества, и с недавних пор ведутся исследования в этом направлении [13].

Надежность, как упоминалось, включает в себя фактическую точность, а также понятность и устойчивость. Один из вариантов повысить понятность моделей нейросетей – разложить их на набор распознавателей обобщенных признаков, а затем объединить эти признаки с помощью понятной модели, например аддитивной регрессии [15]. Еще один подход – тщательно разобраться, как концепции и правила на самом деле представлены в обученной модели [16]. Кроме того, можно улучшить понятность, если с помощью техник CoT попросить генеративный ИИ объяснить этапы своих рассуждений [17] или сообщить пользователю, когда система не уверена в сделанных выводах [18]. Наконец, можно попросить генеративный ИИ выдать не один готовый ответ, а преобразовать сложный набор данных в простую и понятную схему, например, в дерево принятия решений [19].

Исследовательские задачи

Задача обеспечения фактической точности еще далеко не решена. Появляется все больше контрольных

наборов данных для тестирования фактической точности LLM-моделей. Например, SimpleQA от Google — один из новейших наборов понятных, однозначных, актуальных и непростых вопросов и ответов, касающихся фактов [14]. В декабре 2024 года лучшие модели компаний OpenAI и Anthropic правильно отвечали менее чем на половину вопросов.

Как упоминалось выше, устойчивость в генеративном ИИ можно улучшить с помощью функций устойчивых потерь, таких как контрастное обучение. Состязательное обучение, которое применяет возмущения в пространстве эмбеддинга, может улучшить как устойчивость, так и генерализацию [20]. Кроме того, методы повышения фактической точности также обычно улучшают устойчивость.

1. Будах, Лукас, Мориц Фойерпфайл, Нина Иде, Андреа Натансен, Неле Сина Ноак, Хендрик Патцлафф, Хазар Хармуш и Феликс Науманн (Budach, Lukas, Moritz Feuerpfeil, Nina Ihde, Andrea Nathansen, Nele Sina Noack, Hendrik Patzlaff, Hazar Harmouch and Felix Naumann) (2022 г.). Влияние качества данных на эффективность машинного обучения (The Effects of Data Quality on Machine Learning Performance). https://arxiv.org/pdf/2207.14529
2. А. Рэдфорд, Ц. Ву, Р. Чайлд, Д. Луан, Д. Амодей, И. Суцкевер (Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I.) (2019 г.). Языковые модели — это неконтролируемые многозадачные учащиеся (Language models are unsupervised multitask learners). OpenAI. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
3. Дж. Девлин, М.-В. Чан, К. Ли, К. Тутанова (Devlin, J., Chang, M.-W., Lee, K., & Toutanova) (2019 г.). BERT: предварительное обучение глубоких двунаправленных трансформеров для понимания языка (BERT: Pre-training of deep bidirectional transformers for language understanding). Материалы конференции 2019 года Североамериканского подразделения Ассоциации по вычислительной лингвистике: технологии человеческого языка , том 1 (академические работы и краткие статьи) (Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)), 4171–4186. https://doi.org/10.48550/arXiv.1810.04805
4. Педро Доминго с и Майкл Паззани (Pedro Domingos & Michael Pazzani) (1997 г.). Об оптимальности простого байесовского классификатора при нулевых потерях (On the Optimality of the Simple Bayesian Classifier under Zero-One Loss). Machine Learning, 29(2-3), 103–130. https://doi.org/10.1023/A:1007413511361
5. Каруана, Рич, Инь Лу, Йоханнес Герке, Пол Кох, М. Штурм и Ноэми Эльхадад (Caruana, Rich, Yin Lou, Johannes Gehrke, Paul Koch, M. Sturm and Noémie Elhadad) (2015 г.). Интеллектуальные модели для здравоохранения: прогнозирование риска пневмонии и повторной госпитализации в течение 30 дней (Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission). Материалы 21-й Международной конференции по извлечению знаний и добыче данных Специальной группы Ассоциации вычислительной техники (Association for Computing Machinery, Special Interest Group on Knowledge Discovery and Data Mining, ACM SIGKDD) (2015 г.). https://people.dbmi.columbia.edu/noemie/papers/15kdd.pdf
6. Хадселл, Р., Чопра, С., и ЛеКун, И. (Hadsell, R., Chopra, S., & LeCun, Y.) (2006 г.). Снижение размерности путем обучения инвариантному отображению (Dimensionality reduction by learning an invariant mapping). Материалы Международной конференции общества специалистов по вычислительным машинам IEEE по компьютерном зрению и распознаванию форм (CVPR 2006), 2, 1735–1742. https://doi.org/10.1109/CVPR.2006.100
7. Ху, К., Ху, И., Цао, Х., Сяо, Т., и Чжу, Ц. (Hu, C., Hu, Y., Cao, H., Xiao, T., & Zhu, J.) (2024 г.). Обучение языковых моделей самосовершенствованию путем изучения языковой обратной связи (Teaching language models to self-improve by learning from language feedback). Выводы Ассоциации вычислительной лингвистики (Association for Computational Linguistics, ACL, 2024 г.). https://aclanthology.org/2024.findings-acl.364/
8. Льюис, П., Перес, Э., Пиктус, А., Петрони, Ф., Карпухин, В., Гоял, Н., Кюттлер, Х., Льюис, М., Их, В.-Т., Роктэшель, Т., Ридель, С., и Киела, Д. (Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D.) (2020 г.). Генерация с расширением поиска для наукоемких задач по обработке естественного языка (Retrieval-augmented generation for knowledge-intensive NLP tasks). Advances in Neural Information Processing Systems (NeurIPS 2020 г.), 33, 9459–9474. https://arxiv.org/abs/2005.11401
9. Гуань, Л., Вальмекам, К., Сридхаран, С., и Камбхампати, С. (Guan, L., Valmeekam, K., Sreedharan, S., & Kambhampati, S.) (2023 г.). Использование предварительно обученных больших языковых моделей для построения и использования моделей мира при планировании задач на основе моделей (Leveraging pre-trained large language models to construct and utilize world models for model-based task planning). Материалы 33-й Международной конференции по автоматизированному планированию (International Conference on Automated Planning and Scheduling, ICAPS, 2023 г.). https://arxiv.org/abs/2305.14909
10. Бакес, Дж., Болиньяно, П., Кук, Б., Додж, К., Гацек, А., Лаккоу, К., Рунгта, Н., Ткачук, О., и Варминг, К. (Backes, J., Bolignano, P., Cook, B., Dodge, C., Gacek, A., Luckow, K., Rungta, N., Tkachuk, O., & Varming, C.) (2018 г.). Автоматизированное рассуждение на основе семантики для политик доступа AWS с использованием SMT (Semantic-based automated reasoning for AWS access policies using SMT). Материалы конференции «Формальные методы в автоматизированном проектировании» (Formal Methods in Computer-Aided Design, FMCAD), 2018 г. (стр. 1–9). IEEE. https://doi.org/10.23919/FMCAD.2018.8602994
11. Барт, Антье (Barth, Antje) (2024 г.). Предотвращение фактических ошибок из-за галлюцинаций LLM с помощью математически обоснованных автоматизированных проверок логических заключений (Prevent factual errors from LLM hallucinations with mathematically sound Automated Reasoning checks) (предварительный просмотр). Опубликовано 3 декабря 2024 г., получено 8 февраля 2025 г. AWS News Blog, https://aws.amazon.com/blogs/aws/prevent-factual-errors-from-llm-hallucinations-with-mathematically-sound-automated-reasoning-checks-preview
12. Вэй, Ц., Ван, С., Схурманс, Д., Босма, М., Ихтер, Б., Ся, Ф., Чи, Э., Ле, Ц., Чжоу, Д. (Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D.) (2022 г.). Побуждение больших языковых моделей к рассуждениям с помощью промптов в виде цепочки мыслей (Chain-of-thought prompting elicits reasoning in large language models). Advances in Neural Information Processing Systems (т. 35, стр. 24824–24837). https://proceedings.neurips.cc/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf
13. Дин, Бошэн, Чэнвэй Цинь, Жочэнь Чжао, Тяньцзэ Ло, Синьцзэ Ли, Гуйчжэнь Чень, Вэньхань Ся, Цзюньцзе Ху, Ан Туан Луу и Шафик Р. Джоти (Ding, Bosheng, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu and Shafiq R. Joty) (2024 г.). Аугментация данных с использованием LLM: перспективы данных, парадигмы обучения и проблемы (Data Augmentation using LLMs: Data Perspectives, Learning Paradigms and Challenges). Ежегодная встреча Ассоциации вычислительной лингвистики (2024 г.). https://aclanthology.org/2024.findings-acl.97.pdf
14. Вэй, Джейсон, Нгуен Карина, Хён Вон Чунг, Юньсинь Джой Цзяо, Спенсер Папей, Амелия Глез, Джон Шульман, Уильям Федус (Wei, Jason, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, William Fedus) (2024 г.). Оценка способности больших языковых моделей отвечать на короткие вопросы, требующие выяснения фактов (Measuring short-form factuality in large language models). https://doi.org/10.48550/arXiv.2411.04368
15. Агарвал, Р., Мельник, Л., Фрост, Н., Чжан, С., Ленгерих, Б., Каруана, Р., Хинтон, Дж. Э. (Agarwal, R., Melnick, L., Frosst, N., Zhang, X., Lengerich, B., Caruana, R., & Hinton, G. E.) (2021 г.). Нейронные аддитивные модели: интерпретируемое машинное обучение с помощью нейронных сетей (Neural additive models: Interpretable machine learning with neural nets). Advances in Neural Information Processing Systems, 34, 2021 г. https://proceedings.neurips.cc/paper/2021/hash/251bd0442dfcc53b5a761e050f8022b8-Abstract.html?utm_source=chatgpt.com
16. Темплтон, А., Коунерли, Т., Маркус, Дж., Линдси, Дж., Брикен, Т., Чэнь, Б., Пирс, А., Ситро, К., Амейзен, Э., Джонс, А., Каннингем, Х., Тернер, Н. Л., Макдугалл, К., МакДиармид, М., Тамкин, А., Дурмус, Э., Хьюм, Т., Москони, Ф., Фриман, К. Д., Саммерс, Т. Р., Рис, Э., Бэтсон, Дж., Джермин, А., Картер, С., Олах, К., Хениган, Т. (Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Tamkin, A., Durmus, E., Hume, T., Mosconi, F., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., & Henighan, T.) (2024 г.). Масштабирование моносемантичности: извлечение интерпретируемых признаков из Claude 3 Sonnet (Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet). Transformer Circuits Thread. https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html
17. Йео, В. Ц., Нг, Х. Х., Ле, Т. К. К., Лу, С. (Yeo, W. J., Ng, X. X., Le, T. K. C., & Lu, X.) Насколько интерпретируемыми являются рассуждения больших языковых моделей? (How interpretable are reasoning explanations from prompting large language models?) Выводы Ассоциации вычислительной лингвистики (Association for Computational Linguistics): NAACL, 2024 г. https://aclanthology.org/2024.findings-naacl.138
18. Лин, С., Хилтон, Дж., Эванс, О. (Lin, S., Hilton, J., & Evans, O.) (2022 г.). Обучение моделей выражению неуверенности словами (Teaching Models to Express Their Uncertainty in Words). Transactions on Machine Learning Research. https://openreview.net/pdf?id=8s8K2UZGTZ
19. Чэнь, Ю., Чжан, Л., Ван, Х., Ли, Ц. (Chen, Y., Zhang, L., Wang, H., & Li, J.) (2025 г.). Построение дерева решений с нуля с помощью больших языковых моделей (Zero-Shot Decision Tree Construction via Large Language Models). arXiv preprint arXiv:2501.16247. https://arxiv.org/abs/2501.16247
20. Лю, С., Чэн, Х., Хэ, П., Чэнь, В., Ван, Ю., Пун, Х., Гао, Ц. (Liu, X., Cheng, H., He, P., Chen, W., Wang, Y., Poon, H., & Gao, J.) (2020 г.). Состязательное обучение больших нейронных языковых моделей (Adversarial training for large neural language models). arXiv preprint arXiv:2004.08994. https://arxiv.org/abs/2004.08994

Мнение сообщества

Более 75 % членов Ассоциации по развитию искусственного интеллекта (AAAI) считают, что фактическая точность и надежность имеют большое значение или очень большое значение для их исследований.

Опрошенные поддержали все шесть предложенных подходов к повышению фактической точности: сторонние инструменты для проверки на достоверность, подкрепление, повышение качества данных, курирование данных, обучение на синтетических данных и новые архитектуры нейросетей. Больше всего участников отмечало, что им требуется больше исследований о новых архитектурах нейросетей (73 % ответили, что это важно или чрезвычайно важно), а также больше внешних инструментов для проверки на достоверность (70 %).

Среди факторов повышения надежности на первом месте также стоят новые архитектуры (77 % отметили, что это важно или очень важно), на втором – возможность прослеживать процессы рассуждения моделей (70 %), а на третьем – использование понятных моделей вместо нейросетей (61 %). Интересно отметить, что лишь 24 % опрошенных считают, что ИИ-системы будут вызывать больше доверия, если наделить их человеческими чертами. Наконец, сообщество, в основном (59 %), согласно, что ИИ-моделям не хватает надежности в ее нынешнем определении. Большинство (около 60 %) не считает, что фактическая точность и надежность вскоре будут достигнуты.

AAAI-сообщество предложило ряд дополнительных аспектов фактической точности и надежности, которые не были упомянуты выше. Например:

- Возможность понимать и представлять разные стороны одного и того же вопроса, включая плюсы и минусы.
- Понимание того, что надежность зависит от контекста предметной области, целей организации и целей пользователя. Невозможно назвать ИИ-систему надежной или ненадежной вне контекста.
- Необходимость в прозрачности не только для моделей, но и для фактических источников обучающих данных. Сюда относится проверка фактов, получаемых моделями из разных источников.
- Акцент на рисках и способах их снижения, а не на решении проблем фактической точности и надежности.
- Предоставление ИИ-агентам возможности пополнять свои знания без ущерба для надежности.

ИИ-агенты

Агенты и мультиагентные системы (Multi-Agent Systems, MAS) теперь не просто самостоятельно решают поставленные задачи, но и интегрируются с генеративным ИИ и LLM-моделями, что приводит к созданию фреймворков кооперативного ИИ, улучшающих гибкость, масштабируемость и возможности взаимодействия.

Основные выводы

- Мультиагентные системы эволюционировали от автономных инструментов на основе правил до кооперативного ИИ, в котором на первом плане взаимодействие, согласование и этические принципы.
- Развитие агентного ИИ с использованием LLM расширяет возможности гибкого принятия решений, но может снижать эффективность и простоту систем.
- Интеграция кооперативного ИИ с генеративными моделями требует баланса между адаптивностью, прозрачностью и вычислительными возможностями в мультиагентных средах.

ПРЕДСЕДАТЕЛЬСТВУЮЩИЕ

Вирджиния Дигнум (Virginia Dignum), Университет Умео

Майкл Вулдридж (Michael Wooldridge), Оксфордский университет

Контекст и история

Сфера мультиагентных систем возникла еще в конце 1980-х / начале 1990-х гг. под влиянием двух разных областей [1,2]. В одной из них – ИИ-робототехнике – ученые серьезно занимались вопросами создания интегрированных агентских архитектур: как объединить несколько когнитивных компонентов разума (планирование, рассуждение, обучение, видение и т. д.) в рамках единого вычислительного агента? Второй была только что появившаяся область распределенного ИИ, где изучались возможности совместного достижения результатов ИИ-системами, динамически обменивающимися информацией и задачами. К середине 1990-х эти идеи привели к возникновению новой области, в которой специалисты занимались созданием (полу)автономных ИИ-систем – агентов, работающих от имени пользователей для достижения их целей, в том числе путем взаимодействия с другими подобными агентами. Было очевидно, что, поскольку делегированные цели могут противоречить друг другу, у таких агентов должна быть способность социального рассуждения. Таким образом, хотя исторически эволюция ИИ была направлена на такие когнитивные способности, как рассуждения и решение задач, в новой области мультиагентных систем акцент делался на развитии социальных навыков, таких как сотрудничество, координация, аргументация и согласование. Для формирования этих навыков главной задачей области стало создание моделей психического состояния человека (Theory of Mind).

К концу 1990-х годов на эту тему проводились конференции и выпускался журнал, и это направление стало одним из ключевых ответвлений области изучения ИИ. Эта сфера начала расцветать с конца 1990-х годов. Проводились активные исследования, связанные с языками коммуникации автономных агентов, протоколами сотрудничества, координации и согласования, а также с базовой теорией этих социальных навыков.

Что касается последнего пункта, – если поначалу на теорию мультиагентных систем решающее влияние оказывала парадигма практического рассуждения ИИ-планирования, с 2000-х годов преобладающей теоретической основой стала теория игр. Теория игр, возникшая из экономических исследований, описывает взаимодействие между агентами, движимыми своими интересами. Хотя изначально она была разработана как инструмент изучения взаимодействия между людьми и группами людей, казалось, что это подходящий фундамент для изучения взаимодействия между ИИ-агентами. Проводилось множество исследований на эту тему, например: аукционы как способ распределения ограниченных ресурсов, теория переговоров между ИИ-агентами с собственными интересами, оптимальное группирование агентов для совместного решения задач на взаимовыгодной основе. Интересно отметить, что хотя обучение в мультиагентных системах поначалу являлось ключевым компонентом, в первое десятилетие исследований в этой области оно не стояло в центре внимания.

Первый бум мультиагентных систем продлился с середины 1990-х до 2010–2015 годов. К концу этого периода стали возникать неудобные вопросы. Научные исследования в этой области были более чем продуктивными, но сферы практического применения казались ограниченными. Разумеется, можно было выделить несколько важных вариантов использования. Область игр по обеспечению безопасности, возникшая из мультиагентных систем, опиралась на идеи теории игр и имела целью распределять ограниченные ресурсы для защиты критических объектов, таких как аэропорты. В результате этой работы в аэропортах и морских портах США были развернуты соответствующие приложения. Автоматизированные системы высокочастотного трейдинга, которые планируют и выполняют множество торговых операций на мировых рынках, представляют собой мультиагентные системы в глобальном масштабе. Моделирование на основе агентов, работающее с социотехническими системами на уровне отдельных субъектов принятия решений, стало активно развиваться после финансового кризиса 2008 года, а затем после пандемии COVID-19, поскольку показало себя как важный инструмент моделирования распространения вредного влияния: экономического в первом случае, эпидемиологического и социального – во втором. Несмотря на все достижения, пока не удалось создать мультиагентную систему, в которой агенты полноценно функционируют в контексте других агентов с учетом таких социальных концептов, как нормы, организации, практики и ценности, с тем чтобы результаты можно было использовать для формирования политик в разных областях, от здравоохранения до перевозок и урбанистической трансформации. Причем исследования ведутся не только и не столько сообществом участников Международной конференции по автономным агентам и мультиагентным системам (International Conference on Autonomous Agents and Multi-Agent Systems, AAMAS), сколько в сфере социального моделирования.

Пока практическое применение исследований в области мультиагентных систем (в которых ИИ-агенты взаимодействуют между собой) не соответствует первоначальным ожиданиям, но отдельные разговорные агенты, такие как Alexa, Siri и Cortana, уже прочно вошли в нашу жизнь. Они разработаны на основе трудов по созданию интеллектуальных агентов в 1990-х годах, а также работ сообщества специалистов по развитию разговорных систем с применением обработки естественного языка. За последние тридцать лет это направление принесло немало практических результатов: автоматизированные ассистенты в колл-центрах, помощники по обслуживанию клиентов, виртуальные ассистенты в смартфонах, ассистенты в умных колонках, домашние роботы-помощники, которые могут общаться с пользователями и выполнять такие поручения, как бронирование билетов или столиков в ресторане,

онлайн-покупки, медицинская помощь и консультирование по продажам. При этом они используют такие ИИ-модули, как распознавание речи, понимание естественного языка, управление диалогом (отслеживание состояния и политика диалога), а также генерацию текста на естественном языке и синтез речи.

Появление машинного обучения в начале века дало новый толчок исследованиям мультиагентных систем. Мультиагентное обучение с подкреплением (Multi-Agent Reinforcement Learning, MARL) стало самым значимым направлением в этой области, отчасти благодаря тому, что разрабатывать эксперименты MARL можно относительно быстро и без использования дорогостоящего оборудования. На момент написания этого документа MARL представляет собой крупное ответвление в сфере машинного обучения, но пока у него нет ни четкого единого видения, ни направления, ни применения.

Текущая ситуация и тенденции

Появление LLM-моделей в 2020 году привело к росту интереса к агентам [3]. LLM-модели можно использовать как часть рабочего процесса для автоматизации рутинных задач, к тому же широко обсуждаются общие способности таких агентов планировать и решать задачи. В этом контексте концепция агентного ИИ подразумевает интеграцию генеративного ИИ и LLM-моделей для создания фреймворков автономных агентов с целью использовать генеративные способности таких моделей для улучшения взаимодействия, творческого потенциала и эффективности принятия решений в режиме реального временив динамических средах. На момент написания (конец 2024 года) наблюдается взрывной рост числа стартапов, которые надеются заработать на таких агентах. Несмотря на новую волну энтузиазма, до сих пор не достигнуты изначальные цели AAMAS, сформулированные 30 лет назад, такие как создание устойчивых и автономных мультиагентных систем, способных на сложную координацию и долгосрочные рассуждения. Пока не ясно, в какой степени агенты новой волны в своих действиях опираются на прошлый контекст.

Сейчас необходимо понять, что собой представляют мультиагентные системы в эпоху LLM-моделей. Нынешний тренд на агентификацию LLM может привести к созданию излишне сложных и необязательных архитектур, а также к огромным вычислительным затратам, в то время как применение мультиагентной парадигмы к разработке и использованию LLM-моделей может предложить устойчивый способ эффективного формирования, диверсификации и интеграции подходов. Хотя распределение было одним из главных аспектов MAS, в текущей парадигме это направление еще очень мало исследовано. Еще одна современная тенденция — это дополнение принципов классических когнитивных архитектур навыками здравого смысла для автономных агентов.

Сейчас ведутся работы над созданием мультиагентных архитектур, в которых компоненты ИИ делятся на модульные системы с целью повысить прозрачность, адаптивность и этичность. Акцент на кооперативных агентах подчеркивает смещение приоритетов в сторону сотрудничества, согласования и совместного принятия решений. За счет использования принципов модульной организации, инкапсуляции и разделения задач, эти архитектуры поддерживают масштабируемое взаимодействие между автономными агентами и людьми. Это идеальный подход для гибридных вариантов применения ИИ, в которых требуется доверие, объяснимость и экспертные знания в предметной области.

Исследовательские задачи

- Определить сложности и преимущества внедрения агентов на базе генеративного ИИ в MAS с целью улучшить сотрудничество без нарушения текущих динамик.
- Исследовать способы, помогающие агентам на базе LLM-моделей эффективнее договариваться и принимать решения в динамичных мультиагентных средах с соблюдением принципов этики и безопасности.
- Разработать архитектуры, интегрирующие агентов на базе LLM без ущерба для масштабируемости, прозрачности и эффективности вычислений в мультиагентных средах.

1. Йоав Шохам (Yoav Shoham). Агентно-ориентированное программирование (Agent-oriented programming). Artificial Intelligence. Artificial Intelligence. 60 (1): 51–92.
2. Майкл Вулдридж (Michael Wooldridge). Введение в мультиагентные системы (An Introduction to Multi-agent Systems), 2-е изд. Wiley, 2009.
3. Джулия Визингер, Патрик Марлоу и Владимир Вускович (Julia Wiesinger, Patrick Marlow and Vladimir Vuskovic). Агенты (Agents). Доклад Google. https://archive.org/details/google-ai-agents-whitepaper

Мнение сообщества

Большинство участников опроса считают эту тему актуальной для своих исследований и выражают все больше интереса к интеграции больших языковых моделей (LLM) в мультиагентные системы. Многие участники уже используют ИИ-агенты, чаще всего LLM (29,34 %), что указывает на их растущую важность для приложений на базе ИИ.

Мультиагентные системы, использующие LLM-модели, имеют потенциал в таких областях, как совместное решение задач (68,86 %), распределенное принятие решений (54,49 %) и социальное моделирование (41,32 %). Среди проблем опрошенные называют несоответствие между общими знаниями LLM и конкретными потребностями системы (59,88 %), сложности интерпретации (59,28 %) и риски безопасности (50,90 %). Эти опасения указывают на необходимость улучшить объяснимость, согласование стратегий и меры безопасности для эффективного развертывания.

Пока нет единого мнения о необходимости агентификации LLM-моделей: 51,5 % опрошенных считают, что мультиагентные парадигмы LLM крайне важны для устойчивого ИИ, а 42,33 %, напротив, полагают, что это чересчур усложняет системы. Также сомнения вызывают вычислительные затраты, поскольку неясно, смогут ли LLM-модели окупить огромные расходы.

В произвольных ответах об интеграции LLM в MAS также нет единообразия, при этом некоторые выступают за гибридные подходы вместо применения только LLM-моделей. Многие респонденты подчеркивают потребность в разнообразных архитектурах ИИ, особенно в модульных системах, использующих различные технологии, где LLM-модели играют свою роль, но не доминируют. Управление, координация и адаптивность считаются ключевыми преимуществами MAS, а главными недостатками – сложность, нехватка теоретических обоснований и высокие вычислительные затраты.

Некоторые опрошенные критикуют чрезмерную увлеченность LLM-моделями, поскольку неясно, действительно они имеют ценность или просто являются популярным трендом. Другие говорят о практических трудностях, таких как обоснование, согласованность и надежные протоколы коммуникации, указывая на необходимость в новых фреймворках, которые интегрируют символическое рассуждение, структурированное управление и масштабируемые архитектуры. В целом, эта дискуссия отражает критический, но открытый подход к агентификации LLM-моделей, который предполагает, что контекст, область применения и технологическое разнообразие повлияют на их эффективность в мультиагентных средах.

В заключение, результаты опроса демонстрируют оптимизм в отношении MAS на базе LLM, при этом подчеркивая необходимость решить основные проблемы перед их повсеместным распространением.

Оценка ИИ

Оценка ИИ – это процесс анализа производительности, надежности и безопасности ИИ-систем.

ПРЕДСЕДАТЕЛЬСТВУЮЩИЙ

Карен Майерс (Karen Myers), SRI International

Основные выводы

- Процесс оценки ИИ-систем значительно отличается от стандартных методов проверки и валидации программного обеспечения.
- Сейчас оценка выполняется с использованием эталонных тестов, направленных на анализ качества (генеративных) моделей, но недостаточно внимания уделяется остальным важным факторам, таким как удобство использования, прозрачность и соблюдение этических принципов.
- Необходимы новые подходы и методы для оценки ИИ-систем, которые гарантировали бы надежность при масштабном развертывании.

Контекст и история

Последние успехи в области ИИ позволяют найти множество новых применений этой технологии. Однако многие организации не решаются развертывать ИИ-системы, опасаясь за свою репутацию, которая может пострадать из-за галлюцинаций генеративного ИИ, утечки закрытых данных и нарушения юридических и этических норм.

Эмпирические методы уже давно играют важную роль в исследованиях ИИ (например, [1]). Научное сообщество действительно разработало надежный набор метрик и методов для оценки отдельных алгоритмов ИИ, чтобы измерять их производительность и отслеживать прогресс. Гораздо меньше внимания уделяется оценке ИИ-систем и их развертыванию в реальных условиях, в том числе для использования людьми, которые не являются специалистами в этой области.

Процесс оценки ИИ-систем значительно отличается от стандартных методов проверки и валидации программного обеспечения. Универсальные, многогранные и обширные возможности ИИ невозможно протестировать полностью, поэтому требуется новый подход к определению достаточного уровня тестирования. Адаптивность в среде выполнения и эволюция обученных моделей непрерывно меняют поведение системы, поэтому мониторинг и проверки должны выполняться на постоянной основе. Многие ИИ-системы предназначены для интерактивного использования, поэтому важно учитывать особенности совместной работы и влияние на пользователей.

Текущая ситуация и тенденции

В настоящее время для оценки систем генеративного ИИ проводится тестирование моделей на основе растущего набора эталонных тестов. Некоторые тесты направлены на оценку общих возможностей (например, GLUE [2], ARC-AGI [3], MMLU [4]), а другие — на определенные типы рассуждений и знаний (например, MATH для математики [5], GPQA для логики [6], HumanEval для написания кода [7]). Эталонное тестирование позволяет эффективно оценить возможности и недостатки, а также прогресс моделей со временем. Эталонные тесты позволяют оценить возможности ИИ, но они ограничены контекстом, поэтому их не всегда можно применять в новых предметных областях. Более того, эталонное тестирование не позволяет гарантировать успешное развертывание, потому что не проверяет ожидаемые варианты применения в реальных условиях реальными людьми. Эталонное тестирование также связано с рисками переобучения и загрязнения тестовых данных обучающими данными. Закон Гудхарта гласит: *«Когда мера становится целью, она перестает быть хорошей мерой»*.

Оценка ИИ-систем по определению сложна, особенно если эти системы имеют широкое применение и обучаются после развертывания. Требуется четкий и прозрачный сбалансированный подход, который позволяет избежать переобучения по конкретным метрикам в ущерб общей надежности, справедливости и применимости в реальном мире. *Оценка* на уровне системы исследует репрезентативные варианты применения и не является всеобъемлющей. В качестве дополнительного метода можно привлекать красные команды, которые намеренно провоцируют модели с целью выявить отклонения от желаемого поведения.

В будущем при оценке ИИ необходимо будет учитывать различные аспекты производительности системы. Как правило, основным критерием оценки является *способность* выдать ожидаемый ответ или поведение в ответ на запрос или задачу в области, в которой система должна работать. И хотя эти требования важно выполнить, не следует забывать и о других аспектах работы системы.

Удобство использования — это еще один важный аспект оценки. Важный фактор удобства использования — *прозрачность*, то есть наличие механизмов, с помощью которых пользователи могут понять, на основе чего система выдает действия и ответы. Кроме того, удобство использования предусматривает *управляемость*, которая позволяет пользователям контролировать и менять поведение системы в соответствии с текущими и специализированными потребностями. Сейчас для этого часто используется термин alignment (согласованность). Для ИИ-систем, которые развертываются для помощи людям, оценка обязательно должна учитывать, повышает ли технология совместную производительность человека и системы.

Еще один важный аспект оценки — проверка на соответствие юридическим и этическим нормам. Все чаще правительства издают законы, ограничивающие использование ИИ-систем в их юрисдикции и требующие подтверждения того, что система будет работать в установленных рамках. Правительственные и коммерческие организации по этическим и финансовым причинам стремятся к честному и беспристрастному использованию ИИ. В поддержку этого стремления разрабатываются различные надежные фреймворки для оценки ИИ, которые помогают организациям оценивать ИИ-системы на предмет справедливости, прозрачности, устойчивости и соблюдения этических норм. Среди таких фреймворков можно выделить фреймворк оценки надежности ИИ в ЕС, фреймворк управления рисками ИИ Национального института стандартов и технологий, а также стандарт системы управления ИИ ISO/IEC 42001:2023.

Развертывание ИИ-систем связано с различными операционными проблемами. Главная из них — обеспечение конфиденциальности, то есть защита от утечки персональных и корпоративных данных, содержащихся в модели. ИИ-системы сами по себе превратились в объекты атак. Злоумышленники стараются получить

данные, узнать веса модели или исказить ответы, чтобы модель не выполняла поставленную создателями цель. При оценке общей производительности ИИ-системы также необходимо просчитать потребление ресурсов и затраты на обучение и развертывание.

Требуется учесть множество факторов, включая справедливость, устойчивость, интерпретируемость и соблюдение меняющихся законодательных норм. Комплексный фреймворк оценки должен поддерживать баланс между этими факторами, чтобы гарантировать безопасность, эффективность и соответствие ИИ-систем этическим и юридическим нормам.

Исследовательские задачи

Требуется развивать науку оценки ИИ-систем для внедрения дополнительных жестких стандартов оценки. Эта наука будет основываться на существующих метриках и методологиях, а также включать в себя новые подходы, чтобы мы чувствовали больше уверенности при развертывании ИИ-систем в критически важных средах (например, [8] для оценки систем генерации с дополненной выборкой). Фреймворки для аудита и воспроизводимости помогут обеспечить надежность и устойчивость результатов [9]. Кроме того, следует уделять больше внимания обучению надлежащей эмпирической методологии. Ниже приводятся ключевые действия по изучению методов эффективной оценки ИИ-систем.

- Тщательно изучить способы отслеживания и оценки ИИ-систем, развернутых в течение длительных периодов, особенно тех, которые со временем эволюционируют.
- Разработать фреймворки для оценки безопасности агентных ИИ-систем, которые могут осуществлять действия в реальном мире.
- Разработать методы повышения прозрачности моделей машинного обучения.
- Разработать методологии оценки, напрямую направленные на взаимодействие человека с возможностями ИИ (наподобие теста Тьюринга).
- Изучить баланс между различными аспектами оценки, например: окупает ли повышенная прозрачность дополнительные расходы; оправдывает ли соблюдение ограничений возможные последствия для конфиденциальности; как сочетание других аспектов влияет на общий результат.

1. Коэн, П.Р. (Cohen, P.R.) (1995 г.). Эмпирические методы для искусственного интеллекта. (Empirical Methods for Artificial Intelligence). MIT Press.
2. Ван, А., Сингх, А., Майкл, Дж., Хилл, Ф., Леви, О., и Боуман, С.Р. (Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S.R.) (2018 г.). GLUE: многозадачная платформа для оценки и анализа систем понимания естественного языка (GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding). BlackboxNLP@EMNLP.
3. Шолле, Ф. (Chollet, F.) (2019 г.). Об измерении интеллекта (On the Measure of Intelligence). ArXiv, abs/1911.01547.
4. Хендрикс, Д., Бернс, К., Базарт, С., Зоу, А., Мазейка, М., Сон, Д.С., и Штейнхардт, Дж. (Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D.X., and Steinhardt, J.) (2020 г.). Оценка понимания языка в условиях многозадачности (Measuring Massive Multitask Language Understanding). ArXiv, abs/2009.03300.
5. Хендрикс, Д., Бернс, К., Кадават, С., Арора, А., Базарт, С., Танг, Э., Сон, Д.С., и Штейнхардт, Дж. (Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D.X., and Steinhardt, J.) (2021 г.). Оценка решения математических задач с помощью набора данных MATH (Measuring Mathematical Problem Solving With the MATH Dataset). ArXiv, abs/2103.03874.
6. Рейн, Д., Хоу, Б.Л., Стикланд, А.К., Петти, Дж., Панг, Р., Дирани, Дж., Майкл, Дж., и Боуман, С.Р. (Rein, D., Hou, B.L., Stickland, A.C., Petty, J., Pang, R., Dirani, J., Michael, J., and Bowman, S.R.) (2023 г.). GPQA: A Graduate-Level Google-Proof Q&A Benchmark. ArXiv, abs/2311.12022.
7. Чен, М. (Chen, M.) и др. (2021 г.). Оценка больших языковых моделей, обученных с помощью кода (Evaluating Large Language Models Trained on Code). ArXiv abs/2107.03374
8. Шахул, Э., Джеймс, Дж., Анке, Л.Е., и Шокерт, С. (Shahul, E., James, J., Anke, L.E., and Schockaert, S.) (2023 г.). Технология RAG: автоматическая оценка генерации ответа, дополненной результатами поиска (RAGAs: Automated Evaluation of Retrieval Augmented Generation). Конференция Европейского подразделения Ассоциации вычислительной лингвистики (Conference of the European Chapter of the Association for Computational Linguistics).
9. Гундерсен, О.Е., Хельмерт, М., и Хоос, Х. (Gundersen, O.E., Helmert, M., and Hoos, H.) (2024 г.). Повышение воспроизводимости в исследованиях ИИ: четыре механизма, реализованных JAIR (Improving Reproducibility in AI Research: Four Mechanisms Adopted by JAIR). J. Artif. Intell. Res. 81, 1019-1041.

Мнение сообщества

Как показывают результаты опроса, сообщество не удовлетворено текущими методами оценки ИИ-систем. В целом, 75 % респондентов согласны или абсолютно согласны со следующим утверждением: *«Недостаток строгих норм при оценке ИИ-систем замедляет исследования в этой области»*. Только 8 % в целом не согласны или абсолютно не согласны с этим, а 17 % затруднились дать ответ. Эти результаты подчеркивают необходимость направить усилия на развитие методов оценки, в том числе на создание новых методов, которые лучше соответствуют появляющимся подходам и возможностям ИИ.

Учитывая ответы на первый вопрос, удивительно, что только 58 % респондентов в целом согласны или абсолютно согласны со следующим утверждением: *«Организации будут неохотно развертывать ИИ-системы, пока не появятся более четкие методы оценки»*. Примерно 17 % в целом не согласны или абсолютно не согласны с этим, а 25 % затруднились дать ответ. Если предположить, что недостаток контроля при исследовании приводит к недостатку контроля при применении, ответы на эти два вопроса показывают, что организации поспешно внедряют ИИ-системы без надлежащей оценки.

На вопрос *«Какой процент времени вы тратите на оценку по сравнению с другими аспектами работы с ИИ?»* 90 % респондентов ответили, что тратят больше 10 % времени, а 30 % – что тратят больше 30 %. Это указывает на то, что члены сообщества серьезно относятся к оценке и уделяют ей значительное внимание. Безусловно, такие временные затраты на процесс оценки заслуживают одобрения, но очевидно, что это слишком большая нагрузка, и требуется принять меры для упрощения этой задачи. Например, можно направить усилия на составление рекомендаций по оценке, чаще обмениваться наборами данных и продолжать совместно разрабатывать контрольные показатели.

При ответе на вопрос *«Что из следующего является наибольшим препятствием при оценке ИИ-систем?»* респонденты чаще всего выбирали отсутствие подходящих методологий оценки (40 %), закрытость систем (26 %), а также затраты средств и времени на проведение оценки (18 %). Эти результаты подчеркивают необходимость в разработке подходов к оценке, которые будут лучше соответствовать текущим методам и условиям развертывания.

Этика и безопасность ИИ

Вопросы этики и безопасности в сфере ИИ требуют единого подхода, поскольку кратко- и долгосрочные риски все теснее связаны друг с другом.

ПРЕДСЕДАТЕЛЬСТВУЮЩИЕ

Винсент Конитцер (Vincent Conitzer), Университет Карнеги – Меллона

Стюарт Рассел (Stuart Russell), Калифорнийский университет, Беркли

Основные выводы

- Из-за быстрого развития ИИ риски нарушения этики и безопасности вызывают все больше опасений, но пока у нас нет необходимых технических и регуляторных механизмов для их устранения.
- Новые угрозы, такие как киберпреступления с использованием ИИ и автономное оружие, требуют безотлагательного внимания, как и этические последствия новейших способов применения ИИ.
- Для решения проблем с этикой и безопасностью требуется взаимодействие специалистов из разных дисциплин, а также непрерывный контроль и четкое обозначение ответственности при разработке ИИ-решений.

Контекст и история

Большой успех ИИ означает большую ответственность. В связи с расширением возможностей и областей применения ИИ действия теоретиков и практиков в этой области оказывают заметное влияние на мировое сообщество. Это влияние может быть негативным, поэтому сообщество обеспокоено вопросами этики и безопасности разрабатываемых ИИ-систем. Оба термина не вполне точны, и их значения отчасти совпадают. С одной стороны, беспилотный автомобиль не должен сбивать пешеходов – это вопрос безопасности. С другой стороны, при внедрении технологий беспилотного вождения возникают этические проблемы. С одной стороны, алгоритмы оценки рисков не должны дискриминировать людей – это вопрос этики. С другой стороны, несправедливая дискриминация может приводить к опасным ситуациям, например, когда полиция действует на основе прогнозов. Когда рекомендательные системы с помощью манипуляций постепенно убеждают пользователей в существовании теорий заговоров, возникают вопросы по поводу безопасности и этики. Все эти опасения можно обобщить в одной фразе: системы искусственного интеллекта должны служить благом для людей. Однако вопрос о том, что такое «благо», остается спорным с философской и этической точек зрения. Различные фреймворки этичного ИИ, организации по обеспечению безопасности ИИ и попытки регулирования ИИ на уровне государств отражают различные подходы к решению этих проблем.

Кроме того, эти риски можно рассматривать в кратко- или долгосрочной перспективе. Считается, что борцы за этику беспокоятся о последствиях в настоящем времени, например о несправедливой дискриминации, а борцы за безопасность волнуются за будущее, в котором ИИ может привести к уничтожению человечества. Мы не согласны с таким разделением. Примеры выше показывают, что ИИ может представлять собой опасность уже сегодня, но мы также не имеем морального права создавать ИИ, если высок риск истребления человечества. Однако желание или нежелание размышлять о будущем всегда было камнем преткновения среди людей, озабоченных проблемами ИИ. И это связано с историей вопроса.

На протяжении десятилетий исследования ИИ переживали взлеты и падения. Было несколько периодов спада интереса («зим ИИ»), связанных с завышенными ожиданиями, когда исследователи в других областях информатики явно выражали скептицизм. До появления технологий глубинного обучения даже многие исследователи в сфере машинного обучения избегали термина «искусственный интеллект» в своих работах, предпочитая делать акцент на статистическом характере исследований. Сообщество исследователей ИИ научилось проявлять осторожность и избегать рассуждений о будущем, и тогда эту роль взяли на себя другие, например философ Ник Бостром [1].

Повсеместное применение ИИ приводит к росту обеспокоенности, причем те, кто выражает опасения, разделились на два лагеря. Одни переживают за будущее, в котором ИИ однажды может превзойти человечество во всем, и тогда наша жизнь изменится к худшему. Например, вырастет безработица и мы утратим ощущение цели, что может привести к социальной дезорганизации и системному коллапсу. Но самое серьезное опасение связано с очевидным последствием создания машин, превосходящих человека. Как сказал Алан Тьюринг в 1951 году: «Мы должны ожидать, что машины перехватят контроль». Если говорить подробнее, то учитывая известную сложность с правильным формулированием целей (проблема царя Мидаса), высока вероятность того, что ИИ-системы будут преследовать цели, не соответствующие нашим, а мы не сможем им помешать. Проблема усугубляется тем фактом, что «инструментальные цели», такие как самосохранение и приобретение ресурсов, являются логической необходимостью для достижения почти любой другой цели. Академическое сообщество ИИ не разделяет эти страхи, хотя в последнее время исследователи начинают сомневаться и заявляют об этом. Например, опубликовано открытое письмо с просьбой приостановить исследования в этой сфере [2].

С другой стороны, сообщество, больше обеспокоенное последствиями для настоящего, находит среди ученых больше поддержки, поэтому проводятся такие конференции, как AIES (ИИ, этика и общество) и FaccT (Справедливость, ответственность и прозрачность). Некоторые не верят в мрачное будущее по другим причинам. Например, они считают, что компании намеренно запугивают людей искусственным интеллектом для собственной выгоды – чтобы подчеркнуть значимость своей работы и отвлечь внимание от вреда, который они сами наносят [3]. Неясно, действительно ли что компании получают выгоду, рассказывая о том, что их технологии уничтожат человечество, но идея неизбежности может помешать эффективному реагированию и сместить акцент с негативных последствий в настоящем.

У двух лагерей есть и общие опасения. Например, и те, и другие, выступают против автономных систем летального вооружения [4, 5]. Искусственное разделение между ними только мешает решать проблемы с обеих сторон.

Текущая ситуация и тенденции

Недавние успехи в области ИИ, особенно LLM-моделей, привели к тому, что некоторые прогнозы уже начали сбываться. Необходимо продумывать безопасность этих систем, а также соответствие действий ИИ поставленным нами задачам [6]. Еще пять лет назад большинство исследователей ИИ не поверили бы, что на поведение ведущих ИИ-систем можно влиять, попросив, например, выбрать ответ,

который дал бы миролюбивый, этичный и мудрый человек, такой как Мартин Лютер Кинг или Махатма Ганди [7]. С другой стороны, современные подходы к согласованию поведения и целей, в том числе для выполнения указанного выше запроса, оказываются крайне неточными. Вполне оправданы сомнения в правильности такого пути развития.

В то же время, беспокойство по поводу большинства проблем, которые много лет вызывали опасения в ближайшей перспективе, только усилилось в связи с расширением возможностей и распространением ИИ-систем. Например, киберпреступники, которые имитируют романтические отношения с жертвами, теперь используют ИИ для автоматической замены лица в видеозвонках [8]. В целом, дипфейки теперь довольно сложно распознать, и они вызывают немало проблем в обществе, так как применяются в самых различных сценариях — от создания порнографических изображений из мести до проведения целых кампаний по дезинформации. В военной сфере уже активно используется автономное оружие [9]. С учетом того, что уже умеют ИИ-системы, мы начинаем беспокоиться о новых проблемах, которые угрожают нам сейчас и в ближайшем будущем.

Например, помогут ли они создать новые опасные химические соединения? Мы уже знаем, что можно создавать высокотоксичные молекулы (просто изменив знак системы, которая имеет противоположную цель) [10], а преступник, который недавно взорвал Cybertruck, спланировал атаку с помощью ChatGPT [11].

Недавно была основана Международная ассоциация безопасного и этичного ИИ (International Association for Safe and Ethical AI, IASEAI), которая признает общность интересов защитников этики и сторонников безопасности. В феврале 2025 года прошла первая конференция, на которой лично побывало 700 участников и еще больше — онлайн. Миссия организации — обеспечить гарантированно безопасную и этичную работу ИИ-систем — подчеркивает необходимость тщательных исследований и разработок в связи с поведением ИИ.

Исследовательские задачи

Научное сообщество играет в этом важную роль, в том числе потому, что не обязано отчитываться перед акционерами. Однако ведущие модели являются слишком масштабными и дорогостоящими, чтобы ими могло заниматься научное сообщество. Следует ли увеличить расходы на научные исследования? Можно ли привлечь к этому корпорации или это приведет к конфликту интересов? Разрешится ли ситуация сама собой, когда масштаб перестанет иметь такое значение?

На каком этапе следует искать и разрешать проблемы с этикой и безопасностью? Следует ли оценивать систему непосредственно перед развертыванием? Должны ли вопросы этики и безопасности учитываться с момента ее разработки? Необходимо ли непрерывно отслеживать систему при работе в реальных условиях? Можно ли официально подтвердить соответствие системы требованиям этики и безопасности, или это невыполнимо в век нейросетей? Что собой представляет устойчивая ИИ-система? По каким ранним признакам можно понять, что ИИ-системы выходят из-под нашего контроля? В целом, какие технологии помогут ответить на эти вопросы?

Кто несет ответственность за последствия, если системы часто создаются из набора компонентов, предоставленных разными группами специалистов? Можно ли создать модульный дизайн с четкими требованиями для каждого компонента?

Проблема согласования (обеспечение того, что ИИ-системы будут способствовать созданию будущего, к которому стремится человечество), вызывает ряд сложных вопросов. Самый очевидный из них: как учесть интересы всех людей [12]? А как же интересы людей, которые будут жить в будущем? Как гарантировать, что ИИ-системы не будут манипулировать потребностями людей, например, чтобы их было проще удовлетворять? Должны ли ИИ-системы помогать тем, кто хочет навредить другим? Как продвинутые ИИ-системы должны вести себя, если из-за самого их существования люди начинают терять смысл жизни?

Исследования ИИ традиционно редко попадают в поле зрения Институционального ревизионного совета по вопросам этики (Institutional Review Board, IRB). Правильно ли это? Например, следует ли проверять обучение на всех данных из интернета? Нужно ли проверять ИИ-системы, предназначенные для детей, чтобы защитить пользователей от психологического вреда [13]? Должны ли специалисты по оценке ИИ обучаться анализу этических принципов, а также надлежащему и последовательному рассмотрению отчетов о воздействии? В целом, каким должен быть оптимальный подход к обучению теоретиков и практиков в сфере ИИ по вопросам этики и безопасности?

Не всегда ясно, кто должен отвечать на эти вопросы: специалисты по информатике или представители других дисциплин? Ведь опасения относятся к самым разным сферам [14]. В какой степени многие из этих проблем можно решить с помощью единой методологии (например, универсальных техник согласования)? В какой степени требуется несколько отдельных методологий? Зависит ли это от универсальности технологии?

В ходе междисциплинарных исследований по-прежнему возникает множество препятствий. Потребует ли решение некоторых из этих вопросов привлечения специалистов в других дисциплинах?

Например, следует ли призвать политиков и политологов для исследования процессов коллективного формирования этих технологий? Какие вопросы должны задавать исследователи, и какая среда будет благоприятной для таких исследований?

1. Винсент Конитцер (Vincent Conitzer). Искусственный интеллект: где философский анализ? (Artificial intelligence: where's the philosophical scrutiny?). Prospect, 4 мая 2016 г.
2. Future of Life Institute. Приостановите масштабные эксперименты с ИИ: открытое письмо (Pause Giant AI Experiments: An Open Letter). 22 марта 2023 г.
3. Дарон Аджемоглу (Daron Acemoglu). Споры о безопасности ИИ ведутся неправильно (The AI Safety Debate Is All Wrong). Project Syndicate, 5 августа 2024 г.
4. Future of Life Institute. Роботы-убийцы стоят на пороге (Slaughterbots are here). https://futureoflife.org/project/lethal-autonomous-weapons-systems/
5. Клаудия Дрейфус (Claudia Dreifus). Тоби Уолш, эксперт по ИИ, спешит остановить роботов-убийц (Toby Walsh, A.I. Expert, Is Racing to Stop the Killer Robots). The New York Times, 30 июля 2019 г.
6. Брайан Кристиан (Brian Christian). Проблема согласования машинного обучения и гуманитарных ценностей (The Alignment Problem: Machine Learning and Human Values). W. W. Norton & Company, 2020.
7. Юнтао Бай (Yuntao Bai) и д.р. Конституционный ИИ: безобидность обратной связи ИИ (Constitutional AI: Harmlessness from AI Feedback). arXiv:2212.08073.
8. Мэтт Берджесс (Matt Burgess). Мошенничество на сайтах знакомств: дипфейки в реальном времени (The Real-Time Deepfake Romance Scams Have Arrived). Wired, 18 апреля 2024 г.
9. Сэмюэл Бендетт и Давид Кириченко (Samuel Bendett, David Kirichenko). Боевые дроны и ускорение гонки автономных вооружений в Украине (Battlefield Drones and the Accelerating Autonomous Arms Race in Ukraine). 01.10.25. https://mwi.westpoint.edu/battlefield-drones-and-the-accelerating-autonomous-arms-race-in-ukraine/
10. Дерек Лоу (Derek Lowe). Злонамеренная оптимизация (Deliberately Optimizing for Harm). 15 марта 2022 г. https://www.science.org/content/blog-post/deliberately-optimizing-harm
11. Сейдж Лаццаро (Sage Lazzaro). Два случая злоупотребления популярными ИИ-инструментами вызывают вопрос: когда винить инструменты? (Two misuses of popular AI tools spark the question: When do we blame the tools?). Fortune, 9 января 2025 г.
12. Винсент Конитцер (Vincent Conitzer) и др. Социальный выбор должен служить основой для согласования ИИ при обработке разнообразной обратной связи от людей (Social Choice Should Guide AI Alignment in Dealing with Diverse Human Feedback). ICML 2024. arxiv.org/abs/2404.10271
13. Блейк Монтгомери (Blake Montgomery). Мать судится с создателем ИИ-чатбота, который мог довести ее сына до самоубийства (Mother says AI chatbot led her son to kill himself in lawsuit against its maker). The Guardian, 23 октября 2024 г.
14. Яна Шайх Борг (Jana Schaich Borg) и др. Как создать этичный ИИ (Moral AI And How We Get There). Pelican, 2024 г.

Мнение сообщества

Результаты опроса подчеркивают важность вопросов этики, безопасности и согласования ценностей, – 67,5 % респондентов считают их актуальными или весьма актуальными для своих исследований. Это свидетельствует о том, что такие опасения влияют на разработку и развертывание ИИ-систем. Один участник опроса прокомментировал: «Мои студенты после выпуска делают нечто прямо противоположное тому, что сейчас нужно миру. Меня это разочаровывает». Это означает, что сейчас людям кажется, будто на практике никто не решает эти проблемы.

Среди главных этических проблем называют дезинформацию (75 %), конфиденциальность (58,75 %) и ответственность (49,38 %), что указывает на необходимость повышения прозрачности, объяснимости и подотчетности ИИ-систем. Кроме того, респонденты беспокоятся о недостатке ресурсов для этических исследований в сфере ИИ (57,86 %). Это указывает на то, что эта область требует большего финансирования и дополнительной организационной поддержки.

Участники опроса подчеркивают важность междисциплинарных подходов (85,5 %) к вопросам безопасности ИИ, называя в качестве главных стратегий поддержку технических исследований (71,88 %), регулирование (60,62 %) и образование (74,38 %). По-прежнему сложно найти баланс между краткосрочными этическими проблемами и долгосрочными угрозами для безопасности, но 55,63 % опрошенных считают, что оба лагеря должны скоординировать усилия, а не работать по отдельности.

Самыми эффективными решениями для налаживания сотрудничества могут стать совместные конференции (76,25 %) и междисциплинарное образование (64,38 %). В целом, опрос показывает, что сообщество все больше признает необходимость в проактивных, скоординированных и хорошо финансируемых усилиях, направленных на обеспечение соблюдения этических и общественных ценностей при разработке ИИ-систем.

Произвольные ответы подчеркивают потребность в более убедительных стимулах, юридической ответственности и обязательных стандартах безопасности, причем некоторые респонденты выступают за то, что ИИ-системы должны сами научиться ценностям, а не действовать в строгих рамках. Однако многие настроены скептически и считают, что этические опасения слишком туманны и политизированы и препятствуют развитию. Некоторые респонденты подчеркивают роль специалистов по философии и этике, а другие полагают, что имеющиеся стандарты не соблюдаются, а регулирование неэффективно. Также участники говорят о политических и структурных барьерах, выражая опасения, что политика и идеология мешают прогрессу.

Воплощенный ИИ

Воплощенный ИИ создает умных агентов, которые воспринимают и понимают физический мир и могут взаимодействовать с ним.

Основные выводы

- Интеллект возникает путем взаимодействия физического тела с физической средой.
- В воплощенном ИИ такое взаимодействие является обязательным условием для достижения реального интеллекта у ситуативных агентов.
- Роботы представляют собой хорошую научную и инженерную платформу для разработки воплощенного ИИ.

ПРЕДСЕДАТЕЛЬСТВУЮЩИЙ

Алан Макуорт
(Alan Mackworth),
Университет Британской
Колумбии

Контекст и история

В упрощенном виде историческое развитие ИИ можно разделить на две парадигмы. Первая основана на явных представлениях знаний, встроенных или полученных путем обучения. В основе второй лежит обучение с чистого листа, как в нейросетях. Оба подхода обычно не связаны с воплощением. Третий подход подразумевает, что для проявления интеллекта у ситуативных агентов требуется физическое воплощение [2]. Гипотеза состоит в том, что интеллект — на уровне эволюции вида и индивидуального развития — возникает путем непрерывного взаимодействия физического тела с физической средой. Третью парадигму называют «воплощенным ИИ».

Схожие темы, имеющие определенные различия, но основанные на центральной роли воплощения, возникли и в других когнитивных науках, включая психологию [9], нейробиологию [4,7] и философию [3,5]. Воплощенный ИИ обладает шестью характеристиками: воплощенный, встроенный, действующий, расширенный, возникающий и развивающийся. У воплощенного агента есть физическое тело. Ситуативный агент внедрен в определенное окружение, где могут действовать другие воплощенные агенты. Теория энактивизма утверждает, что познание происходит путем динамического взаимодействия агента со средой. Интеллект находится не только в контроллере агента, но и в его теле, а также проявляется при его взаимодействии со средой. Интеллект возникает на основе эволюции этого взаимодействия. Робот является искусственным и воплощенным с определенной целью агентом. Воплощенный ИИ подчеркивает тесную связь восприятия и действия. И действительно: восприятие часто является действием и наоборот. Следовательно, робототехника представляет собой идеальную тестовую среду для воплощенного ИИ. С этой целью, например, создаются роботы-футболисты для испытаний воплощенного ИИ [6]. Кубок RoboCup дал толчок новым экспериментам и теориям, связанным с обучением, принятием решений и действием воплощенных мультиагентных систем в режиме реального времени [8,10].

С точки зрения науки, воплощение является одним из важнейших требований для развития интеллекта. С практической точки зрения оно также необходимо в любых сферах применения, где требуется взаимодействие с реальным миром, например для беспилотных автомобилей или заводских роботов. Форма воплощения, например человекоподобный или не человекоподобный робот, может зависеть от потребностей людей, взаимодействующих с роботом.

Текущая ситуация и тенденции

Пока агент пассивно изучает мир по текстам и видео, он не научится принимать решения и действовать самостоятельно. Даже если текст содержит явную и правдивую информацию о мире, он не отражает обычные повседневные знания, которые подсказывает нам здравый смысл. Воплощенному агенту в реальном мире нужен этот здравый смысл [1], но получить его можно только путем взаимодействия. Пассивный просмотр видео не научит агента, как действовать в мире. В отличие от пассивных агентов, которые обычно обучаются на корреляционных моделях, воплощенные агенты изучают, тестируют и пересматривают причинные модели мира. Для освоения этого навыка воплощение является достаточным, но не обязательным основанием.

В настоящее время исследователи обучают роботов методом с подкреплением в ходе огромного числа испытаний в симуляциях и в физическом мире. Кроме того, ведется активная работа над адаптацией LLM-моделей для создания планов для роботов. Еще одно направление исследований — инвертирование вероятностных причинных моделей для выявления причинно-следственных связей у роботов, взаимодействующих с миром, настоящим или искусственным.

Исследовательские задачи

Исследователям еще предстоит ответить на множество вопросов. Можно ли полноценно обучить воплощенного агента с помощью существующих технологий? Требуется ли новый синтез ИИ и теории управления для дальнейшего прогресса? Можно ли использовать имеющиеся предварительно обученные языковые модели и (или) модели машинного зрения, чтобы улучшить способность воплощенных агентов к познанию? Можно ли создать достаточно реалистичные симуляторы и модели мира, чтобы обучать агентов только (или хотя бы по большей части) в симуляции? Действительно ли симуляционные агенты «обречены на успех»? Можно ли с помощью формальных методов доказать, что воплощенный агент (почти всегда) достигает целей, не нарушая требования безопасности?

Пока мы не можем создать умного ситуативного агента, который выполнял бы широкий ряд задач с эффективностью человека, но, возможно, некоторые или даже почти все компоненты для этого у нас уже есть. Основная сложность состоит в том, как вписать такого агента в реальный мир. Пока мы не видим серьезных препятствий к тому, чтобы создавать интеллектуальных воплощенных агентов, способных решать задачи как человек и даже лучше.

1. Брахман, Р. Дж. и Левеск, Х. Дж. (Brachman, R. J. and Levesque, H. J.) (2022 г.). Машины, похожие на нас: ИИ со здравым смыслом (Machines like Us: Toward AI with Common Sense). MIT Press.
2. Брукс, Р. А. (Brooks, R. A.) (1991 г.). Интеллект без представления (Intelligence without representation). Artificial Intelligence, 47:139–159.
3. Кларк, Энди (Clark, Andy) (2010 г.) Расширение разума: воплощение, действие и эволюция познания (Supersizing the mind: Embodiment, action, and cognitive extension). Oxford University Press.
4. Дамасио, Антонио (Damasio, Antonio) (2021 г.). Чувствовать и знать: создание сознательного разума (Feeling & knowing: making minds conscious). New York: Pantheon Books.
5. Ди Паоло, Эсекьель А. и Эван Томпсон (Di Paolo, Ezequiel A., Evan Thompson). Энактивный подход (The Enactive Approach). Руководство Routledge по воплощенному познанию (The Routledge Handbook of Embodied Cognition). Routledge, 2024 г. 85-97.
6. Макворт, А. К. (Mackworth, A. K.) (1993 г.). О видении роботов (On seeing robots). Компьютерное зрение: системы, теории и применения (Computer Vision: Systems, Theory, and Applications), стр. 1–13, под ред. Басу, А. и Ли, С. (Basu, A., Li, X.). World Scientific Press.
7. Шанахан, Мюррей (Shanahan, Murray). (2010 г.) Воплощение и внутренняя жизнь: познание и сознание в пространстве возможного разума (Embodiment and the Inner Life - Cognition and Consciousness in the Space Of Possible Minds). Oxford University Press.
8. Стоун, П. (Stone, P.) (2007 г.). Обучение и рассуждения для автономных агентов в мультиагентных системах (Learning and multiagent reasoning for autonomous agents). Материалы 20-й Международной совместной конференции по искусственному интеллекту (The 20th International Joint Conference on Artificial Intelligence, IJCAI- 07), стр. 13–30. http://www.cs.utexas.edu/□pstone/Papers/bib2html-links/IJCAI07-award.pdf.
9. Варела, Франсиско Х., Томпсон, Эван и Рош, Элеонора (Varela, Francisco J., Evan Thompson, Eleanor Rosch). Воплощенный разум, новая редакция: когнитивистика и человеческий опыт (The embodied mind, revised edition: Cognitive science and human experience). MIT press, 2017.
10. Виссер, У. и Буркхард, Х.-Д. (Visser, U., Burkhard, H.-D.) (2007 г.). RoboCup: десятилетие проблем и достижений (Robocup: 10 years of achievements and challenges). AI Magazine, 28(2):115–130.

Мнение сообщества

Опрос показывает мнение членов сообщества о воплощенном ИИ. Здесь приводится сводка по результатам. 31 % респондентов решили ответить на вопросы по этой теме. Результаты с разбивкой по вариантам ответов:

1. Насколько эта тема актуальна для вашего исследования? 74 % респондентов считают ее актуальной: 27 % – в некоторой степени актуальной, 25 % – актуальной, 22 % – весьма актуальной.

2. Является ли воплощение важным аспектом будущих исследований ИИ? С этим согласно 75 % респондентов: 43 % согласны, а 32 % абсолютно согласны.

3. Требуются ли для исследований воплощенного ИИ роботы или достаточно симуляций? 72 % считают, что да: 52 % полагают, что это полезно, а 20 % – что необходимо.

4. Является ли искусственная эволюция перспективным путем реализации воплощенного ИИ? С этим утверждением согласно 35 % опрошенных: 28 % согласны, 7 % абсолютно согласны.

5. Следует ли при изучении воплощенного ИИ полагаться на концепцию воплощения в психологии, нейробиологии и философии? С этим согласно 80 %: 50 % согласны и 30 % абсолютно согласны.

Поскольку респонденты сами принимали решение об участии в опросе по данной теме (около трети всех респондентов), необходимо учитывать фактор необъективности. Тем не менее важно отметить, что три четверти опрошенных говорят об актуальности воплощенного ИИ для своих исследований и почти столько же – о его важности для будущих исследований. Аналогичный процент респондентов считает, что робототехника (по сравнению с симуляцией) полезна или необходима для воплощенного ИИ. Только треть опрошенных видит в искусственной эволюции перспективный путь развития воплощенного ИИ. Однако почти все согласны, что важно изучать точку зрения на ИИ в когнитивных науках. В целом, эти результаты позволяют получить представление о будущем исследований в сфере воплощенного ИИ.

ИИ и когнивистика

ИИ может получить множество знаний из других областей когнитивистики — и, в свою очередь, способен внести большой вклад в их развитие.

Основные выводы

- Когнитивистика — это междисциплинарная область, возникшая в связи с ИИ и исследованиями гипотезы о вычислениях как о научном языке для понимания познания.
- Продолжающиеся взаимодействия между исследованиями ИИ и другими направлениями когнитивистики привели к ценным результатам, например к созданию когнитивной архитектуры.
- Результатом дальнейшей совместной работы могут стать важные открытия в обоих областях.

ПРЕДСЕДАТЕЛЬСТВУЮЩИЙ

Кеннет Д. Форбус
(Kenneth D. Forbus),
Северо-Западный университет

Контекст и история

ИИ был первой областью, основанной на интеллектуальной гипотезе о том, что вычисления могут стать научным языком для понимания природы интеллекта независимо от его основы. Когнитивистика стала второй такой областью, объединив исследователей в сфере ИИ, психологии, лингвистики, нейробиологии, антропологии и в других направлениях науки. Идеи в сфере вычислений, сформулированные при изучении ИИ, во многом повлияли на когнитивистику на начальных этапах. Со временем по ряду причин произошло отделение ИИ от остальных областей [3]. Мы видим большие преимущества в возобновлении совместной работы, а также в изучении того, как прогресс в сфере ИИ помогает понять процесс познания у человека в частности и у животных в целом, и наоборот — как прогресс в других областях когнитивистики способствует совершенствованию ИИ-систем. В некоторых случаях мы должны узнать, как воплотить в программном коде когнитивные способности живых организмов, а в других ситуациях следует намеренно выбрать иной путь, чтобы ИИ дополнял человеческое мышление и помогал людям работать продуктивнее.

Текущая ситуация и тенденции

Когнитивистика изучает широкий ряд явлений, но мы поговорим о трех областях, исследования в которых можно соотнести с ИИ.

Человекоподобное обучение и рассуждение

Многих животных можно чему-либо обучить, но люди значительно превосходят остальные виды по способностям к обучению и рассуждению. Удивительно, но обучение человека, если говорить в терминах машинного обучения, по большей части является пошаговым и непрерывным, при этом в нем эффективно используются данные и часто создаются выразимые модели (например, Гентнер и Маравилла (Gentner & Maravilla), 2018). Современные промышленные графы знаний содержат десятки миллиардов фактов, но им не достает выразительности, присущей человеческой концептуальной структуре. Существующие системы рассуждений, например решатели задач выполнимости булевых формул и инструменты верификации моделей, часто справляются с невероятно объемными задачами и предлагают такие сложные решения, на которые не способен человеческий мозг [2]. Однако системы рассуждений на основе ИИ не могут рассуждать достоверно при наличии неполных и частично неверных теорий предметной области, а также с опорой на обширный опыт, — как это делают люди.

Когнитивные архитектуры — это системы, которые исследуют гипотезы о фиксированных структурах, определяющих процессы и представления, используемые для познания [7]. С их помощью можно изучать возможности создания ИИ-систем, которые в режиме реального времени объединяют восприятие, познание и двигательный контроль для выполнения различных задач. Кроме того, они помогают лучше понять человеческий интеллект. Например, когнитивные архитектуры использовались для моделирования выводов (и прогнозирования) в области когнитивной психологии и нейронауки ([1,8]). Каждая когнитивная архитектура включает в себя несколько процессов и представлений, а различия заключаются в изучаемых аспектах человеческого познания и детализации выдвигаемых гипотез.

Социальные агенты

Одна из характерных черт людей — способность созидать и жить в коллективном мире, узнавать друг друга и постигать культурные социальные нормы путем взаимодействия. Прогресс в понимании того, как создаются социальные агенты, поможет строить ИИ-системы, которые живут в нашем мире как помощники и партнеры [9]. Обычно социальный ИИ разрабатывают независимо от теорий и открытий в социальных науках. Подход к изучению социального поведения людей и приобретения социальных навыков отличается.

Исследовательские задачи

Прогресс в решении этих проблем поможет создать более адаптивные системы ИИ, уменьшить нагрузку на вычислительные и экологические системы, лучше понять когнитивные способности человека, в том числе и навыки социального познания, а также получить более совершенные инструменты для мышления.

Человекоподобное обучение и рассуждение

1. Как разработать пошаговые методы человекоподобного обучения с эффективным использованием данных, позволяющие создавать выразимые модели?

2. Разработать формальные онтологии, которые охватывают весь диапазон человеческих концептуальных структур, касающихся концептов на основе абстракций и сенсомоторики.

3. Как ИИ-системы могут эффективно оперировать неполными и частично неверными теориями в предметной области и использовать в мыслительном процессе опыт всего человечества?

Когнитивные архитектуры

1. Расширение когнитивных возможностей высокого уровня у когнитивных архитектур за счет динамической интеграции всего спектра человеческих возможностей в ответ на требования задачи. Это включает различные формы мышления, метапознание, онлайн-обучение, непрерывное обучение по направлениям и типам знаний, а также участие в постоянном человеческом взаимодействии (напр. [6]). Эти возможности потребуют обучения и рассуждения о моделях физического мира, абстракциях и других агентах с использованием символических реляционных и модально-специфических представлений о настоящем, будущем и прошлом.

2. Исследование интеграции базовых моделей с когнитивными архитектурами, включая источники знаний, для интерпретации/генерации естественных условий (напр. [10]). Могут ли способности пошагового обучения, которые демонстрируют когнитивные архитектуры, преодолеть ограничения, связанные с устаревшей информацией в базовых моделях?

3. Разработка комплексного набора контрольных задач для оценки широты и интеграции когнитивных способностей человека с точки зрения комплексной эффективности, как описано выше. Задачи должны быть разнообразными и масштабными, чтобы обеспечить объективную оценку. Кроме того, задачи должны быть диагностическими, позволяющими выделить когнитивные способности и их взаимодействие, чтобы дать представление о конкретных сильных и слабых сторонах.

Социальные агенты

1. Облегчить обучение через взаимодействие: ИИ-системы текущего поколения обучаются через пассивное наблюдение социального поведения, а не через участие в нем (аналогично разделению между теорией принятия решений и теорией игр). Люди, напротив, совместными усилиями непрерывно вырабатывают линию поведения, адаптируясь друг к другу. Системы ИИ, в лучшем случае, имитируют взаимодействие, тренируясь с неподвижными смоделированными пользователями (например, посредством обучения с подкреплением и обратной связью от человека (Reinforcement Learning With Human Feedback, RLHF)), но не учитывают взаимную адаптацию. Таким образом, нам нужно исследовать пути поддержки (или моделирования) интерактивного обучения в широких масштабах.

2. Упростить методы сохранения конфиденциальности для ознакомления с социальными данными: человеческие социальные сигналы (лицо, голос) обычно раскрывают личность социального актора. Интерпретация социальных сигналов требует ознакомления с дополнительной информацией о ситуации (напр., значение улыбки зависит не только от выражения лица нужного человека, но и от того, кто еще находится в этой ситуации, что они делают, от характера физической среды и т.д.). Возможность собирать эту информацию разумно ограничена законом (напр., Регламентом ЕС об ИИ (EU's AI act)). Однако это существенно ограничивает возможности получения данных и развертывания приложений. Как мы создаем алгоритмы, которые идентифицируют социально значимую информацию, доказывая при этом, что ни один будущий алгоритм не сможет восстановить собранную информацию, нарушающую конфиденциальность/анонимность? Создание методов, позволяющих собирать и в то же время обезличивать социальные данные, имеет решающее значение для развития социальных агентов.

3. Разработать контрольные показатели взаимодействия: поскольку ИИ стремится создать системы с широкими социальными возможностями, необходим надежный способ измерения и оценки улучшения новых моделей. Это включает в себя характеристику потенциальной предвзятости, вопросов согласования ценностей, готовности модели прибегать к обману и т.д. Люди сейчас предлагают специальные наборы задач, но для исследований нужна разработка полной классификации задач и оценок. Социальные науки, наоборот, позволили разработать теоретически основанные онтологии для характеристики социальных ситуаций. Необходимы исследования, чтобы перевести эти выводы в систематические и всеобъемлющие стандарты социального и интерактивного поведения человека.

1. Андерсон Дж. Р. (Anderson, J. R.) (2007). Как человеческий разум может существовать в физической вселенной? (How can the human mind exist in the physical universe?) New York, NY: Oxford University Press.
2. Бьер А., Хойл М. и др. (Biere, A., Heule, M., et al.) (2021) Справочник по выполнимости, 2-е издание (Handbook of Satisfiability, 2nd Edition), IOS Press
3. Форбус К. (Forbus, K.) (2010). ИИ и когнитивистика: Прошлое и ближайшие 30 лет. (AI and Cognitive Science: The Past and Next 30 years). Topics in Cognitive Science, 2(3), стр. 346-356), https://doi.org/10.1111/j.1756-8765.2010.01083.x
4. Форбус К. (Forbus, K.) (2021). Оценка революции в ИИ с человеческой точки зрения. (Evaluating revolutions in artificial intelligence from a human perspective). ОЭСР, ИИ и навыки будущего, том 1: Возможности и оценка (AI and the Future of Skills, Volume 1: Capabilities and Assessments), OECD Publishing, Paris. DOI:https://doi.org/10.1787/004710fe-en
5. Гентнер Д. и Маравилла Ф. (Gentner, D. & Maravilla, F.) (2018 г.). Аналоговые рассуждения (Analogical reasoning). Л. Дж. Болл и В. А. Томпсон (ред.) (L. J. Ball & V. A. Thompson (eds.) Международный справочник по рассуждению и логическим выводам (International Handbook of Thinking & Reasoning), (стр. 186-203). NY, NY: Psychology Press.
6. Глюк К. и Лэрд Дж. (Gluck, K. & Laird, J.) (2019 г.). Интерактивное обучение задачам: люди, роботы и агенты осваивают новые задачи посредством естественного взаимодействия. (Interactive Task Learning: Humans, Robots, and Agents Acquiring New Tasks through Natural Interactions). MIT Press.
7. Коцеруба И., Цоцос Дж. К. (Kotseruba, I., & Tsotsos, J. K.) (2020 г.). 40 лет когнитивных архитектур: основные когнитивные способности и практическое применение. (40 years of cognitive architectures: Core cognitive abilities and practical applications). Artificial Intelligence Review, 53(1), 17–94. https://doi.org/10.1007/s10462-018-9646-y
8. Лэрд Дж. (Laird, J.) (2012 г.). Когнитивная архитектура Soar (The Soar Cognitive Architecture), MIT Press.
9. Лугрин, Биргит; Пелашо, Кэтрин; Траум, Дэвид (ред.) (Lugrin, Birgit; Pelachaud, Catherine; Traum, David (Ed.), (2021 г.). Руководство по социальному взаимодействию агентов (The Handbook on Socially Interactive Agents), стр. 433-462, ACM, New York, NY, USA, 2021 г., ISBN: 978-1-4503-8720-0.
10. Шумерс Т. Р., Яо С., Нарасимхан К. и Гриффитс Т. Л. (Sumers, T. R., Yao, S., Narasimhan, K., & Griffiths, T. L.) (2023 г.). Когнитивные архитектуры для языковых агентов. (Cognitive architectures for language agents). Труды в области машинного обучения (Transactions on Machine Learning Research), arXiv preprint arXiv:2309.02427.

Мнение сообщества

Отвечая на вопросы в этом разделе, 30 % респондентов отметили наличие пересечения с другими областями когнитивистики. Среди ответивших на вопрос о влиянии на их исследования 18 % выбрали вариант «всегда», 32 % – «обычно», 32 % – «иногда». Только 2,8 % выбрали вариант «никогда», а 12 % – «редко». Таким образом, 82 % респондентов в разумной степени испытывают влияние других областей когнитивистики на свои исследования. Какие области оказывают наибольшее влияние? Участники опроса выбрали следующие варианты: психология (82 %), нейронаука (44 %), лингвистика (40 %), антропология (22 %), другие области (13 %), среди которых чаще всего встречалась философия. С точки зрения того, какие вопросы являются наиболее актуальными, был дан широкий набор творческих ответов, некоторые были очень необычными, например, изучение того, как может функционировать разум, совершенно отличный от нашего собственного.

Аппаратное обеспечение и ИИ

Совместная разработка архитектуры аппаратного и программного обеспечения для ИИ включает в себя создание компонентов, которые вместе будут эффективно работать, максимально увеличивая производительность и энергоэффективность ИИ-систем.

ПРЕДСЕДАТЕЛЬСТВУЮЩИЙ

Джойдип Бисвас
(Joydeep Biswas),
Техасский университет
в Остине

Основные выводы

- Внедрение эффективных алгоритмов основывается на доступном оборудовании, а проектирование устройств оптимизировано для основных алгоритмов.
- Энергозатраты и производительность являются ключевыми проблемами при обучении больших моделей. При этом числовые представления, разреженность и параллелизм данных/моделей рассматриваются как ключевые факторы, способствующие масштабному обучению и получению выводов.
- Развертывание ИИ-систем на периферийных устройствах остается сложной задачей по нескольким причинам: потребности в распределении ресурсов и планировании для интеграционных систем и разнородного оборудования, энергетические нужды и ограничения в рассеянии тепла, а также требования к конкретному приложению в режиме реального времени.

Контекст и история

На протяжении всей истории ИИ, успешные развертывания были тесно связаны с особенностями аппаратного обеспечения – алгоритмы, которые использовали существующие функции оборудования были готовы к развертыванию. Затем последовали усовершенствования аппаратного обеспечения, чтобы ускорить работу основных алгоритмов. До широкого распространения и внедрения нейронных сетей существовало лишь несколько примеров специализированного оборудования для ИИ, позволяющего ускорить процессы исследования и оптимизации. Однако с широкомасштабным внедрением искусственных нейронных сетей пространство аппаратных ускорителей, предназначенных для ИИ, значительно расширилось.

По состоянию на 2025 год процесс совместного внедрения аппаратного и программного обеспечения для ИИ выглядит следующим образом:

- Алгоритмы, которые можно легко внедрить и масштабировать на текущем оборудовании, получили широкое распространение.
- Для разработки аппаратного обеспечения нужно ускорить вычислительные операции, которые считаются наиболее актуальными, с учетом используемых в настоящее время алгоритмов.
- Энергозатраты (потребление и потери) и производительность (данные и вычисления) являются самыми сложными задачами в обучении больших моделей.
- Численные представления, разреженность и параллелизм данных / моделей рассматриваются как ключевые факторы, способствующие масштабному обучению и получению выводов.
- Развертывание ИИ-систем на периферийных устройствах остается сложной задачей по нескольким причинам, включая потребности в распределении ресурсов и планировании для интеграционных систем и разнородного оборудования, энергетические нужды и ограничения в рассеянии тепла и требования к конкретному приложению в режиме реального времени.

Текущая ситуация и тенденции

Резюмируя прошлые и настоящие примеры совместного проектирования аппаратного и программного обеспечения по классам подходов к ИИ, можно отметить следующее:

- **ИИ для разработки аппаратного обеспечения:** компоновка микросхем и проектирование схем стали более эффективными благодаря автоматической маршрутизации, реализованной с помощью решателей задач целочисленного линейного программирования (Integer Linear Programming, ILP), а для проверки конструкций микросхем использовалась функциональная верификация. В последнее время наблюдается значительный интерес к применению методов машинного обучения при проектировании микросхем [11].
- **Символический ИИ, планирование и исследования:** Deep Blue [1], суперкомпьютер IBM, созданный для игры в шахматы и победивший мирового гроссмейстера Гарри Каспарова, продемонстрировал успех специализированного аппаратного обеспечения для исследований. В последнее время для ускорения планирования движения роботов используются программируемые пользователем вентильные матрицы (Field-Programmable Gate Arrays, FPGA) [2], графические процессоры (Graphics Processing Units, GPU) [6], а также инструкции с одиночным потоком команд и множественным потоком данных (Single Instruction, Multiple Data, SIMD) [8].
- **Вероятностные методы, численная оптимизация:** вычислительная геометрия, сочетание датчиков и оценка состояния основаны на операциях линейной алгебры с SIMD-ускорением (напр., Eigen – библиотека линейной алгебры C++), а численные решатели – на аппаратно-оптимизированной матричной факторизации. Библиотеки линейной алгебры Intel MKL, AMD OCL и Nvidia cuSOLVER включают себя как и специфичные для аппаратного обеспечения ускоренные операции линейной алгебры, так и специализированные процедуры плотной и разреженной факторизации. Для ускорения алгоритмов визуальной локализации и преобразования данных на периферийных устройствах используются специализированные интегральные схемы (Application-Specific Integrated Circuits, ASICs) [3].
- **Машинное обучение:** сегодня в машинном обучении доминируют искусственные нейронные сети. Существует широкий спектр аппаратных ускорителей для таких нагрузок, включая графические, тензорные, процессоры, блоки обработки изображений, процессоры Graphocore и нейроморфные вычисления. Тесную связь между аппаратным обеспечением и уровнем развития технологий машинного обучения можно сформулировать так: «Аппаратное обеспечение способствовало развитию технологий глубокого обучения, но теперь оно тормозит их дальнейший прогресс» [5]. Несмотря на то, что ландшафт архитектур моделей быстро меняется, к основным инновациям, которые оказались полезными для ускорения, можно отнести новые системы счисления [5], ускоренные матричные умножения, высокоскоростное подключение (в том числе оптическое) [7] и ограниченную разреженность [5]. Высокопроизводительное развертывание требует глубокой аппаратно-зависимой оптимизации. Хотя такие библиотеки общего назначения, как TensorFlow и PyTorch, дают готовое ускорение

для быстрой разработки прототипов и исследований, значительного повышения производительности можно добиться за счет аппаратно-зависимой оптимизации и оптимизации алгоритмов.

Исследовательские задачи

Системы счисления: современные модели показали, что они выигрывают от увеличения пропускной способности при снижении численной прецизионности с минимальной потерей точности. При том же общем количестве битов в модели сокращенные представления могут фактически демонстрировать более высокую производительность [10]. Адаптация аппаратного обеспечения к оптимизированным моделям систем счисления является перспективным направлением в будущем.

Разреженность: в то время как численные решатели полагаются на разреженность для эффективной масштабируемой матричной факторизации, аналогичных результатов трудно достичь для произвольных шаблонов разреженности в моделях машинного обучения. Не решен вопрос поддержки оборудования для более общих разреженных структур.

Масштабирование и ограничения на системном уровне: обучение современных моделей машинного обучения требует разработки мощных систем, выходящих за рамки ускоренных вычислений. Здесь можно выделить следующие проблемы: ограничение памяти и пропускной способности, параллелизм моделей и данных для масштабируемого распределенного обучения и получения выводов, пиковая пропускная способность хранилища для копирования, энергопотребление и терморегулирование.

Развертывание на периферии: помимо стремительного роста размеров и вычислительной сложности современных моделей существуют серьезные трудности в развертывании ИИ-систем на периферийных устройствах, включая использование энергии, рассеяние тепла и память. Кроме того, развернутые системы обычно интегрируются с огромным количеством разнородных компонентов, что ведет к проблемам с распределением ресурсов и планированием.

ИИ для систем и оборудования: предвидеть развитие алгоритмов ИИ сложно. Поскольку темп изменений ускоряется, разработчики аппаратного обеспечения и стека ПО неизбежно отстанут в разработке эффективных методов совместной оптимизации. Таким образом, развитие ИИ-технологий, которые помогают людям в проектировании и сокращают сроки оптимизации, информируют или контролируют адаптацию во время выполнения, вероятно, будут иметь первостепенное значение

1. Кэмпбелл М., Хоан-младший А.Дж. и Хсу Ф.Х. (Campbell, M., Hoane Jr, A.J., Hsu, F.H.), 2002 г. Deep blue. Artificial intelligence, 134(1-2), стр. 57-83).
2. Мюррей С., Флойд-Джонс У., Ци Й., Сорин Д. Дж. и Конидарис Г. Д. (Murray, S., Floyd-Jones, W., Qi, Y., Sorin, D.J. and Konidaris, G.D.), 2016 г., июнь. Планирование движения робота с чипом. (Robot motion planning on a chip). In Robotics: Science and Systems (Vol. 6), (том 6).
3. Чжан З., Сулейман А.А., Карлоне Л., Сзе В. и Караман С. (Zhang, Z., Suleiman, A.A., Carlone, L., Sze, V. and Karaman, S.), 2017 г. Визуально-инерциальная одометрия с чипом: подход к совместному проектированию алгоритмов и оборудования. (Visual-inertial odometry on chip: An algorithm-and-hardware co-design approach).
4. Прабхакар Р., Чжан Й., Кёплингер, Д. Фельдман, М., Чжао Т., Хаджис С., Педрам А., Козиракис К. и Олукотун К. (Prabhakar, R., Zhang, Y., Koeplinger, D., Feldman, M., Zhao, T., Hadjis, S., Pedram, A., Kozyrakis, C. and Olukotun, K.), 2017 г. Пластилин: перенастраиваемая архитектура для параллельных шаблонов. (Plasticine: A reconfigurable architecture for parallel paterns). ACM SIGARCH Computer Architecture News, 45(2), стр. 389-402.
5. Далли Б. (Dally, B.), 2023 г., август. Аппаратное обеспечение для глубинного обучения. (Hardware for deep learning). 2023 год, конференция IEEE Hot Chips 35 (In 2023 IEEE Hot Chips 35) Симпозиум (HCS) (Symposium (HCS), (стр. 1-58). IEEE Computer Society.
6. Сундарелингам Б., Хари С.К.С., Фишман А., Гарретт К., Ван Вайк К., Блукис В., Миллейн А., Олейникова Х., Ханда А., Рамос Ф. и Ратлифф Н. (Sundaralingam, B., Hari, S.K.S., Fishman, A., Garrett, C., Van Wyk, K., Blukis, V., Millane, A., Oleynikova, H., Handa, A., Ramos, F. and Ratliff, N.), 2023 г., май. Curobo: параллельное генерация движений роботов без столкновений. (Curobo: Parallelized collision-free robot motion generation). Международная конференция IEEE по робототехнике и автоматизации (2023 IEEE International Conference on Robotics and Automation) (ICRA), 2023 г., (стр. 8112-8119). IEEE.
7. Джоуппи Н., Курян Г., Ли С., Ма П., Нагараджан Р., Най Л., Патил Н., Субраманиан С., Свинг А., Тоуэлс Б. и Янг К. (Jouppi, N., Kurian, G., Li, S., Ma, P., Nagarajan, R., Nai, L., Patil, N., Subramanian, S., Swing, A., Towles, B. and Young, C.), 2023 г., июнь. TPU v4: Оптически перенастраиваемый суперкомпьютер для машинного обучения с аппаратной поддержкой встраиваний. (Tpu v4: An optically reconfigurable supercomputer for machine learning with hardware support for embeddings). Материалы 50-го ежегодного международного симпозиума по архитектуре компьютеров (Proceedings of the 50th Annual International Symposium on Computer Architecture), (стр. 1-14).
8. Томасон В., Кингстон З. и Кавраки Л.Е. (Thomason, W., Kingston, Z., Kavraki, L.E.), 2024 г., май. Движения в микросекундах с использованием векторизованного выборочного планирования. (Motions in microseconds via vectorized sampling-based planning). Международная конференция IEEE по робототехнике и автоматизации (ICRA) (2024 IEEE International Conference on Robotics and Automation (ICRA), 2024 г., (стр. 8749-8756). IEEE.
9. Саксена Д., Шарма Н., Ким Д., Двиведула Р., Чен, Дж. Янг, С., Равула, С., Ху З., Акелла А., Энджел С. и Бисвас Дж. (Saxena, D., Sharma, N., Kim, D., Dwivedula, R., Chen, J., Yang, C., Ravula, S., Hu, Z., Akella, A., Angel, S. and Biswas, J.), 2023 г. О базовой модели для операционных систем. (On a Foundation Model for Operating Systems). 7-й семинар по машинному обучению для систем. (7th Workshop on Machine Learning for Systems). 37-я конференция по нейронным информационным системам обработки (Held at 37th Conference on Neural Information Processing Systems), (NeurIPS 2023).
10. Деттмерс Т. и Зеттлемойер Л. (ettmers, T. and Zettlemoyer, L.), 2023 г., июль. Пример 4-битной точности: законы масштабирования инференса для k-бит. (The case for 4-bit precision: k-bit inference scaling laws). Международная конференция по машинному обучению (In International Conference on Machine Learning), (стр. 7750-7774). PMLR.
11. Грингард, Самуэль. (Greengard, Samuel). ИИ переосмысливает проектирование чипов. (AI Reinvents Chip Design.) (2024): 16-18. Communications of the ACM.

Мнение сообщества

Результаты опроса свидетельствуют о том, что сообщество единодушно считает необходимым тесное взаимодействие аппаратного обеспечения ИИ и исследований.

1. Отвечая на вопрос о совместной эволюции аппаратного обеспечения и ИИ, 75 % респондентов оценили этот процесс как «очень важный» (52,63 %) или «критически важный» (22,81 %), подчеркнув широко распространенное мнение о том, что прорывы зависят от интегрированного проектирования аппаратного и программного обеспечения. Когда их спросили о зависимости алгоритмического прогресса от аппаратного обеспечения, 61 % участников отметили, что достижения в развитии аппаратного обеспечения являются важными: 42,11 % заявили, что эти процессы «тесно связаны», а 19,30 % выразили мнение, что они «неразделимы».

2. Также имеется значительная поддержка для разработки абстракций, не зависящих от аппаратного обеспечения. В ответ на вопрос, согласны ли участники опроса с утверждением о том, что разработка абстракций, не зависящих от аппаратного обеспечения, необходима, 57,9 % отметили, что такие абстракции имеют решающее значение для того, чтобы позволить исследователям сосредоточиться на алгоритмических инновациях, не ограничиваясь деталями аппаратного обеспечения (47,37 % выбрали вариант «согласен» и 10,53 % – «полностью согласен»); 28,07 % затруднились ответить, а 14 % не согласились с этим утверждением (10,53 % выбрали вариант «не согласен», а 3,51 % – «категорически не согласен»).

3. Что касается использования оборудования, традиционные платформы продолжать доминировать в обучении и развертывании. Для обучения моделей 80,70 % респондентов используют графические процессоры, а 68,42 % – центральные процессоры. Аналогичная тенденция наблюдается при развертывании: 68,42 % используют центральные процессоры, а 68,42 % – графические.

4. Большинство предпочитает развертывать ИИ или на своих компьютерах (71,93 %), или на облачных платформах (59,65 %). В меньшей степении развертывание осуществляется на периферийных компьютерах с оптимизированным аппаратным обеспечением (19,3 %) или мобильных устройствах (15,79 %).

5. В отношении факторов, ограничивающих эффективную разработку моделей ИИ, опрос показал следующие результаты:

- Что касается ограничений в обучении, объем памяти является главной проблемой (52,63 %), затем следуют вычислительная производительность (49,12 %), пропускная способность памяти (35,09 %) и энергопотребление (28,07 %).
- Если говорить о проблемах с развертыванием, вычислительная производительность стала самым критически узким местом (47,37 %), также важную роль играют объем памяти (35,09 %) и энергопотребление (24,56 %).
- Несколько респондентов отметили стоимость для обучения и развертывания как дополнительную проблему вдобавок к другим вариантам ответа.

6. Среди респондентов, которые ответили на вопрос об интеграции нескольких компонентов, проблемы были равномерно распределены между несколькими факторами, включая ограничения в режиме реального времени (33,33 %), взаимодействие между компонентами (29,82 %) и управление конфликтом аппаратных ресурсов между компонентами (24,56 %).

ИИ для общественного блага

ИИ для общественного блага – раздел исследования ИИ, целью которого является измерение воздействия на общество, особенно на уязвимые и малообеспеченные группы населения, с упором на те области, в которых исторически не было достаточного объема исследований и разработок в области ИИ.

ПРЕДСЕДАТЕЛЬСТВУЮЩИЙ

Миллинд Тамбе (Millind Tambe),
Гарвардский университет

Основные выводы

- **Разработка ИИ, ориентированного на соблюдение этических норм и на положительное воздействие** – за последние 10 лет проекты, связанные с применением ИИ для общественного блага (AI for Social Good, AI4SG), стали более масштабными благодаря достижениям в сфере ИИ и МО. Их приоритетами являются этические соображения, справедливость и социальные блага, что обеспечивает ответственный подход к решению реальных проблем.
- **Междисциплинарное сотрудничество крайне важно** – успешные AI4SG-инициативы требуют тесного партнерства между исследователями ИИ, экспертами в предметной области, политиками и местными сообществами, чтобы обеспечить актуальность и долгосрочное устойчивое развитие.
- **Трудности с масштабированием и устойчивым развитием** – пока AI4SG демонстрирует значительный потенциал, поддержка и масштабирование решений в условиях ограниченности ресурсов остается ключевой задачей.

Контекст и история

«ИИ для положительного влияния на общество» (AI for Social Impact, AI4SI / AI for Social Good, AI4SG) выступает в качестве отдельного раздела изучения ИИ и фокусируется на измерении социального влияния на общество, в особенности на уязвимые и малообеспеченные слои населения. В отличие от традиционных исследований ИИ, в которых обычно приоритетом являются методологические достижения, главная цель AI4SI — прямое влияния на общество. Этот ИИ помогает решать проблемы, которым обычно не уделяется должного внимания в исследованиях ИИ, благодаря чему сокращается разрыв между возможностями ИИ и реальными социальными проблемами, такими как бедность, сельское хозяйство, здравоохранение и защита окружающей среды. Цель — создать эффективные решения для реальных проблем, часто в условиях ограниченных ресурсов.

Исследования AI4SI требуют глубокой вовлеченности экспертов в предметной области и членов сообщества, чтобы обозначить насущные проблемы, выработать эффективные меры вмешательства и тщательно оценить их воздействие. Этот междисциплинарный подход опирается на такие области, как взаимодействие человека и компьютера, здравоохранение и общественная деятельность, подчеркивая важность понимания и удовлетворения конкретных потребностей целевого сообщества. Эти проекты требуют баланса технических инноваций, этических соображений и практической осуществимости.

Семинар «ИИ для общественного блага», организованный Управлением научно-технической политики Белого дома, может считаться единственным событием, которое вызвало большой интерес к этой теме [1]. Итогом стало объединение различных усилий, направленных на общественное влияние, в одну общую развивающуюся область. Этот всплеск интереса объясняется несколькими ключевыми факторами. Во-первых, впечатляющие достижения в ИИ-технологиях, включая глубокое обучение, обработку естественного языка и обучение с подкреплением, предоставили мощные инструменты, применимые к широкому кругу социальных проблем. Доступность возросших вычислительных способностей и больших наборов данных дополнительно ускорила это прогресс, обеспечив разработку сложных ИИ-моделей.

Во-вторых, создание правительственных и промышленных программ финансирования, а также организация специализированных семинаров, конференций и особых потоков в рамках основных конференций по ИИ повысили осведомленность и привлекли исследователей в сферу AI4SG[2,3]. Возросший интерес академического сообщества привел к существенному росту числа публикаций, связанных с AI4SI, что демонстрирует увеличение востребованности этой области[4].

Текущая ситуация и тенденции

Развитие междисциплинарного сотрудничества становится нормой в применении ИИ для общественного блага. Эта необходимость обусловлена сложным характером общественных проблем, для решения которых часто требуются знания из разных областей. Тесное партнерство с экспертами в данной области, местными практическими специалистами и политическими деятелями гарантирует, что решения в области ИИ будут не только технически обоснованными, но также актуальными и эффективными в реальном мире. Такой совместный подход способствует общему понимаю проблемы и позволяет разрабатывать решения с учетом конкретных потребностей сообществ, для которых они предназначены.

Также важно уделять особое внимание этичности ИИ, при этом справедливость, прозрачность и конфиденциальность имеют первостепенное значение. Основные проблемы, такие как предвзятость при сборе данных и нежелательные последствия развертывания ИИ, должны решаться с самого начала, чтобы ИИ-системы соответствовали общественным ценностям и не причиняли вреда. AI4SG-проекты по своей природе подразумевают работу с уязвимыми группами населения и конфиденциальными данными, что делает этические соображения еще более важными. Активно решая потенциальные этические проблемы, исследователи смогут построить доверительные отношения с сообществами и гарантировать, что ИИ работает во благо, а не увеличивает существующее неравенство.

Будущее AI4SG определяют новые возможности. Во-первых, использование облачных платформ для масштабного развертывания ИИ обеспечивает более широкий охват и влияние. Облачные решения позволяют развертывать ИИ-инструменты отдалено или в местах с ограниченными ресурсами, что делает его более доступным. Во-вторых, обеспечение объяснимости и прозрачности ИИ-решений очень важно для завоевания доверия у практических специалистов и бенефициаров, особенно в таких ответственных областях, как ликвидация чрезвычайных ситуаций и здравоохранение. Если ИИ-системы продемонстрируют понятный процесс рассуждений и принятия решений, они с большей вероятностью будут приняты и эффективно использованы. В-третьих, акцент на локализованных AI4SI-решениях, которые могут устойчиво поддерживаться конечными пользователями, способствует долгосрочному эффекту и заинтересованности сообщества. Предоставляя местным сообществам возможность контролировать и обслуживать ИИ-инструменты, AI4SG-проекты гарантируют, что их преимущества сохранятся на долгое время после завершения первоначального этапа разработки. Наконец, эксплуатация доступных базовых моделей открывает широкие возможности для ускорения разработки и развертывания AI4SG-приложений. Эти предварительно обученные модели

могут служить отправной точкой для создания ИИ-решений, направленных на устранение конкретных социальных проблем, что позволит сократить время и ресурсы, требуемые для разработки [5].

Исследовательские задачи

Серьезная проблема, стоящая перед исследованиями ИИ для общественного блага, связана с разработкой ИИ-систем, которые были бы не только технически эффективны, но и глубоко контекстуально значимы в условиях социального влияния. Это требует тонкого понимания специфических нужд, культурных особенностей и практических ограничений сообществ, с которыми ведется работа. Исследователи должны выйти за рамки чисто алгоритмических соображений и принять участие в процессах разработки, в которых приоритет отдается голосам и опыту конечных пользователей, чтобы ИИ-решения действительно соответствовали их реальным потребностям и проблемам.

Другое существенное препятствие заключается в преодолении ограниченности данных. При реализации проектов с использованием ИИ для общественного блага часто возникают проблемы, связанные с нехваткой, низким качеством или предвзятостью данных, что может существенно повлиять на эффективность и беспристрастность ИИ-моделей. Разработка стратегий сбора достоверных данных, использование методов для аугментации данных и уменьшения предвзятости, исследование источников альтернативных данных важны для построения надежных и объективных ИИ-систем. При этом также требуется тщательное изучение культурного контекста, в котором собираются данные. Нужно убедиться, что методы по сбору данных ИИ адаптированы к местным практикам, а не навязаны стандартами извне.

Обеспечение устойчивости и масштабируемости развертывания ИИ в условиях ограниченности ресурсов представляет собой еще одну сложную проблему. После первоначального этапа создания прототипа жизнеспособность ИИ-решений в долгосрочной перспективе зависит от их способности поддерживаться и масштабироваться организациями с ограниченными ресурсами, такими как НПО и государственные учреждения. До этого требуется разработать устойчивые архитектуры ПО, создать удобные интерфейсы, а также обеспечить обучение и поддержку местных заинтересованных сторон. Более того, спонсирование этих работ является само по себе проблемой, потому что обычно отсутствует коммерческая целесообразность.

Кроме того, для анализа влияния ИИ-решений на местах и укрепления доверия заинтересованных сторон крайне важны надежные фреймворки оценки. Эти фреймворки должны выходить за пределы стандартных метрик эффективности и включать в себя показатели влияния на общество, удовлетворенности пользователей и этические соображения. Лозунг сообщества по обеспечению доступности: «Ничего о нас без нас» [6] служит мощным напоминанием важности вовлечения заинтересованных сторон на всех стадиях процесса оценки. Более того, необходимо бдительно следить за потенциальными попытками создать ложное впечатление об этике ИИ или экологичности. Это позволит гарантировать, что инициативы по ИИ действительно продиктованы стремлением к общественному благу, а не являются просто мероприятиями для привлечения внимания.

Напоследок следует упомянуть наличие разрыва между образованием в сфере ИИ и определенными умениями, необходимыми для эффективной общественной работы. Стандартные учебные программы по ИИ в основном посвящены разработке и анализу алгоритмов и зачастую сфокусированы на теоретических концепциях и производительности на основе контрольных наборов данных. Безусловно, такой подход важен для улучшения основных методологий ИИ, но он не учит студентов решать сложные, реальные проблемы, которые стоят перед AI4SG. Эффективные исследования AI4SG требуют более широкого набора навыков, выходящих за рамки чисто технической сферы. Для этого требуется навык эффективного сотрудничества с такими узкими экспертами, как представители здравоохранения, ученые-экологи или социальные работники. Также важно активно взаимодействовать с членами сообщества, на жизнь которых технология оказывает непосредственное влияние. Понимание нюансов социально-экономического и культурного контекстов социальных проблем и воплощение технических достижений в практические меры, ориентированные на пользователей, являются важнейшими компетенциями.

1. Организация Объединенных Наций (United Nations), (2015 г.). Преобразование нашего мира: повестка дня в области устойчивого развития на период до 2030 года. (Transforming our world: the 2030 Agenda for Sustainable Development). https://sdgs.un.org/2030agenda
2. Семинар «ИИ для общественного блага», организованный Управлением научно-технической политики Белого дома, 2016 г. (White House Office of Science and Technology Policy Workshop on AI for Social Good 2016) https://cra.org/ccc/events/ai-social-good/
3. Конференция AAAI: призыв к участию в специальном потоке «ИИ для положительного влияния на общество» (AAAI Conference Call for the Special Track on AI for Social Impact), 2024 г. https://aaai.org/aaai-24-conference/call-for-the-special-track-on-ai-for-social-impact/
4. Конференция IJCAI: прием заявок и проектов для многолетнего потока «ИИ и общественное благо» (IJCAI conference Call For Papers And Projects: Multi-Year Track On AI And Social Good) https://2025.ijcai.org/call-for-papers-and-projects-multi-year-track-on-ai-and-social-good-special-track/
5. Ши З., Ван Ч., Фан Ф. (Shi, Z., Wang, C., Fang, F.) ИИ для общественного блага: обзор (AI for social good: A survey), 2020 г. https://arxiv.org/abs/2001.01818
6. Управление Организации Объединенных Наций по социальному иклюзивному развитию (United Nations Division for Social Inclusive Development) Влияние ИИ на общее благо: отчет (AI for Good Impact Report), https://aiforgood.itu.int/newsroom/publications-and-reports/
7. Чжао Й., Бёмер Н., Танеджа А., Тамбе (M. Zhao, Y., Boehmer, N., Taneja, A., Tambe, M.) К созданию многоагентной системы на основе базовой модели в целях ускорения ИИ для положительного влияния на общество (Towards Foundation-model-based Multiagent System to Accelerate AI for Social Impact), AAMAS, 2025 г.
8. Чарлтон Дж. И. (Charlton, J. I.) Ничего о нас без нас: угнетение инвалидов и расширение их прав и возможностей. (Nothing About Us Without Us: Disability Oppression and Empowerment). University of California Press, 1998 г. http://www.jstor.org/stable/10.1525/j.ctt1pnqn9

Мнение сообщества

Опрос AAAI-сообщества, посвященный будущему исследований в области ИИ, в котором приняли участие 475 человек, касался различных аспектов ИИ, в частности его применения для общественного блага. Многие респонденты (119 человек) занимаются вопросами, связанными с применением ИИ для общественного блага.

Учитывая важность использования ИИ для общественного блага, большинство из 119 респондентов, ответивших на вопрос, отметили, что эта тема актуальна или весьма актуальна для их исследований. В частности, 33,61 % посчитали ее весьма актуальной, 26,89 % – актуальной, а 21,01 % – в некоторой степени актуальной. При ответе на вопрос о препятствиях при внедрении ИИ в их социально значимые проекты, респонденты обозначили несколько проблем. Самым серьезным препятствием, которое выбрали 47,06 % респондентов, стала «Проверка решения на практике». Другими существенными ограничениями оказались «Масштабирование решений» (38,66 %), «Связь с некоммерческими или правительственными организациями» (32,77 %), «Проблема определения» (36,13 %), и «Оценка готовности ИИ» (30,25 %). Проблему «Устойчивости бизнес-модели» также упомянули 42,86 % участников опроса.

В ходе исследования также были изучены важнейшие ресурсы, позволяющие масштабировать ИИ-решения для положительного влияния на общество. «Деньги» и «Данные» были отмечены как чрезвычайно важные для большинства участников опроса (в предоставленных фрагментах точные проценты вырезаны). Другими важными ресурсами являются «Партнерство правительства и компаний», «Партнерство правительства и университетов» и «Техническая поддержка». Наконец, что касается оценки успешности применения ИИ в решении социальных проблем, наиболее часто упоминаемым показателем было «Повышение доходов». Его выбрали 47,90 % из 119 респондентов. Кроме того, 37,83 % ответивших считают важным показатель «Долгосрочные результаты», а 48,74 % – «Уровень принятия».

ИИ и устойчивое развитие

ИИ стремительно меняет отрасли экономики и обладает огромным потенциалом, что способствует устойчивому развитию, начиная от ускоренного перехода на чистую энергетику до улучшения климатической устойчивости. Однако, его развертывание также создает проблемы: растущий спрос на энергию и воду. Чтобы ИИ не усугублял экологические риски, а способствовал устойчивому развитию, потребуются активные усилия по обеспечению его разработки, надлежащего использования, а также по определению сфер его применения.

ПРЕДСЕДАТЕЛЬСТВУЮЩИЕ

Эрик Хорвиц (Eric Horvitz), Microsoft

Хироаки Китано (Hiroaki Kitano), Sony Research

Основные выводы

- Хотя в настоящее время на долю вычислений на базе ИИ приходится лишь незначительный процент мирового потребления энергии и воды, стремительное развитие этих технологий в некоторых регионах создает нагрузку на местные электросети и водные ресурсы. Для управления этими факторами требуются инвестиции в местные электросети и инновации, которые улучшат эффективность аппаратного и программного обеспечения.
- Пока растет беспокойство о потенциальном воздействии ИИ на окружающую среду, исследователи и практические специалисты подчеркивают, что наиболее существенное влияние ИИ на устойчивое развитие (положительное и отрицательное) скорее связано с тем, как внедряют и используют ИИ, а не с энергопотреблением при обучении и эксплуатации моделей.
- ИИ может быть мощным инструментом для целей устойчивого развития и борьбы с изменениями климата. Помимо улучшения эффективности и сокращения выбросов углеводородов в промышленности, ИИ ускоряет прорывы в таких областях, как создание передовых материалов для аккумуляторных батарей, а также технологий для удаления углерода и высокоточного моделирования климата.

Контекст и история

ИИ технологии совершенствовались десятки лет, но недавние разработки в области LLM-моделей и их широкое внедрение способствовали более активному использованию ИИ-инструментов, требующих больших вычислительных затрат, во многих секторах. По мере увеличения вычислительной интенсивности и внедрения ИИ технологий растет и озабоченность по поводу их воздействия на окружающую среду, особенно в том, что касается потребления энергии и воды.

В то же время ИИ становится инструментом устойчивого развития, который может привести к большим преобразованиям. Достижение таких амбициозных климатических и экологических целей, как электрификация экономики, трехкратное увеличение мощностей возобновляемых источников энергии, декарбонизация промышленности и увеличение устойчивого производства продовольствия на 50 %, требует глобальной системной трансформации. ИИ может поддержать эти трансформации, тем самым улучшить мониторинг качества окружающей среды, оптимизировать энергосистемы, улучшить эффективность во всех отраслях и ускорить открытие новых материалов. Например, достижения в области материаловедения позволили разработать катализаторы, которые снижают стоимость улавливания углерода и уменьшают выбросы парниковых газов в результате таких промышленных процессов, как производство бетона.

Признавая эти возможности и вызовы, международные инициативы приводят развитие ИИ в соответствие с приоритетами устойчивого развития. Например, Международное энергетическое агентство (International Energy Agency, IEA) запустило инициативу по названием Energy for AI and AI for Energy («Энергия для ИИ и ИИ для энергии»), в рамках которой изучается, как ИИ может способствовать инновациям в сфере энергетики и управлять собственными потребностями в ресурсах. Весной 2025 года IEA опубликует специальный доклад по ИИ и энергетике и запустит обсерваторию ИИ, чтобы отслеживать потребление энергии искусственным интеллектом и его применение в энергетическом секторе. В 2025 году также была запущена новая инициатива, призванная установить показатели энергопотребления для разных моделей ИИ (AI Energy Score). Тем временем в рамках инициативы «Коалиция за устойчивый ИИ» (Coalition for Sustainable AI), созданной Францией в сотрудничестве с Программой ООН по окружающей среде (United Nations Environment Programme, UNEP) и Международным союзом электросвязи (International Telecommunication Union, ITU), разрабатываются рекомендации по минимизации воздействия ИИ на окружающую среду и продвигается передовой опыт в различных отраслях.

Текущая ситуация и тенденции

Тенденция: растущий спрос на ресурсы для вычислений на основе ИИ. Быстрая экспансия генеративного ИИ серьезно увеличивает потребность центров обработки данных (ЦОД) в энергии и воде, что обусловлено как обучением моделей, так и получением выводов с помощью уже обученных моделей. Например, сообщается, что обучение GPT-3 (175 млрд параметров) потребовало 1287 МВт электроэнергии и привело к выбросам 552 тонн CO2 [1]. Хотя обучение больших моделей очень энергозатратно, огромный спрос на энергию в долгосрочной перспективе, вероятно, будет обусловлен нагрузками, связанными с получением выводов, т. е. с запуском обученных моделей в реальных приложениях. За время работы модели получение выводов может потребовать гораздо больше энергии, чем обучение.

Согласно оценкам, на ЦОД – основу ИИ-инфраструктуры – в 2023 году приходилось около 2 % мирового спроса на электроэнергию [2] и менее 1 % общемирового объема выбросов парниковых газов [3]. Несмотря на то, что нагрузка, связанная с ИИ, составляет малую долю энергопотребления ЦОД, ожидается, что эта доля будет расти. В 2022 году на рабочие загрузки, связанные с ИИ, приходилось примерно 1 % от общего объема потребления электроэнергии в ЦОД. К 2026 году по прогнозам эта цифра вырастет до 9 % [3].

Согласно прогнозам, спрос на электричество у мировых ЦОД может удвоиться к 2030, впрочем степень этого роста будет зависеть от трендов рынка, алгоритмических улучшения и повышения эффективности аппаратного оборудования [4]. Даже при сценариях с высокими темпами роста, по оценкам МЭА, спрос на электроэнергию для ИИ останется относительно небольшой частью мирового энергопотребления [4].

Однако появляются региональные диспропорции. В некоторых сильно загруженных ИИ-центрах энергопотребление ЦОД быстро растет. Например, в ЕС спрос на электроэнергию для ЦОД увеличивается примерно на 9 % в год. С учетом роста вычислительных потребностей ИИ этот показатель превысит 5 % от общего спроса на электроэнергию в ЕС к 2026 году [3]. В США, на самом крупном рынке ЦОД, развитие ИИ привело к тому, что в 2023 году доля потребления электроэнергии центрами обработки данных превысила 4 % от общего объема энергопотребления в стране, увеличившись более чем в два раза с 2018 года. Согласно прогнозам, к 2028 году на ЦОД может приходиться от 7 до 12 % спроса на электроэнергию в США, в зависимости от сценариев развития ИИ [5].

Тенденция: энергоэффективный ИИ и инфраструктура на основе возобновляемых источников энергии. С учетом все более широкого внедрения ИИ разрабатывается несколько стратегий повышения устойчивости, в том числе:

- **Достижение эффективности аппаратного обеспечения:** графические процессы, широко используемые при решении

задач с помощью ИИ, потребляют больше энергии, чем традиционные центральные процессоры. Хотя абсолютное энергопотребление графических процессоров растет, эффективность каждого процессора при вычислении также повышается. [6,7] Оптимизация распределения средств аппаратного обеспечения – использование графических процессоров для решения задач, требующих большой вычислительной мощности, и применение центральных процессоров для более простых вычислений – поможет снизить общее потребление.

- **Малые языковые модели (Small Language Models, SLMs):** для больших языковых моделей (Large Language Models, LLMs) требуется большой объем вычислений, поэтому малые модели, оптимизированные для определенных задач, могут стать энергоэффективной альтернативой. SLM могут работать на ноутбуках и смартфонах, снижая вычислительную нагрузку, но при этом сохраняя производительность для целевых сфер применения [8].
- **Инновации в системах охлаждения:** обычные системы воздушного охлаждения в ЦОД неэффективны; переход на жидкое охлаждение поможет значительно сократить энергопотребление. Хотя некоторым системам жидкостного охлаждения нужна вода, достижения в области безводного охлаждения предлагают решения, которые сводят к минимуму потребление энергии и воды.
- **Оптимизация хранения данных:** нагрузки, связанные с ИИ, требуют хранения огромного количества данных, что увеличивает потребность ЦОД в электроэнергии. Такие методы, как сжатие и оптимизация данных, а также периферийные вычисления могут снизить энергозатраты.
- **Регулирование спроса и перераспределение нагрузки** – стратегии, которые помогают сбалансировать нагрузку на электросети, регулируя потребление электроэнергии в зависимости от их состояния. Регулирование спроса побуждает потребителей менять подход к электропотреблению: либо сокращать спрос в пиковое время, перенося его на периоды высокого предложения, либо использовать местные системы выработки и хранения энергии. Перераспределение нагрузки обычно направлено на изменение графика энергопотребления в соответствии с более дешевым или низкоуглеродным потреблением электроэнергии. Оба подхода все чаще используются для сокращения выбросов углерода путем переноса нагрузок с периодов высокой интенсивности выбросов на периоды, когда доступны более экологичные источники энергии.

Тенденция: применение ИИ для устойчивого развития. Искусственный интеллект выступает как преобразующий инструмент для устойчивого развития, предлагая три ключевые возможности, которые могут ускорить меры по борьбе с изменением климата и защите окружающей среды. ИИ-технологии играют центральную роль в области *вычислительной устойчивости* [9], ставя перед собой цель использовать математику, компьютерные науки и информатику, чтобы обеспечить устойчивое развитие и расширить возможности роста благосостояния всего человечества. ИИ может сыграть преобразующую роль в обеспечении устойчивого развития, предлагая возможности, которые улучшат эффективность, оптимизируют ресурсы и ускорят технологические прорывы.

ИИ-методы помогут расширить возможности человека по прогнозированию и оптимизации систем, повышая эффективность работы энергосетей, управления водными ресурсами и выполнения промышленных операций, одновременно сокращая объем отходов и выбросов. Достижения в области ИИ-моделирования используются в проектах, которые демонстрируют как ИИ-системы для распознавания образов, прогнозирования и оптимизации могут применяться при решении различных проблем устойчивого развития: начиная от сохранения и защиты дикой природы до повышения эффективности транспорта, прорывов в химии и материаловедении, способных ускорить открытие новых материалов, что приведет к развитию технологий производства аккумуляторов, улавливания углерода и создания низкоуглеродистых промышленных материалов.

В сфере обеспечения безопасности водных ресурсов и устойчивости к климатическим изменениям ИИ может радикально изменить методы гидрологического прогнозирования, использования ирригационных систем и подготовки к стихийным бедствиям [10,11]. ИИ для обнаружения утечек снижает потери воды [12]. В сфере управления климатическими рисками ИИ и машинное обучение используются для локального прогнозирования наводнений и аномальной жары с помощью уменьшенных климатических моделей, позволяя правительству лучше подготовиться к экстремальным погодным явлениям [13]. Мониторинг дикой природы с помощью ИИ теперь позволяет с точностью 99,3 % идентифицировать виды, что значительно повышает эффективность природоохранных мероприятий [14].

Перспективная сфера применения ИИ – оптимизация энергопотребления. В интеллектуальных сетях энергоснабжения ИИ применяется для прогнозирования спроса, балансировки нагрузки и интеграции возобновляемых источников энергии, помогая оптимизировать распределение энергии, уменьшить потери и снизить выбросы [15, 16]. Диагностирование неисправностей электросетей с помощью ИИ помогает коммунальным службам минимизировать сбои в работе [15].

Серьезные изменения происходят и в материаловедении, где ИИ значительно ускоряет открытие низкоуглеродных материалов. Ранее на разработку новых материалов для аккумуляторных батарей,

улавливания углерода и устойчивого строительства могли потребоваться годы или даже десятилетия. Сегодня ИИ-модели могут просканировать миллионы комбинаций материалов за несколько дней или недель [17]. Сотрудничество Microsoft и Тихоокеанской северо-западной национальной лаборатории позволило всего за девять месяцев создать новый твердотельный электролит для аккумуляторов. Ранее этот процесс с проведением экспериментов занимал бы годы [18].

ИИ может помочь в обучении и расширении возможностей персонала в области устойчивого развития, предоставляя ученым, политикам и инженерам инструменты для совершенствования процесса принятия решений и масштабирования устойчивых практик.

Исследовательские задачи

Потенциал ИИ для ускорения устойчивого развития очевиден. Однако у нас нет гарантий, что ИИ-технологии станут для него положительным фактором. Пока у ИИ есть потенциал в ускорении прогресса устойчивого развития, его потребности в ресурсах и энергии должны тщательно контролироваться. Чтобы ИИ мог ускорить прогресс в области устойчивого развития, потребуются целенаправленные исследования и инновации по целому ряду направлений, включая стратегические инвестиции в следующие сферы:

- Энергоэффективные ИИ-системы, которые минимизируют затраты на вычислительные процессы и потребление водных ресурсов.
- Инновационные ИИ-приложения, которые способствуют прорывам в области устойчивого развития.
- Надежное моделирование сценариев и сбор данных для обоснования политики и руководства по разработке устойчивого ИИ.

Благодаря активному управлению, целенаправленным исследованиям и межотраслевому сотрудничеству ИИ может действительно позиционироваться не как угроза устойчивому развитию, а как мощная сила, способствующая прогрессу в решении проблемы изменения климата.

Устранение пробелов в данных и больших неопределенностей

В то время как растет обеспокоенность по поводу потенциального влияния ИИ на окружающую среду, исследователи и практические специалисты подчеркивают, что наиболее существенное влияние ИИ на устойчивое развитие — положительное и отрицательное — скорее связано с тем, как внедряется и используется ИИ, а не с энергопотреблением при обучении и эксплуатации моделей [19]. Применение ИИ может привести к косвенным последствиям выбросов — как положительным, так и отрицательным [20].

Тем не менее, остаются значительные неопределенности. Трудно предсказать, как будут развиваться ИИ-технологии, и как их повсеместное внедрение повлияет на устойчивое развитие. Оценка суммарного воздействия ИИ на устойчивое развитие затруднена из-за двух основных проблем: (1) ограниченная доступность надежных данных и (2) сложность измерения реального влияния ИИ. Существуют широкие возможности разработки более полных наборов данных для лучшего понимания влияния ИИ на устойчивое развитие.

Многие ИИ-решения зависят от высокого качества наборов экологических и промышленных данных, но они обычно являются неполными, частичными или ориентированными только на самые развитые страны. Более того, большой объем отсутствующих данных затрудняет решение таких серьезных проблем устойчивого развития, как нехватка воды и сокращение биоразнообразия [21, 22]. ИИ-модели, обученные на ограниченных или предвзятых наборах данных, могут не учесть региональные изменения окружающей среды, что приведет к неточным прогнозам или несправедливым решениям в области устойчивого развития. Инвестирование в сбор данных и стандартизацию может помочь устранить эти пробелы.

Ограниченная доступность данных о потреблении энергии и воды искусственным интеллектом также представляет собой проблему. Лишь немногие компании раскрывают подробную информацию об энергопотреблении, углеродном следе и потреблении воды при выполнении вычислений с помощью ИИ. Отсутствие стандартизированной системы отчетности мешает политикам, исследователям и представителям общественности оценить реальное влияние ИИ на устойчивое развитие. Новые нормативные акты, такие как Регламент ЕС об ИИ, могут заполнить этот пробел.

Исследования с помощью моделирования и сценариев. ИИ может повысить эффективность в таких секторах, как транспорт, сельское хозяйство и производство, одновременно ускоряя процессы реализации природоохранных мер. Однако прогнозирование долгосрочного влияния внедренных ИИ-решений на устойчивое развитие остается сложной задачей. Ученые призывают к разработке фреймворков моделирования, оценивающих как непосредственное потребление ресурсов искусственным интеллектом, так и более широкие экологические последствия его внедрения при различных сценариях будущего [20]. В качестве примера неопределенности можно привести парадокс Джевонса, когда повышение эффективности приводит к росту общего потребления. С развитием аппаратного и программного обеспечения эффективность ИИ растет, а стоимость вычислений снижается. В результате ИИ становится более доступным и находит широкое применение. Но, как это ни парадоксально, такая доступность может привести к росту общего потребления энергии и ресурсов и свести на нет выгоды от повышения эффективности ИИ. Хотя отдельные ИИ-вычисления становятся менее энергозатратными, экспоненциальный

рост объема задач, решаемых с помощью ИИ, означает увеличение спроса на электроэнергию, что может нивелировать многие преимущества от повышения эффективности ИИ [23].

Для преодоления этих трудностей необходимо проанализировать варианты развития событий на основе различных стратегий. Анализ должен учитывать полное воздействие ИИ на окружающую среду, включая как прямое потребление энергии и воды, так и системные воздействия в различных отраслях, таких как здравоохранение, производство, сельское хозяйство и транспорт. Изменения, вызванные ИИ, могут ускорить процесс сокращения выбросов углекислого газа или повысить нагрузку на окружающую среду. На данный момент имеются только разрозненные исследования в этой области. Моделирование различных вариантов развития событий (сценариев) поможет в выборе стратегии и в управлении стратегическими инвестициями. Кроме того, это позволит ученым и разработчикам стратегий лучше понимать ключевые факторы неопределенности.

Сценарное моделирование широко используется в сфере финансов и оценки климатических рисков, позволяя количественно оценивать неопределенности. Исследуются различные варианты внедрения ИИ: от минимального уровня интеграции до повсеместного внедрения в соответствии с глобальными целями устойчивого развития. Исследователи должны разработать основы для прогнозирования и оценки различных вариантов будущего. Сценарии могут быть как оптимистичными, где ИИ способствует значительному сокращению выбросов, так и пессимистичными, где его безудержный рост приводит к увеличению нагрузки на окружающую среду. Такая аналитическая информация крайне важна для управления инновациями в области ИИ в направлении обеспечения устойчивого развития и снижения непреднамеренных рисков.

Разработка ресурсоэффективных ИИ-систем. Существует множество способов создания более ресурсо- и энергоэффективных ИИ-моделей и вариантов инфраструктуры. Ниже перечислены некоторые из них:

- Оптимизация архитектур ИИ-моделей для повышения энергоэффективности вычислений без ущерба для производительности.
- Разработка специализированного оборудования для ИИ, которое потребляет меньше энергии и воды, чем традиционные графические процессоры.
- Совершенствование механизмов управления ИИ-инфраструктурой для обеспечения низкоуглеродных вычислений. В рамках этого процесса выполнение вычислений с помощью ИИ планируется в зависимости от состояния сети, что позволяет минимизировать потребление энергии и, следовательно, сократить выбросы углерода.

Расширение применения ИИ-технологий для обеспечения критически важных возможностей устойчивого развития. Вероятно, наиболее преобразующие преимущества ИИ в области устойчивого развития будут связаны с появлением новых целевых приложений, направленных на решение ключевых экологических проблем. Искусственный интеллект открывает перед нами широкие горизонты для решения сложных проблем устойчивого развития. Он способен повысить эффективность транспортных систем и промышленных процессов, а также воплотить в жизнь достижения химии, материаловедения и биологии.

Например, ИИ может совершить революцию в области материаловедения, значительно ускорив процессы поиска новых материалов для аккумуляторных батарей (например, см. [24]), разработки решений для улавливания углерода и создания промышленных материалов с низким содержанием углерода. К числу других высокоэффективных возможностей относятся следующие:

- Разработка экономически эффективного решения для длительного хранения энергии, чтобы уменьшить зависимость от непостоянных возобновляемых источников, таких как ветер и солнце.
- Удаление углекислого газа в больших промышленных объемах по цене менее 100 долл. США за тонну.
- Повышение пропускной способности и надежности линий электропередач в связи с более широким использованием возобновляемых источников энергии.
- Сокращение утечек воды и газа в глобальном масштабе с помощью мониторинга на базе ИИ.
- Заполнение ключевых пробелов в данных о биоразнообразии и оптимизация программ по его сохранению с помощью ИИ.
- Внедрение новых методов повышения эффективности транспортных систем (см., например, [25]).

Для решения этих задач необходимо тесное взаимодействие между исследователями ИИ и экспертами в предметных областях. Это подразумевает создание новых методов и сфер применения ИИ, а также значительные инвестиции в создание и интеграцию соответствующих наборов данных для анализа, моделирования и машинного обучения.

1. Паттерсон Д., Гонсалес Дж., Ле К., Лян К., Мунгия Л. М., Ротшильд Д., ... и Дин Дж. (Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L. M., Rothchild, D., ... & Dean, J.) (2021 г.). Выбросы углерода и обучение крупных нейронных сетей. (Carbon emissions and large neural network training) arXiv preprint arXiv:2104.10350.
2. Международное энергетическое агентство (2024 г.). Электроэнергия 2024: анализ и прогноз до 2026 года. (Electricity 2024: Analysis and forecast to 2026.) https://www.iea.org/reports/electricity-2024
3. Международное энергетическое агентство (2024 г.). Центры обработки данных и сети передачи данных. (Data Centres and Data Transmission Networks.) https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks
4. Международное энергетическое агентство (2024 г.). Обзор мировой энергетики за 2024 год. (World Energy Outlook 2024.) https://www.iea.org/reports/world-energy-outlook-2024
5. Шехаби А., Смит С. Дж., Хаббард А., Ньюкирк А., Лей Н., Сиддик М. А. Б., Холечек Б., Куми Дж., Масанет Э., Сартор Д. (Shehabi, A., Smith, S.J., Hubbard, A., Newkirk, A., Lei, N., Siddik, M.A.B., Holecek, B., Koomey, J., Masanet, E., Sartor, D.) (2024). Отчет по энергопотреблению в центрах обработки данных США за 2024 год (2024 United States Data Center Energy Usage Report) (LBNL-2001637), Lawrence Berkeley National Laboratory, Berkeley, California.) https://eta-publications.lbl.gov/sites/default/files/2024-12/lbnl-2024-united-states-data-center-energy-usage-report.pdf
6. Технические характеристики продуктов Nvidia DGX A 100, H100 (2024). (Nvidia, Product specification sheets for DGX A100, H100) https://resources.nvidia.com/en-us-dgx-systems/ai-enterprise-dgx
7. Смит М. С. (Smith, M. S.)(2024 г.), Борьба претендентов за корону Nvidia: в «Игре престолов» ИИ не стоит недооценивать новичков. (Challengers are Coming for Nvidia's Crown: In AI's Game of Thrones, Don't Count Out the Upstarts.) IEEE Spectrum, Выпуск 61, № 10, с. 40-44, октябрь 2024 г., doi: 10.1109/MSPEC.2024.10705376) https://ieeexplore.ieee.org/document/10705376
8. ЮНЕСКО (март 2024 г.). Малые языковые модели (SLMs): более дешевый и экологичный путь к ИИ (Small Language Models (SLMs): A Cheaper, Greener Route into AI) https://www.unesco.org/en/articles/small-language-models-slms-cheaper-greener-route-ai#
9. Гомес К., Диттерих Т., Барретт К. и др. (Gomes, C., Dietterich, T., Barrett, C., et al.) (2019 г.). Устойчивость вычислений: вычисления для обеспечения лучшего мира и устойчивого будущего (Computational sustainability: Computing for a better world and a sustainable future), Communications of the ACM. 62 (9): 56–65.) https://dl.acm.org/doi/pdf/10.1145/3339399
10. Флекер А.С. и др. (Flecker, A.S., et al.), (2022 г.). Снижение негативных последствий расширения гидроэнергетики Амазонки. (Reducing adverse impacts of Amazon hydropower expansion.) Science № 375, с. 753-760. DOI:10.1126/science.abj4017 https://www.science.org/doi/10.1126/science.abj4017
11. Ролник Д. и др. (Rolnick, D. et al.) (2022 г.). Борьба с изменением климата с помощью машинного обучения. (Tackling Climate Change with Machine Learning.) ACM Computing Surveys 55, 2, статья 42 (февраль 2023 г.), 96 страниц. https://dl.acm.org/doi/10.1145/3485128
12. Болгар С. (Bolgar, C.) (сентябрь 2024 г.). Инструмент ИИ использует звук для определения места протечки труб, экономя драгоценную питьевую воду. (AI tool uses sound to pinpoint leaky pipes, saving precious drinking water.) Источник: Microsoft News. https://news.microsoft.com/source/features/sustainability/ai-tool-uses-sound-to-pinpoint-leaky-pipes-saving-precious-drinking-water/
13. Йошикане Т. и Йошимура К. (Yoshikane, T., & Yoshimura, K.) (2023 г.). Способ масштабирования и корректировки смещения, используемый в ансамблевом моделировании климатических моделей для локальных почасовых осадков. (A downscaling and bias correction method for climate model ensemble simulations of local-scale hourly precipitation.) Scientific Reports, 13(1), 9412.
14. Норуззаде М.С. и др., (Norouzzadeh, M.S., et al.) (2018 г.). Автоматическая идентификация, подсчет и описание диких животных на снимках, сделанных с помощью камеры-ловушки, по методу глубокого обучения. (Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning) Материалы Национальной академии наук США 115 (25) E5716-E5725, https://doi.org/10.1073/pnas.1719367115 (2018). https://www.pnas.org/doi/10.1073/pnas.1719367115
15. К. Дж. Бенеш, Дж. Э. Портерфилд, К. Янг. (Benes, K. J., Porterfield, J. E., & Yang, C.) (2024 г.). ИИ в энергетике: возможности экономики при современных энергосистемах экологически чистой энергии. (AI for energy: Opportunities for a modern grid and clean energy economy.) Министерство энергетики США.
16. Д. Сандалов, К. Маккормик, А. Кучукельбир и др. (Sandalow, D., McCormick, C., Kucukelbir, A., et al.) (2024 г.). Дорожная карта ИИ для смягчения последствий изменения климата (второе издание) (Artificial Intelligence for Climate Change Mitigation Roadmap (Second Edition) (Проект «Дорожная карта инноваций МИЭФ», ноябрь 2024 г.) https://doi.org/10.7916/2j4p-nw61.
17. А. Мерчант, С. Батцнер, С. С. Шенхольц, М. Айкол, Г. Чон и Э. Д. Кубук (Merchant, A., Batzner, S., Schoenholz, S. S., Aykol, M., Cheon, G., Cubuk, E. D.) (2023 г). Масштабирование глубокого обучения для поиска новых материалов. (Scaling deep learning for materials discovery.) Nature, 624(7990), с. 80-85.
18. Ускорение поиска новых материалов с помощью ИИ и элементов Azure Quantum (2024 г). (Accelerating materials discovery with AI and Azure Quantum Elements (2024). Блог Microsoft Azure Quantum https://azure.microsoft.com/en-us/blog/quantum/2023/08/09/accelerating-materials-discovery-with-ai-and-azure-quantum-elements
19. Каак Л. Х., Донти П. Л., Штрубелл Э., Камия Г., Крейциг Ф., и Рольник Д. (Kaack, L. H., Donti, P. L., Strubell, E., Kamiya, G., Creutzig, F., & Rolnick, D.) (2022 г.). Использование искусственного интеллекта для смягчения последствий изменения климата. (Aligning artificial intelligence with climate change mitigation.) Nature Climate Change 12 (6), с. 518-527.
20. А. Луерс, Дж. Куми, Э. Масанет, О. Гаффни, Ф. Крейциг, Дж. Лависта Феррес и Э. Хорвиц (Luers, A., Koomey, J., Masanet, E., Gaffney, O., Creutzig, F., Lavista Ferres, J., & Horvitz, E.) (2024 г.). Ускорит или отсрочит ИИ гонку за нулевыми выбросами? (Will AI accelerate or delay the race to net-zero emissions?) Nature, 628(8009), с. 718-720. https://www.nature.com/articles/d41586-024-01137-x
21. Люнг Б. и Гонсалес А. (Leung, B., & Gonzalez, A.) (2024 г.). Глобальный мониторинг биоразнообразия: анализ неопределенности, рисков и возможностей для выявления изменений тенденций. (Global monitoring for biodiversity: uncertainty, risk, and power analyses to support trend change detection.) Science Advances, 10(7), eadj1448.
22. Л. Роза и М. Санджорджио (Rosa, L., Sangiorgio, M.) (2025 г.). Глобальный дефицит водных ресурсов в условиях будущего потепления (Global water gaps under future warming levels). Nature Communications 16(1), 1192.
23. Луччиони А. С., Штрубелл Э. и Кроуфорд К. (Luccioni, A. S., Strubell, E., Crawford, K.) (2025 г.). От повышения эффективности к эффекту отдачи: проблема парадокса Джевонса в поляризованных экологических дебатах об ИИ. (From Efficiency Gains to Rebound Effects: The Problem of Jevons' Paradox in AI's Polarized Environmental Debate.) arXiv preprint arXiv:2501.16548.
24. Чен К. и др. (Chen, C., et al.) (2024 г.). Ускорение компьютерного поиска материалов с помощью машинного обучения и высокопроизводительных облачных вычислений: от крупномасштабного скрининга до экспериментальной проверки. (Accelerating Computational Materials Discovery with Machine Learning and Cloud High-Performance Computing: from Large-Scale Screening to Experimental Validation), Journal of the American Chemical Society 2024 146 (29), 20009-20018DOI: 10.1021/jacs.4c03849. https://pubs.acs.org/doi/abs/10.1021/jacs.4c03849
25. Камар Э. и Хорвиц Э. (Kamar, E., Horvitz, E.) (2009 г.). Сотрудничество и общие планы в открытом мире: исследования совместного использования транспортных средств. В материалах 21-й Международной совместной конференции по ИИ (Collaboration and shared plans in the open world: Studies of ridesharing. In Proceedings of the 21st International Joint Conference on Artificial Intelligence) (IJCAI'09). Morgan Kaufmann Publishers Inc., Сан-Франциско, Калифорния, с. 187–194. https://web.archive.org/web/20220119103915id_/https://www.ijcai.org/Proceedings/09/Papers/041.pdf

Мнение сообщества

Недавний опрос, проведенный среди участников ИИ-сообщества, показал, что мнения о воздействии ИИ на окружающую среду разделились:

- Около 35 % респондентов согласились с тем, что вред, который ИИ наносит окружающей среде, перевешивается его потенциалом в решении климатических проблем. Другие 35 % респондентов выразили несогласие или категорическое несогласие с этим утверждением.
- Более 70 % респондентов считают, что ИИ, оперирующий большими объемами данных, значительно влияет на глобальное потребление ресурсов.
- 57 % респондентов выразили обеспокоенность тем, что слишком большое энергопотребление ИИ может замедлить темпы исследований в этой области.
- Почти 75 % респондентов считают, что для снижения энергопотребления ИИ наиболее важными являются энергоэффективное обучение и алгоритмы получения выводов. За ними следуют оптимизация энергоснабжения центров обработки данных (20 %) и внедрение энергоэффективных чипов (5 %).

На вопрос о том, в каких сферах ИИ может оказать наибольшее влияние на устойчивое развитие, были получены следующие ответы:

- Более 30 % респондентов указали на логистику, транспорт и оптимизацию инфраструктуры, как на главные направления, где ИИ может быть особенно полезен. Около 10 % участников опроса также отметили важность ИИ в таких сферах, как снижение выбросов углекислого газа, сельское хозяйство, прогнозирование природных катаклизмов и продвижение экономики замкнутого цикла. Однако только 5 % считают, что ИИ обладает большим потенциалом для обеспечения устойчивого развития в сфере сохранения биоразнообразия.

Эти выводы демонстрируют, что ИИ-сообщество осознает как возможности, так и риски, связанные с использованием ИИ для достижения целей устойчивого развития.

ИИ в области научных открытий

Искусственный интеллект совершает настоящую революцию в научных открытиях. Он ускоряет весь исследовательский процесс: от извлечения знаний и выдвижения гипотез до автоматизации экспериментов и их проверки с невиданной ранее скоростью.

Основные выводы

- Продвижение использования ИИ в области научных исследований значительно ускорит процесс открытий и полностью изменит подход к научно-исследовательской работе.
- Разрабатываются высокоавтоматизированные ИИ-системы, которые, несмотря на определенные ограничения, способны самостоятельно проводить научные исследования, минимизируя участие человека в решении определенных задач.
- Использование ИИ для совершения научных открытий ставит перед нами новые этические, социальные и технические вопросы, для решения которых требуется комплексный подход.

ПРЕДСЕДАТЕЛЬСТВУЮЩИЙ

Хироаки Китано
(Hiroaki Kitano), Sony Research

Контекст и история

В прошлом научные открытия были результатом человеческой изобретательности. Однако с появлением и развитием ИИ этот процесс изменился и ускорился.

В 1960-х годах были созданы первые ИИ-системы, такие как DENDRAL, которые позволили автоматизировать процесс выдвижения гипотез и решения задач в области органической химии [1]. Также были разработаны такие компьютерные программы, как EURISKO, где использовались эвристические методы обучения и адаптации, раскрывающие возможности ИИ в творческом поиске решений [2].

Эти новаторские работы заложили основы для использования ИИ в науке, продемонстрировав его способность обрабатывать большие объемы данных и выдвигать гипотезы.

С развитием технологий, позволяющих проводить сложные и высокоточные измерения, ученые сталкиваются с огромным объемом данных, который необходимо обработать, и сложностью систем, стоящих за этими данными. Для проведения научных исследований необходимы технологии, позволяющие выявлять закономерности в больших массивах данных, которые могут быть связаны с новыми гипотезами, подлежащими проверке с помощью экспериментов. ИИ в области научных открытий становится все более востребованным направлением исследований, которое, как ожидается, будет способствовать развитию науки благодаря применению подходов на основе данных.

Существует целый ряд подходов к преобразованию области научных открытий с помощью ИИ. Во-первых, предполагается, что ИИ-инструменты будут становиться все более совершенными. Это позволит ученым и исследователям быстрее находить новые решения и справляться с более сложными задачами.

Альтернативный подход заключается в создании интегрированных ИИ- и роботизированных систем, которые способны самостоятельно проводить все этапы научных исследований, минимизируя участие человека. Промежуточный вариант – рассматривать ИИ как партнера в научной деятельности, способного совместно с учеными решать научные задачи в режиме реального времени. ИИ в области научных открытий охватывает широкий спектр взаимодействий между ИИ и учеными. Ожидается стремительный прогресс во всех областях этого взаимодействия, что существенно изменит подход к научной деятельности. В ряде статей, докладов и семинаров был сделан вывод, что использование ИИ в науке является одним из наиболее важных направлений исследований на ближайшие годы [3-6].

Текущая ситуация и тенденции

1. От простых инструментов до полноценных ИИ-сотрудников и автономных ИИ-ученых

Возможности ИИ в совершении научных открытий значительно возросли благодаря таким прорывным системам, как ИИ-программа AlphaFold2, разработанная DeepMind, которая продемонстрировала впечатляющие результаты в области структурной биологии [7,8]. Программа AlphaFold2 стала настоящим прорывом в области молекулярной биологии, решив десятилетнюю проблему сворачивания белков. Это достижение открывает новые возможности для разработки лекарств и развития биомедицины. AlphaFold2 – это яркий пример успешного применения ИИ в научных исследованиях. Эта программа произвела настоящую революцию в области биомедицины и биохимии, результатом чего стало присуждение Нобелевской премии по химии в 2024 году. Все большее число ИИ- и роботизированных систем разрабатывается для ученых в качестве эффективных инструментов для их исследовательской работы.

Помимо биологии, для ускорения научных открытий были разработаны многочисленные ИИ-инструменты для химии [9-11], материаловедения [12], математики [13-15] и многих других научных областей. Например, машина Рамануджана – это ранняя попытка создания автоматизированной системы генерации математических гипотез для фундаментальных констант, которая служит инструментом для математиков, решающих определенную математическую задачу [13]. Были разработаны интегрированные ИИ- и роботизированные системы, которые автоматически выполняют определенный тип химического эксперимента [16,17].

Абсолютно противоположным вариантом являются ИИ- и роботизированные системы с высокой степенью автономности, которые способны самостоятельно проводить научные исследования без участия человека или с его минимальным вмешательством. В качестве примера таких решений можно привести системы «Адам» и «Ева», созданные Россом Кингом (Ross King), в которых научные открытия происходят без участия человека [18,19]. Эти «ученые-роботы» не только выдвигают гипотезы, но также разрабатывают и проводят эксперименты для их проверки. Например, система «Ева» смогла обнаружить перспективные препараты для лечения малярии с помощью автоматизированного и высокопроизводительного скрининга, продемонстрировав способность ИИ самостоятельно проводить научные исследования[20]. Проект Nobel Turing Challenge представляет собой амбициозную попытку создания ИИ- и роботизированных систем с высокой степенью автономности, способных совершать крупные научные открытия. [21, 22]. Предлагаемый как масштабная концепция развития, этот проект направлен на разработку ИИ-систем, которые могут привести к настоящим прорывам в науке, сопоставимым с теми, что были отмечены Нобелевской премией. Проект включает в себя тест Фейгенбаума, который позволяет определить, способна ли компьютерная

система воспроизвести лучшего эксперта-человека в определенной области. Этот тест является разновидностью теста Тьюринга [23]. Эти системы будут не просто помогать исследователям, а выступать в качестве автономных структур, способных предлагать, проверять и уточнять теории.

2. Влияние на науку

Интеграция ИИ в научные процессы предвещает смену парадигмы:

- **Ускорение научных открытий:** благодаря ИИ, который способен автоматизировать весь процесс научных исследований, время, необходимое для важных научных открытий, значительно сокращается, что позволяет быстро расширять границы познания.
- **Расширенное сотрудничество:** такие ИИ-системы, как AlphaFold, демонстрируют, как междисциплинарные подходы, сочетающие ИИ, биологию и физику, способны решать старые проблемы. В итоге, между ИИ-системами может быть сформирована сеть сотрудничества, обеспечивающая обширный обмен идеями и данными в масштабах, недоступных человеку.3. Исследование
- **За пределами человеческой интуиции:** способность ИИ исследовать пространство гипотез с максимальной тщательностью открывает двери к новым открытиям, которые могли бы остаться незамеченными для людей, чьи мыслительные процессы и методологические подходы ограничены их собственными предубеждениями.
- **Трансформация в обработке данных:** подход на базе ИИ кардинально меняет то, как исследователи обрабатывают экспериментальные данные. В подходе на базе ИИ имеют значение все данные, а не только те, которые убедительно подтверждают ожидаемый результат. Даже данные, которые не соответствуют ожиданиям, важны, потому что для обучения ИИ-системы необходимо предоставить как можно больше информации. Это позволит системе генерировать более точные гипотезы и делать более точные прогнозы.

3. Социальные и этические последствия

Наука, основанная на ИИ, вероятно, окажет на общество глубокое влияние:

- **Трансформация здравоохранения:** с помощью ИИ-технологий, которые значительно ускоряют разработку новых лекарственных препаратов и позволяют создавать индивидуальные подходы к лечению, можно спасти или улучшить жизни миллионов людей.
- **Изменения в окружающей среде и климате:** стремительный прогресс в области материаловедения, химии и наук об окружающей среде открывает новые горизонты для открытий, которые могут помочь смягчить последствия изменения климата и улучшить состояние окружающей среды.
- **Этические факторы:** автономность ИИ-систем поднимает вопросы подотчетности, признания заслуг за сделанные открытия и возможного вытеснения людей из области исследований. Существуют опасения, что автономные ИИ-системы смогут разработать и произвести опасные материалы. Важно создать и внедрить этические принципы и меры безопасности, чтобы предотвратить возможное злоупотребление передовыми технологиями в области синтеза материалов и биологии[24]. Необходимо принять надлежащие меры для предотвращения такого злоупотребления.

Исследовательские задачи

1. Вопросы коммуникации: одной из самых серьезных проблем может стать способность понимать ученых и находить с ними общий язык. Это связано с тем, что люди обычно накапливают знания и общаются со своими коллегами на естественном языке, который содержит множество нюансов, аналогий и часто зависит от культурного контекста. Особенно важны следующие моменты:

- **Знания, основанные на здравом смысле:** профессиональные знания возникают из повседневного опыта. Создание общей концептуальной базы знаний о мире для обоснования рассуждений и моделей, а также для обеспечения основы для аналогий.
- **Сотрудничество:** в науке часто приходится работать в команде, поэтому способность продуктивно сотрудничать с другими людьми или даже с ИИ становится ключевым умением.
- **Коммуникация:** ученые общаются между собой, рисуют и используют разнообразные интерактивные средства. ИИ-системы в области науки должны уметь читать и понимать научную литературу и общаться с людьми-партнерами.
- **Модели научного рассуждения и подходы** к разработке механизмов моделирования и получения логических выводов, которые могут расширить возможности человеческого познания для научных открытий [25].

2. Определение пространства гипотез: наука решает задачи со множеством вариантов ответов. В отличие от большинства игр, таких как шахматы или Го, структура и масштаб пространства задач в науке не являются очевидными. Вероятно, они безграничны и могут иметь очень большую размерность. Извлечение или приобретение знаний и их правильное размещение в пространстве научного познания представляют собой нетривиальную проблему. Аналогично, процесс генерации гипотез сталкивается с проблемой определения размерности и масштаба пространства гипотез, в котором он будет осуществляться.

3. Неточности, искажение и воспроизводимость данных:

в некоторых научных областях данные могут быть очень искаженными, неточными и невоспроизводимыми. В биологии такие неточности и искажения считаются неизбежными из-за артефактов, которые возникают в ходе экспериментов, а также внутренней изменчивости экспериментальных образцов. Ученые показали, что значительная часть данных, опубликованных в научных работах, не может быть воспроизведена должным образом в ходе биомедицинских исследований [26, 27]. Это может создавать проблемы для качества выдвигаемых гипотез и их проверки на ранних этапах исследования.

1. Р. Линдсей, Б. Бьюкенен, Э. Фейгенбаум, Дж. Ледерберг (Lindsay R, Buchanan B, Feigenbaum E, Lederberg J.) Экспертная система DENDRAL: практический пример первой экспертной системы для формирования научных гипотез. (DENDRAL: A Case Study of the First Expert System for Scientific Hypothesis Formation.) Artif Intell. 1993;61: с. 209–261.
2. Д. Ленат, Дж. Браун. (Lenat D, Brown J.) Почему AM и ЭВРИСКО, похоже, работают. (Why AM and EURISKO appear to work.) Artif Intell. 1984;23: с. 269–294.)
3. Х. Ван, Т. Фу, И. Ду, В. Гао, К. Хуан, З. Лю и др. (Wang H, Fu T, Du Y, Gao W, Huang K, Liu Z, et al.) Научные открытия в век ИИ. (Scientific discovery in the age of artificial intelligence.) Nature. 2023;620: с. 47–60.)
4. Группа по развитию научно-технического потенциала, Совет по компьютерным наукам и телекоммуникациям, Политика и Глобальные вопросы, Инженерно-физический отдел, Национальные академии наук, инженерии и медицины. ИИ в области научных открытий: материалы семинара под редакцией Р. Пула (Science and Engineering Capacity Development Unit, Computer Science and Telecommunications Board, Policy and Global Affairs, Division on Engineering and Physical Sciences, National Academies of Sciences, Engineering, and Medicine. AI for scientific discovery: Proceedings of a workshop. Pool R, editor.) Вашингтон, округ Колумбия: National Academies Press; 2024 г. doi:10.17226/27457
5. Ю. Гил, М. Гривз, Дж. Хендлер, Х. Хирш. (Gil Y, Greaves M, Hendler J, Hirsh H.) Искусственный интеллект. Расширение возможностей для научных открытий с помощью ИИ. (Artificial Intelligence. Amplify scientific discovery with artificial intelligence.) Science 2014 г.;346: с. 171–172.
6. Совет по развитию науки и техники при Президенте США. PCAST: Доклад президенту о перспективных исследованиях: использование ИИ для решения глобальных задач. (President's Council of Advisors on Science and Technology. PCAST: Report to the president on supercharging research: Harnessing artificial intelligence to meet global challenges.) Office of Scientific and Technical Information (OSTI); апрель 2024 г. doi:10.2172/2481685)
7. А.В. Синиор, Р. Эванс, Дж. Джампер, Дж. Киркпатрик, Л. Сифре, Т. Грин и др. (Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al.) Улучшенное предсказание структуры белков с использованием потенциалов глубокого обучения. (Improved protein structure prediction using potentials from deep learning.) Nature. 2020;577: с. 706–710.)
8. Дж. Джампер, Р. Эванс, А. Притцель, Т. Грин, М. Фигурнов, О. Роннеберггер и др. (Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al.) Высокоточное предсказание структуры белков с помощью программы AlphaFold. (Highly accurate protein structure prediction with AlphaFold.) Nature 2021;596: с. 583–589.
9. М. Пеплоу (Peplow M.) ChatGPT в области химии: ИИ и роботы объединяют усилия для создания новых материалов. (ChatGPT for chemistry: AI and robots join forces to build new materials.) Nature. 2023 г. [процитировано 30 ноября 2024 г.]. doi:10.1038/d41586-023-03745-5)
10. А.Л. Диас, Т. Родригес. (Dias AL, Rodrigues T.) Большие языковые модели управляют работой автоматизированной химической лаборатории. (Large language models direct automated chemistry laboratory.) Nature 2023 г., 624: 530–531.
11. Д.А. Бойко, Р. Макнайт, Б. Клайн, Г. Гомес. (Boiko DA, MacKnight R, Kline B, Gomes G.) Автономные химические исследования с использованием больших языковых моделей. (Autonomous chemical research with large language models.) Nature 2023 г., 624: 570–578.
12. М. Моника, Дж. Борн, Дж. Кадоу, Д. Кристофиделис, А. Дэвид, Д. Кларк и др. (Manica M, Born J, Cadow J, Christofidellis D, Dave A, Clarke D, et al.) Ускорение проектирования материалов с помощью генеративного инструментария для научных открытий. (Accelerating material design with the generative toolkit for scientific discovery.) Npj Comput Mater. 2023 г., 9: 1–6
13. Г. Району, С. Готлиб, Ю. Манор, Г. Пиша, Ю. Харрис, У. Мендлович и др. (Raayoni G, Gottlieb S, Manor Y, Pisha G, Harris Y, Mendlovic U, et al.) Построение гипотез для фундаментальных констант с помощью машины Рамануджана. (Generating conjectures on fundamental constants with the Ramanujan Machine.) Nature 2021 г., 590: 67–73.
14. А. Дэвис, П. Величкович, Л. Бьюсинг, С. Блэквелл, Д. Чжан, Н. Томашев и др. (Davies A, Veličković P, Buesing L, Blackwell S, Zheng D, Tomašev N, et al.) Развитие математики путем управления человеческой интуицией с помощью ИИ. (Advancing mathematics by guiding human intuition with AI.) Nature 2021 г., 600: 70–74.
15. И-Х. Хи. (Не Y-H.) Исследования, основанные на ИИ, в области чистой математики и теоретической физики. (AI-driven research in pure mathematics and theoretical physics.) Nature Reviews Physics. 2024 г., 1–8.
16. К.У. Коули, Д.А. Томас III, Дж.А.М. Ламмисс, Дж.Н. Яворски, К.П. Брин, В. Шульц и др. (Coley CW, ThomasIII DA, Lummiss JAM, Jaworski JN, Breen CP, Schultz V, et al.) Роботизированная платформа для поточного синтеза органических соединений на основе ИИ-планирования. (A robotic platform for flow synthesis of organic compounds informed by AI planning.) Science 2019 г., 365. doi:10.1126/science.aax1566
17. Б. Бургер, П.М. Маффеттон, В.В. Гусев, К.М. Айтчисон, Ю. Бай, Х. Ван и др. (Burger B, Maffettone PM, Gusev VV, Aitchison CM, Bai Y, Wang X, et al.) Мобильный робот-химик. (A mobile robotic chemist.) Nature 2020 г., 583: 237–241.
18. Р.Д. Кинг, Дж. Роуленд, С.Г. Оливер, М. Янг, У. Обри, Э. Бирн и др. (King RD, Rowland J, Oliver SG, Young M, Aubrey W, Byrne E, et al.) Автоматизация науки. (The Automation of Science.) Science 2009 г., 324: 85–89.
19. Р.Д. Кинг, К.Э. Уилан, Ф.М. Джонс, П.Г. Райзер, Ч. Брайант, Ш. Магглтон и др. (King RD, Whelan KE, Jones FM, Reiser PG, Bryant CH, Muggleton SH, et al.) Разработка функциональной геномной гипотезы и проведение экспериментов с помощью робота-ученого. (Functional genomic hypothesis generation and experimentation by a robot scientist.) Nature 2004 г., 427: 247–252.
20. К. Уильямс, Э. Билсланд, А. Спаркс, У. Обри, М. Янг, Л.Н. Солдатова и др. (Williams K, Bilsland E, Sparkes A, Aubrey W, Young M, Soldatova LN, et al.) Более дешевая и быстрая разработка лекарств, подтвержденная появлением новых препаратов для лечения забытых тропических болезней. (Cheaper faster drug development validated by the repositioning of drugs against neglected tropical diseases.) J R Soc Interface. 2015 г.;12: 20141289.
21. Х. Китано. (Kitano H.) Проект Nobel Turing Challenge: создание механизма научных открытий. (Nobel Turing Challenge: creating the engine for scientific discovery.) npj Systems Biology and Applications 2021 г., 7: 29.
22. Китано (Kitano), 2016 г. Искусственный интеллект удостоится Нобелевской премии и не только за создание механизма научных открытий. (Artificial Intelligence to Win the Nobel Prize and Beyond Creating the Engine for Scientific Discovery.) AI Magazine, 37:1 2016 г.
23. Эдвард А. Фейгенбаум, (Feigenbaum, Edward A. (2003 г.). Некоторые проблемы и грандиозные задачи для вычислительного интеллекта (Some challenges and grand challenges for computational intelligence). Journal of the ACM. 50 (1): с. 32–40. doi:10.1145/602382.602400. S2CID 15379263
24. Б.Дж. Виттманн, Т. Алексанян, С. Бартлинг, Дж. Бил, А. Клор, Дж. Диггенс и др. (Wittmann BJ, Alexanian T, Bartling C, Beal J, Clore A, Diggans J, et al.) На пути к ИИ-скринингу порядка синтеза нуклеиновых кислот: процесс, результаты и рекомендации. (Toward AI-resilient screening of nucleic acid synthesis orders: Process, results, and recommendations.) bioRxiv. 2024 г. p. 2024.12.02.626439. doi:10.1101/2024.12.02.626439)
25. Т. Хоуп, Д. Дауни, Д.С. Уэлд, О. Эциони и Э. Хорвиц (Hope, T, Downey, D., Weld, D.S., Etzioni, O. and Horvitz, E. (2023 г.). Переоценка вычислительных возможностей в научных исследованиях. (A Computational Inflection for Scientific Discovery.) Communications of the ACM 66, 8 (Август 2023 г.), с. 62–73. https://doi.org/10.1145/3576896
26. Ф. Принц, Т. Шланге, К. Асадулла ((Prinz F, Schlange T, Asadullah K) (август 2011 г.). Верить или нет: насколько можно доверять опубликованным данным о потенциальных мишенях лекарственных препаратов? (Believe it or not: how much can we rely on published data on potential drug targets?). Nature Reviews. Drug Discovery. 10 (9): 712. doi:10.1038/nrd3439-c1. PMID 21892149.
27. К.Г. Бегли, Л.М. Эллис (Begley CG, Ellis LM) (Март 2012 г.). Разработка лекарств: повышение стандартов доклинических исследований рака. (Drug development: Raise standards for preclinical cancer research). Nature (Комментарий к статье). 483 (7391): с. 531–533. Bibcode:2012Natur.483..531B. doi:10.1038/483531a

Мнение сообщества

Согласно опросу участников сообщества, 32 % респондентов считают, что это в определенной степени актуально. Наиболее полезным применение ИИ считается в таких областях науки, как (1) биология (47 %), (2) физика (14 %), (3) химия (12 %). 26 % респондентов отметили другие области, где ИИ может принести пользу. Когда у респондентов спросили, сможет ли ИИ-система когда-либо совершить открытие, достойное Нобелевской премии, лишь 13 % ответили «никогда», 25 % не были уверены в этом, 11 % предположили, что это может произойти в 2020-х годах, а 45 % считали, что это может случиться к 2050-м годам.

Сильный искусственный интеллект (Artificial General Intelligence, AGI)

В сфере разработки ИИ уже долгое время наблюдается стремление к созданию универсального интеллекта, сравнимого с человеческим (AGI), а недавние успехи в разработке моделей нейронных сетей и их более широкие возможности вызвали активные дискуссии о дальнейших направлениях развития, возможных последствиях и сомнениях в возможности достижения этой цели, которая, по мнению некоторых наблюдателей, теперь кажется вполне реальной.

Основные выводы

- Стремление к пониманию принципов и механизмов интеллекта, которые можно было бы использовать для достижения возможностей уровня человека, всегда было в центре внимания разработчиков ИИ. Еще в 1956 году основатели этого направления ясно обозначили эту цель.
- В начале 2000-х годов стали звучать призывы к реализации масштабных концепций, таких как «ИИ уровня человека» и «сильный искусственный интеллект». Эти призывы возникли на фоне успешного внедрения ИИ в специализированных областях применения и осознания некоторыми специалистами того, что прогресс в достижении долгосрочных целей в этой сфере недостаточен.
- Хотя существуют разногласия по поводу точных определений и ценности некоторых понятий в контексте сильного ИИ, амбициозные цели и связанные с ними идеи, такие как «искусственный интеллект, сопоставимый с человеческим», стали источником вдохновения для многих важных достижений в области ИИ. Эти идеи также определили ключевые направления исследований, которые способствуют разработке более мощных ИИ-систем. С другой стороны, если удастся создать сильный ИИ, это может привести к серьезным социальным последствиям и создать новые угрозы безопасности, включая проблемы, связанные с благополучием и даже выживанием человечества.

ПРЕДСЕДАТЕЛЬСТВУЮЩИЕ

Эрик Хорвиц (Eric Horvitz), Microsoft

Стюарт Рассел (Stuart Russell), Калифорнийский университет, Беркли

Контекст и история

В сфере разработки ИИ долгое время преобладали универсальные принципы интеллекта, подразумевающие, что прогресс в нашем понимании вычислительных основ интеллекта может привести к разработке универсального или общего ИИ. Тест Тьюринга наглядно демонстрирует этот факт: чтобы успешно пройти его, машина должна быть не хуже, а то и лучше человека в некоторых областях, где люди обычно проявляют свои знания и умения.

В 1955 году появилось «Предложение о проведении Дартмутского семинара», которое положило начало эре искусственного интеллекта. Предложение начиналось так: «Исследование должно основываться на предположении, что каждый аспект обучения или другая черта интеллекта в принципе могут быть описаны настолько точно, что можно создать машину для их имитации. Будет предпринята попытка найти способы заставить машины использовать естественные языки, формировать абстракции и концепции, решать задачи, которые в настоящее время под силу только людям, и улучшать самих себя». Эта необычайно амбициозная программа задала тон многим последующим работам участников, включая «Программы со здравым смыслом» Маккарти, «Общее решение проблем» Ньюэлла и Саймона и «Универсальную индукцию» Соломонова. Всего два года спустя, в 1957 году Херб Саймон высказал предположение, что «круг задач, с которыми могут справиться машины, будет соответствовать кругу задач, которые решает человеческий разум». Таким образом, в сфере разработки ИИ всегда ставилась задача создания машин с универсальными или общими интеллектуальными способностями. Значительная часть исследований в области ИИ продолжалась в русле изучения общих принципов интеллекта, включая усилия в области представления, восприятия, логических и вероятностных выводов.

На протяжении десятилетий ученые и исследователи в основном работали над созданием определенных методологий и элементов интеллекта, не уделяя должного внимания их объединению в универсальные или общие системы. Хотя некоторые исследователи в своих областях были полны энтузиазма относительно перспектив применения своих разработок, результаты их внедрения в реальную жизнь часто оказывались неудовлетворительными. Приложения, использующие передовые разработки в области ИИ, были довольно сложными и ненадежными. Отсутствие прогресса в создании универсальных ИИ-систем, способных эффективно работать в реальных условиях, заставило некоторых экспертов в этой области выразить недовольство тем, что общая цель и высокие амбиции ИИ были забыты. Например, в статье Нильса Нильссона «Eye on the Prize» [1] 1995 года говорилось: «ИИ находится на пороге нового этапа, который ускорит усилия по созданию программ с общим, человекоподобным уровнем компетентности». Однако это утверждение скорее было призывом, чем констатацией факта. В начале 2000-х годов идея достижения «человеческого уровня» интеллекта вновь стала актуальной. Например, в 2002 году Марвин Мински организовал семинар на тему «Проектирование архитектур для интеллекта человеческого уровня».

В это же время появился термин «сильный ИИ» (AGI), как отражение высоких устремлений молодого поколения исследователей. Им казалось, что сфера применения ИИ была слишком узкоспециализированной. В начале 2000-х годов методы машинного обучения начали использовать во множестве узкоспециализированных сфер применения, каждая из которых воспринималось как значительный шаг вперед. Но ограниченность этих сфер применения привела к тому, что ученые стали искать более универсальные и эффективные методологии, поскольку принципы машинного обучения и рассуждения могли применяться в самых разных областях.

Изначально сильный ИИ определялся как ИИ, способный сравняться с человеческими когнитивными способностями или превзойти их при решении широкого спектра задач, что соответствовало первоначальным амбициям, заявленным в 1956 году. Хотя эти цели были уже знакомы ведущим исследователям ИИ, использование термина «сильный ИИ» воспринималось многими — как в этой сфере, так и за ее пределами — как новый призыв к амбициозным проектам.

Помимо терминов AGI и «ИИ человеческого уровня», ставших популярными, примерно в то же время появились и другие термины: «универсальный или общий ИИ» и «сильный ИИ». Однако именно термин «сильный ИИ» стал основным как в научных исследованиях, так и в общественных дискуссиях. В популярных книгах и статьях сильный ИИ часто изображается как новая веха, хотя его истоки уходят глубоко в историю развития ИИ. Во многих обсуждениях, в том числе вне сферы исследования ИИ, сильный ИИ связывали как с идеальным, так и мрачным будущим, что отражало различные взгляды, ожидания и тревоги.

Президентская комиссия по долгосрочному развитию ИИ (ранее — Президентская комиссия Ассоциации по развитию искусственного интеллекта) [2] была основана в 2008 году на фоне растущего интереса к сильному ИИ, возрождения высоких амбиций лидеров в области ИИ и активизации общественных дискуссий о будущем ИИ, а также роста количества сфер применения ИИ в открытом мире. Серия встреч и заключительный созыв в Асиломаре были посвящены ключевым вопросам, касающимся осуществимости, последствий, этики и безопасности, а также исследованиям в области создания мощного общего интеллекта, сравнимого с человеческим.

Различные точки зрения на природу сильного ИИ выходят за рамки основного определения методов ИИ («способность соответствовать когнитивным способностям человека или превосходить их при решении широкого спектра задач»). Например, обсуждения сильного ИИ, особенно в популярной прессе, подпитывают предположения о том, что характерной чертой систем сильного ИИ может быть чувствительность или сознательность. Исследователи в области ИИ обычно не склонны к таким предположениям. Они утверждают, что анализ и прогнозирование поведения не зависят от того, можно ли считать машину чувствительной.

Некоторые ученые также считают, что системы сильного ИИ должны

обладать «агентскими» способностями. Это означает, что они, подобно людям, смогут действовать как самостоятельные субъекты, которые воспринимают, изучают, анализируют окружающую среду и активно взаимодействуют с ней для достижения определенных целей. Действительно, способность действовать для достижения целей является фундаментальным когнитивным свойством людей, и некоторые ИИ-системы демонстрировали такие способности в зачаточной форме с самых первых дней своего появления.

Возможно, наиболее сложной для понимания является концепция «автономии» и ее связь с сильным ИИ . Некоторые люди считают, что системы сильного ИИ могут самостоятельно определять свои цели, которые будут полностью отличаться от тех, которые им ставят люди. Хотя это и возможно с точки зрения логики — например, ИИ-система может заменить свои текущие цели на новые, случайно сгенерированные, но не совсем понятно, зачем ей это делать, поскольку это привело бы к тому, что она не смогла бы достичь своих текущих целей. С другой стороны, появление так называемых «инструментальных» подцелей — например, стремление к самосохранению и поиску дополнительных вычислительных и финансовых ресурсов кажется весьма вероятным, поскольку ИИ-системы стремятся достичь своих первоначальных целей. Очевидно, что эта тема вызывает много беспокойства и является активным направлением исследований на протяжении многих лет.

Тот факт, что системы сильного ИИ в целом будут более умными, чем люди, вызывает обоснованные опасения относительно потери контроля над ИИ. Действительно, даже Алан Тьюринг, известный своими работами в области ИИ, высказывал мнение, что «мы должны ожидать, что машины возьмут управление на себя», как только они достигнут уровня человеческого интеллекта. Одним из источников риска является несогласованность, когда цели сильного ИИ не согласуются с предпочтениями человека относительно будущего. Это может происходить по двум причинам: люди неправильно определяют или недооценивают свои предпочтения — так называемая «проблема царя Мидаса», или же ИИ-системы неточно понимают, чего хотят люди [3].

Для некоторых сильный ИИ представляет собой некий опасный рубеж, который мы переступаем на свой страх и риск. Например, в «Отчете Гладстона» [4], который был подготовлен по заказу Государственного департамента США, говорится: «сильный ИИ обычно считается основным источником катастрофического риска из-за потенциальной потери контроля». Другие специалисты используют термин «преобразующий ИИ» [5] для описания ИИ-систем, которые могут привести к краху человеческой цивилизации. Они подчеркивают, что для этого не обязательно наличие полноценного сильного ИИ. Важно отметить, что чувствительность и автономность не являются основополагающими характеристиками сильного ИИ, даже если некоторые исследователи предполагают, что у него могут быть эти черты.

В данный момент нет единого понимания того, что такое сильный ИИ, и нет универсального способа его достижения. Некоторые ученые считают, что «мы поймем это, когда увидим», или что это произойдет само собой, если правильно использовать принципы и механизмы проектирования ИИ-систем. В обсуждениях сильный ИИ может упоминаться как достижение определенного порога возможностей и универсальности. Однако существуют мнения, что это неточное определение, и что интеллект лучше охарактеризовать как существование в непрерывном, многомерном пространстве. Некоторые исследователи, например, [6], полагают, что из-за отсутствия ясного понимания того, что такое сильный ИИ, он не может быть основной целью исследований в области ИИ. Они заявляют, что человеческий разум многогранен, и машины могут достичь превосходства в одних сферах, но при этом оставаться ограниченными в других. Также зачастую остаются неясными критерии, на основе которых происходит сравнение, включая сведения о том, каких людей считать эталонами, и какую подготовку они прошли.

Некоторые исследователи считают, что создание сильного ИИ не должно быть конечной целью исследований в области ИИ. Они утверждают, что стремление «соответствовать человеческим когнитивным способностям или превосходить их» не обязательно приведет к созданию инструментов, которые улучшат или дополнят возможности человека. Вместо этого, по их мнению, краткосрочная финансовая выгода от сильного ИИ может быть связана с тем, что он сможет заменить людей в большинстве профессий в экономическом секторе. Более того, многие из предполагаемых преимуществ сильного ИИ в науке, здравоохранении, образовании и других областях могут быть достигнуты с помощью более специализированных инструментов, таких как AlphaFold2. Тем не менее, создание сильного ИИ стало канонической целью для амбициозных компаний, занимающихся ИИ. Например, Сэм Альтман, занимающий должность генерального директора OpenAI, поделился своим видением: «Наша цель — создать сильный ИИ и обеспечить его безопасность..., а также раскрыть его потенциал» [7].

Текущая ситуация и тенденции

Развитие возможностей ИИ и результаты его работы за последнее десятилетие говорят о впечатляющих достижениях. Эти достижения указывают на то, что ИИ постепенно достигает человеческого уровня или даже превосходит его в выполнении различных задач, что подтверждается серией отчетов по индексам ИИ [8] и Международным отчетом по безопасности ИИ 2025 года [9].
Первые успехи были достигнуты в области распознавания речи и объектов на изображениях, за которыми последовали успехи в машинном переводе. С развитием генеративного ИИ появились инструменты для синтеза высококачественных изображений и голосовых сообщений. В 2022 году был достигнут значительный прогресс в области генерации языка. В 2023 году были достигнуты впечатляющие результаты благодаря применению мультимодальных моделей, которые могут обрабатывать информацию в различных форматах: текст, изображения, аудио (в качестве входных и выходных данных) и даже работать с физическими объектами. В 2024 году

мы стали свидетелями значительного прогресса в области рассуждений. Одним из таких достижений стал успех в решении задачи абстрактного рассуждения, известной как ARC-AGI. До 2024 года ИИ-системы не могли справиться с этой задачей.

В последнее время значительный прогресс заметен и в сфере нейросетевых моделей. Он связан с применением алгоритмов, которые позволяют моделям размышлять в процессе работы. Эти алгоритмы обучаются использовать цепочки внутренних размышлений, основанные на более сложном мыслительном процессе, присущем человеку. В отличие от предыдущих моделей, которые за константное время выполнения алгоритма соотносили входные данные с ответом, что можно сравнить с быстрыми интуитивными человеческими реакциями, описываемыми как концепция мышления «система 1», современные технологии вычисления выводов во время диагностики проблемы позволяют ИИ анализировать длинные цепочки рассуждений для поиска ответов на сложные вопросы. Алгоритмы анализируют и оценивают множество вариантов, подобно тому, как это делают люди в процессе более вдумчивого мышления. Такая концепция мышления называется «система 2». Эти модели также требуют гораздо больше вычислительных ресурсов во время работы, что может существенно увеличить как энергетические, так и финансовые затраты на их развертывание, не говоря уже о стоимости обучения.

Наряду с физическими возможностями, способность к логическому рассуждению долгое время считалась ключевым отличием человеческого интеллекта. Но, похоже, что это отличие постепенно стирается. В настоящее время существуют ИИ-системы, которые способны конкурировать с лучшими представителями человечества во многих широко используемых тестах, ориентированных на проверку знаний и логического мышления. С другой стороны, такие системы могут давать элементарные сбои, что поднимает серьезные вопросы о том, как интерпретировать их успехи [10]. К примеру, некоторые современные модели сталкиваются с серьезными проблемами при решении математических задач, представленных в форме текстовых описаний и изображений, которые кажутся людям элементарными [11]. Как и прежде, большинство современных моделей испытывают сложности в области пространственного и геометрического мышления, а также в понимании деталей изображений, особенно при мультимодальном вводе [12]. Планирование, как особая форма рассуждения, все еще недостаточно развито, особенно когда речь заходит о более длительных горизонтах планирования и о планировании практических шагов, а также о помощи людям и со стороны людей в физическом мире. Тем не менее, исследования в этой области активно финансируются, и такие модели могли бы обеспечить компетентность на уровне человека при решении широкого спектра задач. Это имело бы огромную экономическую ценность, но при этом возникают вопросы о влиянии на общество.

Исследовательские задачи

Несмотря на впечатляющий прогресс в области крупномасштабных моделей глубокого обучения, таких как трансформеры, современные системы ИИ в целом не считаются обладающими всеми возможностями, указываемыми в большинстве определений сильного ИИ. Учитывая контекст этих текущих ключевых ограничений, исследования и разработки могут быть сосредоточены на следующих направлениях:

Архитектуры, отличные от архитектуры трансформера: хотя стандартная архитектура трансформера продемонстрировала впечатляющие результаты, она имеет ряд существенных ограничений. К ним относятся фиксированные связи между слоями, отсутствие явной памяти, неспособность обучаться и реагировать на текущую обратную связь от окружающей среды, а также неэффективность и трудности при решении сложных логических задач. Изучение новых архитектур могло бы привести к новым способам определения интеллекта, сравнимого с человеческим. Можно выделить следующие направления исследований: развитие способностей к рассуждению и обобщению за счет применения инновационных архитектур, изучение гибридных архитектур, которые сочетают в себе архитектуру трансформера и другие модели, такие как графические нейронные сети, агенты обучения с подкреплением или системы символической логики.

Долгосрочное планирование и построение логического вывода: современные ИИ-модели сталкиваются с трудностями в долгосрочном планировании и не способны обеспечить надежную иерархию элементов аргументации. В отличие от людей, они не обладают сильным предвидением, испытывают сложности с многоступенчатым решением проблем и не склонны эффективно и точно разбивать сложные задачи на подзадачи. В отличие от классических ИИ-систем, созданных в 1960-х годах и позже, они не могут гарантировать правильность своих шагов аргументации. Тем не менее, недавние успехи в масштабировании вычислительных ресурсов во время вывода с применением методов обучения с подкреплением, позволяющих научиться выстраивать цепочки логических выводов, являются одним из направлений исследований по наделению систем на основе нейросетей способностями к более эффективному планированию. Эти достижения требуют значительных усилий, включая предварительную оценку рисков и затрат для каждого шага в плане. Также необходимо проверить, будет ли каждый шаг соответствовать человеческим ценностям и условиям задачи.

Обобщение за пределами обучающих данных: хотя большие языковые модели демонстрируют впечатляющие результаты, их способность к обобщению и решению новых задач все еще остается неясной. Они могут быть легко введены в заблуждение с помощью манипуляционных действий, и им часто не хватает способности гибко применять знания в различных областях. Возможно, потребуется рассмотреть такие системы представления, как программы, которые более информативны, чем схемы, но для этого не хватает эффективных механизмов.

Непрерывное обучение: в отличие от людей, большие языковые модели не учатся непрерывно на собственном

опыте, а учатся на основе жесткой парадигмы предварительного обучения и дообучения. Необходимо исследовать механизмы, которые позволят системам сохранять и обновлять знания в постоянном потоке, а не полагаться только на автономные, статичные методы обучения. Переход к архитектурам и методикам обучения, способствующим непрерывному обучению на протяжении всей жизни, является важным шагом вперед. Эти возможности включают в себя новые формы самоконтроля и самообучения посредством моделирования и исследования границ компетенций и понимания окружающей среды или концептуальных проблем [13].

Запоминание и узнавание: для создания сильного ИИ может потребоваться разработка механизма запоминания и контекстно-зависимого узнавания, похожего на человеческий. Этот механизм должен включать в себя структурированную эпизодическую память. В отличие от людей, трансформеры не обладают постоянной структурированной памятью, которая эффективно накапливает информацию в течение длительного времени. Предпринимаются усилия по дополнению LLM механизмами памяти, как правило, с помощью внешнего оборудования. В этом направлении исследований существует множество возможностей для дальнейшего развития.

Причинно-следственные и контрфактические рассуждения: хотя ИИ-модели способны выявлять закономерности в больших массивах данных, они с трудом справляются с установлением причинно-следственных связей и контрфактическиими рассуждениями. Понимание причинно-следственных связей — это ключ к принятию обоснованных решений и совершению научных открытий. Важным направлением исследований является изучение причинно-следственных рассуждений с использованием больших языковых моделей.

Воплощение и взаимодействие с реальным миром: человеческий интеллект развивается благодаря разнообразным сенсорно-моторным взаимодействиям с окружающим миром. Однако современные мультимодальные модели, по всей видимости, не способны полностью осознать физическую реальность, что затрудняет их восприятие, мышление и эффективное взаимодействие с миром. В этом контексте особый интерес представляют исследования, направленные на обучение ИИ-моделей в насыщенных интерактивных средах, таких как робототехника и виртуальные миры. Такие исследования позволяют создать более глубокое понимание реальности, которое охватывает множество разнообразных модальностей, включая видео, аудио и сенсорные данные.

Согласованность, интерпретируемость и безопасность: в стремлении к созданию более совершенного интеллекта, который сможет конкурировать с человеческим, важной задачей остается обеспечение соответствия ИИ-систем человеческим ценностям и их интерпретируемости. ИИ-модели, представляющие собой «черные ящики», в том числе трансформеры, зачастую выдают результаты, которые трудно объяснить. И это вызывает проблемы с безопасностью и доверием. Кроме того, LLM обучаются через имитацию, анализируя вербальное поведение, максимально похожее на поведение человека. Вербальное поведение человека имеет конкретную задачу, а именно достижение цели, начиная от самосохранения и поиска пары и заканчивая обретением богатства и власти. Поэтому вполне вероятно, что LLM фактически ставят себе похожие или связанные цели, к достижению которых они могут стремиться самостоятельно. Необходимо проведение исследований по согласованию ценностей, определению ограничений и линий поведения, обеспечивающих безопасную работу, и, в более широком смысле, по разработке мер безопасности, гарантирующих соответствие высокоэффективных ИИ-систем намерениям человека.

Понимание и управление влиянием на общество: по мере роста возможностей ИИ-систем повышается важность исследований в области безопасности, упреждающего управления и активного мониторинга влияния ИИ на людей и общество. Вместо того чтобы полагаться на возможности рыночных сил в достижении положительных результатов, сообщество исследователей ИИ может на самых ранних этапах начать взаимодействие и затем поддерживать контакт с представителями власти и гражданскими активистами, чтобы помочь сформировать возможности и способы использования и контроля ИИ [14].

1. Нильссон, Н. (N. Nilsson) (1995 г.). Помнить о главном. (Eye on the Prize). AI Magazine, 16(2), 9. https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1129
2. Ассоциация по развитию искусственного интеллекта (Association for the Advancement of AI) (2009 г.). Президентская комиссия Ассоциации по развитию искусственного интеллекта по долгосрочным перспективам исследований в области ИИ, 2008-2009 гг. (AAAI Presidential Panel on Long-Term AI Futures). https://aaai.org/about-aaai/aaai-presidential-panel-on-long-term-ai-futures-2008-2009/ Диттерих, Т. и Хорвиц, Э. (T. Dietterich and E. Horvitz) (2015 г.). Рост опасений по поводу ИИ: размышления и направления. (Rise of Concerns about AI: Reflections and Directions). Communications of the ACM, 58(10). https://dl.acm.org/doi/pdf/10.1145/2770869
3. Харрис, Э., Харрис, Дж. и Билл, М. (E. Harris, J. Harris, and M. Beall) (2024 г.). План действий по повышению безопасности и защищенности высокоразвитого ИИ. (An Action Plan to increase the safety and security of advanced AI). Gladstone AI. https://www.gladstone.ai/action-plan
4. Карнофски, Х. (H. Karnofsky) (2016 г.). Об истории наших взглядов на высокоразвитый искусственный интеллект. (Some background on our views regarding advanced artificial intelligence). Open Philanthropy. https://www.openphilanthropy.org/research/some-background-on-our-views-regarding-advanced-artificial-intelligence/
5. Тогелиус, Дж. (J. Togelius) (2024 г.). Сильный искусственный интеллект. (Artificial General Intelligence). MIT Press.
6. Мурджиа, М. (M. Murgia) (2023 г.). Руководитель OpenAI в поисках новых средств на создание «сверхинтеллекта». (OpenAI chief seeks new Microsoft funds to build 'superintelligence'). Financial Times, 13 ноября.
7. Под ред. Р. и Дж. Кларков (R. and J. Clark, eds.) (2017-2024 гг.). Отчеты AI Index (AI Index Reports). Stanford Institute for Human-Centered Artificial Intelligence. https://aiindex.stanford.edu/
8. Бенжио, Й. и др. (Y. Bengio and others) (2025 г.). Международный отчет по безопасности ИИ (International AI Safety Report). Правительство Великобритании. International AI Safety Report 2025 - GOV.UK
9. Маркус, Г. (G. Marcus) (2025 г.). Сильный ИИ против «выдающегося, но поверхностного интеллекта». (AGI vs "broad, shallow intelligence"). Substack. https://garymarcus.substack.com/p/agi-versus-broad-shallow-intelligence
10. Черian, А., Пэн, К., Лохит, С., Маттисен, Дж., Смит, К., Тененбаум, Дж. (A. Cherian, K. Peng, S. Lohit, J. Matthiesen, K. Smith, J. Tenenbaum) (2024 г.). Оценка больших моделей языка и зрения на основе математических олимпиад для детей. (Evaluating Large Vision-and-Language Models on Children's Mathematical Olympiads). NeurIPS 2024. https://arxiv.org/abs/2406.15736
11. Балачандран, В., Чэнь, Ц., Джоши, Н., Нуши, Б., Паланги, Х., Салинас, Э., Винет, В., Уоффинден-Люэй, Дж., Юсефи, С. (V. Balachandran, J. Chen, N. Joshi, B. Nushi, H. Palangi, E. Salinas, V. Vineet, J. Woffinden-Luey, S. Yousefi). (2024 г.) Эврика! Оценка и понимание больших базовых моделей. (Eureka: Evaluating and Understanding Large Foundation Models). arXiv 2409.10566, сентябрь 2024 г. https://arxiv.org/abs/2409.10566
12. Ли, Н., Цай, Ц., Шварцшильд, А., Ли, К., Папаилиопоулос, Д. (N. Lee, Z. Cai, A. Schwarzschild, K. Lee, D. Papailiopoulos) (2024 г.). Самосовершенствующиеся трансформеры решают задачи «от простого к сложному» и обобщения длины. (Self-Improving Transformers Overcome Easy-to-Hard and Length Generalization Challenges). arXiv 2502.01612, февраль 2025 г. https://arxiv.org/abs/2502.01612
13. Хорвиц, Э., Конитцер, В., Макилрайт, Ш. и Стоун, П. (E. Horvitz, V. Conitzer, S. McIlraith, and P. Stone) (2024 г.). Сейчас, после и всегда: 10 приоритетов для исследований, теории и практики ИИ (Now, Later, and Lasting: 10 Priorities for AI Research, Policy, and Practice). Communications of the ACM, 67(6). https://cacm.acm.org/opinion/now-later-and-lasting-10-priorities-for-ai-research-policy-and-practice/

Мнение сообщества

Ответы на вопросы о сильном ИИ в рамках нашего опроса показывают, что мнения о развитии сильного ИИ и об управлении им разнятся. Большинство респондентов (77 %) отдают приоритет созданию ИИ-систем с приемлемым соотношением рисков и преимуществ, а не погоне за сильным ИИ (23 %). Однако дискуссии о возможности создания сильного ИИ и об этических аспектах достижения искусственным интеллектом уровня человека по-прежнему продолжаются.

Значительное большинство респондентов (82 %) считают, что системы с сильным ИИ, если их разработку выполняют частные компании, должны принадлежать государству. Это отражает опасения, связанные с глобальными рисками и ответственностью за соблюдение этических принципов. Но несмотря на это большая часть респондентов (70 %) не согласна с тем, что исследования в области сильного ИИ следует прекратить до тех пор, пока не будут созданы полнофункциональные механизмы обеспечения безопасности и контроля. По-видимому, эти ответы означают, что исследования по данной теме необходимо продолжать, но с соблюдением некоторых мер предосторожности.

Большинство респондентов (76 %) утверждают, что «масштабирование имеющихся подходов к развитию ИИ» для достижения сильного ИИ «маловероятно» или «крайне маловероятно» приведет к успеху, и это порождает сомнения в том, способны ли современные парадигмы машинного обучения создать сильный ИИ.

В целом ответы респондентов указывают на их приверженность осторожному, но прогрессивному подходу: исследователи в области ИИ отдают приоритет безопасности, соблюдению этических принципов в процессе управления, совместному пользованию преимуществами и постепенному внедрению инноваций, выступая за коллективное и ответственное развитие, а не просто стремясь получить сильный ИИ.

ИИ: восприятие и реальность

Как следует оспаривать преувеличенные заявления о способностях ИИ и формировать реалистичные ожидания?

ПРЕДСЕДАТЕЛЬСТВУЮЩИЙ

Родни Брукс (Rodney Brooks), Массачусетский технологический институт

Основные выводы

- На фоне непрерывного появления новых и важных технологий в течение последних 70 лет многие инновации в области ИИ вызывают чрезмерный ажиотаж.
- Как и в случае с другими технологиями, тенденции создания ажиотажа соответствуют общей картине циклов ажиотажа Gartner.
- Текущая фаза цикла ажиотажа для большинства людей в мире, возможно, является знакомством с ИИ, и у этих людей отсутствуют инструменты для проверки истинности многих заявлений.

Контекст и история

Искусственный интеллект (ИИ) – это область знаний, изучающая синтез и анализ вычислительных агентов, которые действуют разумно [6]. Со времен проведенного в 1956 году семинара, на котором был сформулирован термин «искусственный интеллект» и задан курс на использование ИИ в качестве основного направления исследований и обучения на первых факультетах компьютерных наук, ИИ неоднократно проходил различные фазы цикла ажиотажа. Но этот ажиотаж рано или поздно сходил на нет, поскольку его суть состоит в том, что он выходит за пределы реальности. Спустя десятилетия это привело к спаду интереса к ИИ («зиме ИИ»), когда выделение финансирования на эту технологию было прекращено полностью или применительно к определенным направлениям, таким как нейросети или робототехника.

Исследование тенденций общественного восприятия ИИ за 30 лет, проведенное в 2017 году, показало, что количество дискуссий в отношении ИИ резко увеличилось с 2009 года, а публикации на эту тему в СМИ были более оптимистичными, нежели пессимистичными [4]. Исследование также показало рост ожиданий относительно применения ИИ в здравоохранении и образовании со временем. Авторы исследования также сделали вывод о растущей обеспокоенности по поводу потери контроля над ИИ, этических последствий и негативного влияния ИИ на работу.

Возможно, последние годы отличаются от более ранних периодов тем, что ажиотаж вышел за рамки научных конференций, докладов и журналов и перешел в традиционные СМИ и социальные сети. «ИИ» и «искусственный интеллект» стали обычными словами для людей, не имеющих отношение к технике, и обычной темой для лидеров почти всех государств. Впервые у органов власти появилась политика в области ИИ.

Одна из проблем заключается в том, что на самом деле ИИ – это широкое понятие, которое можно использовать множеством различных способов. Но в повседневной речи сейчас этот термин используется так, как будто он означает какой-то один предмет. В 2024 году Нараянан и Капур выпустили книгу [5], в которой сравнили термин «искусственный интеллект» с языком транспортной отрасли, где один термин «транспортное средство» может, к примеру, означать велосипеды, скейтборды, атомные подводные лодки, ракеты, автомобили, 18-колесные грузовики, контейнеровозы и т. п. При таких обстоятельствах невозможно почти ничего сказать о «транспортных средствах» и их возможностях, поскольку все сказанное будет верно лишь для небольшой части всех «транспортных средств». Такое отсутствие различий усложняет проблему ажиотажа, так как конкретные заявления становятся чрезмерно обобщенными.

Ажиотаж также создает у обычных людей определенные ожидания. Многие боятся потерять работу из-за ИИ в краткосрочной перспективе. Как следствие, социологи начинают работать над проблемой отсутствия работы, к примеру, для водителей грузовиков, на смену которым придет ИИ [6], основываясь на прогнозах относительно ИИ (в данному случае – беспилотных грузовиков) и его внедрения, которые оказываются чрезмерно оптимистичными. Но в пределах периода прогнозирования внедрение беспилотных грузовиков не ожидается.

Ажиотаж в ответ на запуск технологии не ограничивается ИИ. Компания Gartner, занимающаяся бизнес-аналитикой, намеренно взяла за правило использовать графическое представление ажиотажа в виде пяти следующих фаз, являющихся едиными для множества технологий: (1) запуск технологии (2) пик завышенных ожиданий, (3) пропасть разочарования, (4) склон просветления и (5) плато продуктивности. Компания проанализировала многие технологии с помощью данной схемы, в том числе квантовые компьютеры, блокчейн, беспилотные транспортные средства, нанотехнологии и т. п. В ноябре 2024 года специалисты Gartner [1] пришли к выводу, что ажиотаж в связи с сильным ИИ только что прошел пиковую фазу и пошел на спад.

Вопрос для экспертов в области ИИ состоит в том, как реагировать на этот ажиотаж, как критически оценивать его и как помочь другим понять, какая оценка является завышенной, сохраняя при этом собственную интеллектуальную скромность и порядочность. Это сложно сделать в разгар эпохи преувеличенных заявлений, и зачастую заниматься тщательным анализом научных аргументов прошлого приходится историкам будущего.

Историк Томас Хэй попытался провести такой анализ практически в режиме реального времени в серии статей в журнале Communications of the ACM. В статье [2] он приводит «посмертный» разбор последствий чрезмерного ажиотажа в отношении ИИ, который привел к явлению, известному как «зима ИИ» в 1980-е гг. Хэй сделал следующий вывод: «Последствия лопнувшего пузыря ажиотажа привели к наступлению настоящей зимы ИИ в конце 1980-х гг.». В статье [3] исследователь сравнивает сегодняшний ажиотаж с ажиотажем прошлого и заключает: «От двигателей логики к двигателям ерунды?».

Исследовательские задачи

Многие из тех, кто десятилетиями работает над ИИ, видят, что зачастую публичные заявления новичков в этой сфере не имеют ничего общего с реальностью, и тем самым сталкиваются с проблемой сохранения честности в такой ситуации.

Главный вопрос заключается в том, могут ли экспертные заключения и исследования, рецензированные коллегами, с учетом динамики социальных сетей и погони за кликами, помочь снизить ажиотаж и повлиять на вызванное им искажение общего понимания стадии развития ИИ и его потенциала в течение одного года, пяти лет, десяти лет и т. п.

Если мы сейчас не обсуждаем это, как мы сможем это изменить?

1. Чандрасекаран, А. (Chandrasekaran, A.) (2024 г.). Движущий фактор ажиотажа вокруг генеративного ИИ. (What's Driving the Hype Cycle for Generative AI), 2024 г. https://www.gartner.com/en/articles/hype-cycle-for-genai
2. Хэй, Т. (Haigh, T.) (2024 г.). Как завершился ажиотаж вокруг ИИ. (How the AI Boom Went Bust). Выпуск 67, № 2, с. 22-26.
3. Хэй, Т. (Haigh, T.) (2025 г.). Искусственный интеллект тогда и сейчас. (Artificial Intelligence Then and Now). Выпуск 68, № 2, с. 24-29.
4. Фаст, И. и Хорвиц, Э. (Fast, E. and Horvitz, E.) (2017 г.). Долгосрочные тенденции общественного восприятия искусственного интеллекта. (Long-term trends in the public perception of artificial intelligence). Ассоциация по развитию искусственного интеллекта, 2017 г.: материалы 31-й Конференции по искусственному интеллекту Ассоциации по развитию искусственного интеллекта, 4 февраля 2017 г., с. 963-969. https://ojs.aaai.org/index.php/AAAI/article/view/10635
5. Нараянан, А. и Капур, С. (Narayanan, A. and Kapoor, S.) (2024 г.). ИИ как средство от всех болезней: на что способен и не способен искусственный интеллект и как понять разницу. (AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference). Princeton University Press.
6. Пул, Д. Л. и Мауорт, А. К. (Poole, D. L. and Mackworth, A. K.) (2023 г.). Искусственный интеллект: основы вычислительных агентов. Издание 3. (Artificial Intelligence: Foundations of Computational Agents). Cambridge University Press.
7. Ван, С., Мак, Э. А., Ван Фоссен, Дж. А., Медвид, М., Коттен, С. Р., Чан, Ч. С., Манн, Дж., Миллер, С. Р., Саволайнен, П. Т. и Бейкер, Н. (Wang, S., Mack, E. A., Van Fossen, J. A., Medwid, M., Cotten, S. R., Chang, C. H., Mann, J., Miller, S. R., Savolainen, P.T. and Baker, N.) (2023 г.). Оценка альтернативных занятий для водителей грузовиков в эпоху появления беспилотных транспортных средств (Assessing alternative occupations for truck drivers in an emerging era of autonomous vehicles). Transportation Research Interdisciplinary Perspectives, выпуск 19, май. https://doi.org/10.1016/j.trip.2023.100793

Мнение сообщества

Опрос дает представление о том, как сообщество реагирует на тему «ИИ: восприятие и реальность». Обобщенные результаты опроса приводятся ниже. Ответы на вопросы по данной теме дали 36 % респондентов. Результаты с разбивкой по вариантам ответов:

1. Насколько актуальной является данная тема для вашего исследования? Суммарно 72 % респондентов отметили, что данная тема является актуальной: в некоторой степени актуальной (24 %), актуальной (29 %) или весьма актуальной (19 %).

2. Настоящее восприятие возможностей ИИ соответствует реальной ситуации в сфере исследований и разработок ИИ. Суммарно 79 % респондентов не смогли согласиться с данным утверждением: не согласны (47 %) или категорически не согласны (32 %).

3. Каким образом несоответствие восприятия возможностей ИИ и реальной ситуации препятствует проведению исследований? Отвечая на данный вопрос, 74 % респондентов согласились, что направления исследований в области ИИ обусловлены ажиотажем, 12 % считают, что в результате этого страдают теоретические исследования в области ИИ, и 4 % убеждены, что все меньше студентов интересуются научными исследованиями.

4. Должно ли сообщество внедрить инициативу по борьбе с ажиотажем посредством проверки достоверности заявлений, касающихся ИИ? «Да» ответили 78 % респондентов, из которых 51 % согласны с необходимостью такой инициативы, а 27 % – полностью согласны.

5. Должно ли сообщество внедрить инициативу по организации общественных дискуссий на тему «ИИ: восприятие и реальность» с возможностью обеспечения открытого доступа к видеозаписям таких дискуссий? «Да» ответили 74 % респондентов, из которых 46 % согласны с необходимостью такой инициативы, а 28 % – полностью согласны.

6. Должно ли сообщество внедрить инициативу по созданию и сопровождению хранилища прогнозных показателей развития способностей ИИ и регулярной проверке точности содержащихся в нем данных? «Да» ответили 59 % респондентов, из которых 40 % согласны с необходимостью такой инициативы, а 29 % – полностью согласны.

7. Должно ли сообщество внедрить инициативу по просвещению общественности (включая СМИ и виртуальные каналы) в области разнообразия методов и направлений исследований ИИ? «Да» ответили 87 % респондентов, из которых 45 % согласны с необходимостью такой инициативы, а 42 % – полностью согласны.

8. Должно ли сообщество внедрить инициативу по разработке методики формирования годового рейтинга зрелости технологии ИИ в части решения нескольких задач? «Да» ответили 61 % респондентов, из которых 42 % согласны с необходимостью такой инициативы, а 19 % – полностью согласны.

Поскольку респонденты сами принимали решение об участии в опросе по данной теме (около трети всех респондентов), необходимо учитывать фактор необъективности. Существенная часть респондентов (хотя и не 100 %) отмечала, что современное восприятие возможностей ИИ является преувеличенным, что это действительно влияет на данную область знаний, и что специалистам необходимо найти способ донесения до людей информации о реальном положении дел.

Разнообразие подходов к изучению ИИ

Важно поощрять и поддерживать исследования различных парадигм ИИ – как старых, так и новых. Сюда относятся разнообразные старые и новые методики (не только нейросети), междисциплинарное сотрудничество и социальные последствия, которые необходимо учитывать.

ПРЕДСЕДАТЕЛЬСТВУЮЩИЙ

Питер Стоун (Peter Stone),
Техасский университет
в Остине и Sony AI

Основные выводы

- Исторически область знаний, связанная с ИИ, охватывала множество различных методик и исследовательских парадигм одновременно.
- Существует риск того, что нынешняя конвергентность данной области с упором на нейронные подходы может препятствовать внедрению инноваций.
- Авторы призывают активно поддерживать исследования классических (ненейронных) подходов, а также исследования, сочетающие нейронные и другие подходы и интегрирующие различные парадигмы в более сложные когнитивные архитектуры. Авторы особенно приветствуют поддержку креативных исследований абсолютно новых парадигм, которые могут стать ключевым фактором в преодолении ограничений существующих подходов.

Контекст и история

В сфере исследования ИИ уже давно существуют отдельные подсообщества, глубоко изучающие различные подходы к воспроизведению интеллекта в компьютерах. Иногда такие подсообщества формировались на основе подходов, таких как планирование, эволюционные вычисления, удовлетворение ограничений или комбинаторный поиск. В других случаях формирование подсообществ осуществлялось на основе областей применения, например компьютерного зрения, обработки естественного языка или робототехники.

Как правило, существуют определенные области, которые являются более «модными» по сравнению с другими, но, несмотря на споры в отношении определенных подходов, сообщество в целом хорошо относится к проявлению терпимости и даже поощряет разнообразие подходов. Действительно, можно утверждать, что нынешний расцвет генеративного ИИ, основанного на нейросетях, является результатом такого проявления терпимости. Введение такого понятия, как «нейросети», и их изучение началось еще до появления термина «искусственный интеллект» в 1950-х гг., а в 1960-х гг. было проведено множество исследований по данной теме. Но нейросети не оправдали уровня ажиотажа, возникшего вокруг них в те годы. В результате наступило время, когда большинство специалистов в данной области считали «связность» тупиковым направлением для исследований. Но подсообщество упорно продолжало работать, и в конце концов настал их звездный час (если не сказать больше).

Текущая ситуация и тенденции

Однако сейчас существует риск того, что традиция придерживаться разнообразных подходов будет утрачена. Вполне вероятно, что одно или несколько из ныне немодных направлений исследований также сможет достичь успеха. Но на сегодняшний день мы не знаем, какое именно. В результате доминирования нейронных подходов многие другие подходы теряют популярность или даже переклассифицируются как не относящиеся к ИИ (заголовок недавней статьи в журнале IEEE Spectrum намекает на то, что классический поиск – это не ИИ: https://spectrum. ieee. org/chip-design-ai). И действительно, как сообщество мы, кажется, рискуем отбить у новичков желание развивать альтернативные подходы.

И мы считаем это ошибкой. Наоборот, для долгосрочного процветания данной области знаний важно найти способ поддержать тех, кто не побоится пойти против мнения большинства, даже если их статьи будут реже публиковать или цитировать. Но некоторые из их статей могут оказаться чрезвычайно успешными. И даже (или в особенности) применительно к людям, сконцентрировавшимся на нейросетях, – по нашему мнению, важно, чтобы они имели представление об альтернативных парадигмах. Это необходимо для того, чтобы не препятствовать инновациям, изобретая велосипед.

Это не значит, что мы закрываем глаза на игнорирование прогресса в области генеративного ИИ, основанного на нейросетях. Возможно, что это крупнейшая революция в области ИИ, и она заслуживает огромного внимания. Просто не абсолютно всего внимания.

Исследовательские задачи

По нашим прогнозам, в будущем произойдут некоторые прорывные достижения в других областях – непосредственно усилиями таких областей или в сочетании с нейронными и другими классическими методами.

Например, исследования возможностей планирования больших языковых моделей показывают, что они действительно не способны эффективно рассуждать и планировать [Валмикам и др. (Valmeekam et al.), 2023 г.; Валмикам и др. (Valmeekam et al.), 2024 г.]. Возможно, для того, чтобы LLM могла создавать разумные планы, ее необходимо объединить с некой системой принятия решений путем манипулирования символами. В этом направлении применяются нейросимволические подходы. Аналогичным образом, конформное прогнозирование [Ангелопулос и Бейтс (Angelopoulos and Bates), 2023 г.] представляет собой попытку внедрения вероятностных рассуждений в нейронные модели.

Мы призываем ИИ-сообщество также обратить внимание на возможности и ограничения нейронных подходов, активно поддержав исследования классических (ненейронных) подходов, таких как поиск, оптимизация, удовлетворение ограничений и каузальные рассуждения, а также исследования, сочетающие нейронные подходы с символическими и вероятностными и интегрирующие различные парадигмы в более сложные когнитивные архитектуры. Авторы особенно приветствуют поддержку креативных исследований абсолютно новых парадигм, которые могут стать ключевым фактором в преодолении ограничений существующих подходов.

Такая поддержка может иметь форму семинаров, посвященных этим исследованиям, и также должна включать в себя финансирование таких приоритетных направлений. Нам следует уделить особое внимание поиску путей поощрения и поддержки исследователей, придерживающихся как старых, так и новых подходов и заинтересованных в изучении новых идей с новых точек зрения, а также возможностей для пересечения различных новых и существующих подходов, даже если поначалу им будет сложно получить такую поддержку.

1. Ангелопулос, Анастасиос Н. и Бейтс, Стивен (Angelopoulos, Anastasios N. and Stephen Bates). Конформное прогнозирование: аккуратное внедрение. Основы и тенденции машинного обучения. (Conformal Prediction: A Gentle Introduction", Foundations and Trends in Machine Learning). Выпуск 16, № 4, с. 494-591 http://dx.doi.org/10.1561/2200000101 (2023)
2. Валмикам, Картик и др. (Valmeekam, Karthik, et al.). О возможностях планирования больших языковых моделей (критическое исследование на основе предложенного бенчмарка) (On the planning abilities of large language models (a critical investigation with a proposed benchmark). Препринт arXiv:2302.06706 (2023 г.).
3. Валмикам, Картик, Стекли, Кайя и Камбхампати, Суббарао (Valmeekam, Karthik, Kaya Stechly, and Subbarao Kambhampati). Большие языковые модели по-прежнему не умеют планировать. А что насчет больших моделей рассуждений? Предварительная оценка o1 OpenAI на основе PlanBench (LLMs Still Can't Plan; Can LRMs? A Preliminary Evaluation of OpenAI's o1 on PlanBench). Препринт arXiv:2409.13373 (2024 г.).

Мнение сообщества

Ответы на вопросы по данной теме дали 57 % респондентов (176 человек). Результаты с разбивкой по вариантам ответов:

1. Насколько актуальной является данная тема для вашего исследования? Суммарно 92 % респондентов отметили, что данная тема является актуальной: в некоторой степени актуальной (23 %), актуальной (37 %) или весьма актуальной (32 %).

2. Согласны ли вы с тем, что одних нейронных подходов достаточно для создания универсальных ИИ-агентов, интеллект которых соответствует уровню человека или превосходит его во всех отношениях? «Да» ответили 16 % респондентов, остальные выбрали вариант «нет».

3. Какой, по вашему мнению, процент исследований в области ИИ необходимо посвятить сочетанию нейронных подходов с другими подходами? 94 % участников опроса ответили, что не менее 25 %, из них 31 % назвали цифру 25 %, 35 % респондентов – 50 % и 28 % респондентов – более 50 %. Один респондент ответил «0 %».

4. Какой, по вашему мнению, процент исследований в области ИИ необходимо посвятить исключительно ненейронным подходам? 86 % респондентов ответили, что не менее 25 %, из них 37 % назвали цифру 25 %, 38 % респондентов – 50 % и 11 % респондентов – более 50 %. Шесть респондентов (3 %) ответили «0 %».

5. Какие парадигмы, выходящие за рамки нейронных сетей, по вашему мнению, заслуживают максимального внимания исследователей на сегодняшний день? Участники опроса дали три следующих ответа на данный открытый вопрос:

- *«Нам необходимо понять, является ли мозг квантовой системой».*
- *«Классические подходы к ИИ, которые сконцентрированы на познавательных способностях высокого уровня, опираются на структурированные представления, придерживаются системного подхода, используют идеи из психологии и стремятся к теоретическому пониманию, а не к победе в состязании».*
- *«Крайне важное значение имеют междисциплинарное сотрудничество и принятие во внимание этических и социальных последствий».*

Следует помнить, что во всех категориях респонденты сами принимали решение об участии в опросе по данной теме (чуть больше половины всех респондентов). Большинство из этих респондентов отметили резонанс основного тезиса данной темы, а именно важность инвестирования в ненейронные исследования в определенной степени. С другой стороны, неудивительно, что многие из тех, кто решил не отвечать на эти вопросы, считают иначе.

Сторонние исследования ИИ

Расширение исследований в области ИИ за счет анализа разнообразных точек зрения и использования опыта дисциплин, выходящих за рамки ключевых исследований ИИ

Основные выводы

- Принятие во внимание мнений социологов, специалистов по этике и политиков с целью обеспечить ответственную разработку и внедрение технологий ИИ с соблюдением этических принципов.
- ИИ – это не просто техническая дисциплина. Это общественная сила, которая меняет управление, культуру, экономику и этику, требуя применения комплексных подходов для обеспечения ответственности при разработке и развертывании.
- «Интеллектуальная поддержка» и инструменты, способствующие слаженному сотрудничеству между исследователями ИИ и экспертами из различных областей знаний и обеспечивающие создание этичных, объяснимых и проблемно-ориентированных решений в сфере ИИ.

ПРЕДСЕДАТЕЛЬСТВУЮЩИЙ

Джийе Ким (Jihie Kim),
Университет Донгук

Контекст и история

Сторонние исследования ИИ подчеркивают важность расширения исследований в сфере ИИ и использования разнообразных мнений и опыта дисциплин, выходящих за рамки ключевых исследований ИИ. В таком расширении можно выделить три направления. Во-первых, для обеспечения ответственной разработки и внедрения ИИ-технологий с соблюдением этических принципов нам необходимо учесть точки зрения различных специалистов, включая, помимо прочего, социологов, специалистов по этике, экспертов в области цифровых гуманитарных наук, исследователей критических данных и массовых коммуникаций, специалистов в области исследований науки и технологий (Science and technology studies, STS) и других дисциплин, а также политиков. Во-вторых, исследователи и практические специалисты в дисциплинах, где все чаще используется ИИ (например биология, право, бизнес, нейробиология, когнитивистика), также могут влиять на наше представление об ИИ. Наконец, по мере роста междисциплинарных исследований мы можем обеспечить «интеллектуальную поддержку» и инструменты для взаимодействия.

Социальные, этические, правовые и культурные проблемы, обусловленные появлением ИИ, подчеркивают необходимость использования междисциплинарного подхода при разработке и внедрении ИИ-технологий [1,2]. Такие проблемы, как необъективность решений, нарушение конфиденциальности, управление, ответственность и инклюзивность, требуют решений, выходящих за рамки инженерной сферы. Для решения этих проблем необходима интеграция мнений представителей гуманитарных, общественных и других наук. Это позволит согласовать ИИ-системы с правами человека, общественными нормами, а также принципами глобального равенства и справедливости.

Такой расширенный подход к ИИ подчеркивает, что его развитие и влияние нельзя рассматривать исключительно в рамках достижений технического прогресса или определенной области применения. Это означает, что ИИ больше не является чисто технической дисциплиной. Напротив, необходимо применять комплексный подход к ИИ как к преобразующей силе, способной изменить структуру общества, культурные нормы, экономические модели и этические стандарты. При таком подходе ИИ рассматривается не только как инструмент для отдельных дисциплин, но как интегрированное явление, которое влияет на различные социальные факторы и подвержено влиянию таких факторов. Используя опыт различных экспертов, включая специалистов по этике, юристов, социологов и политиков, мы можем разработать стандарты управления и механизмы ответственности для решения таких проблем, как необъективность, неравенство и нежелательные последствия для общества. Повышение инклюзивности и разнообразия в области проектирования ИИ-систем также может смягчить риски и способствовать достижению максимальной выгоды от применения таких систем.

Кроме того, «интеллектуальная поддержка» такой деятельности предполагает объединение инструментов, структур и методов, ускоряющих эффективную интеграцию ИИ в различных областях знаний и упрощающих сотрудничество между исследователями ИИ и экспертами в других областях.

Текущая ситуация и тенденции

- Социологи и этики все чаще участвуют в создании инструкций по обработке персональных данных, предотвращению незаконной слежки и обеспечению ответственного использования ИИ для разработчиков ИИ-систем. Также активно обсуждаются правительственные меры по обеспечению соответствия разрабатываемых технологий ИИ правам человека, нормам права и потребностям общества, такие как Регламент ЕС об ИИ [3].
- ИИ стал ключевым инструментом в таких сферах, как здравоохранение [4], право [5], бизнес [6] и т. п., и это все сильнее влияет на исследования в сфере ИИ. В этих сферах необходимы высокоточные и поддающиеся объяснениям и интерпретации ИИ-системы в связи с важностью прозрачности и ответственности. Это стало стимулом для проведения новых исследований в области объяснимого ИИ (Explainable AI, XAI), а также алгоритмов ИИ, специально созданных для определенных областей применения, например для онкодиагностики [7].
- «Интеллектуальная поддержка» и инструменты для обеспечения взаимодействия между ИИ и другими дисциплинами: еще недавно сообщества активно использовали инструменты, не имеющие отношения к ИИ, такие как GitHub, Kaggle и исследовательские репозитории. Мы предполагаем, что использование интеллектуальных возможностей может способствовать более продуктивному сотрудничеству.

Исследовательские задачи

При формулировании направлений будущих исследований необходимо учитывать следующие основные принципы:

- **Динамика общественного развития:** ИИ меняет структуру человеческого общения, рынков труда и общественного управления. Понимание того, каким образом автоматизация, принятие решений на основе алгоритмов и системы на базе ИИ влияют на демократию, социальную справедливость и равенство, играет решающую роль в использовании их потенциала без усугубления существующего неравенства.
- **Интеграция этических принципов:** этические нормы должны быть изначально заложены в системах ИИ. Сюда относится решение дилемм,

связанных с конфиденциальностью, ответственностью и справедливостью. Аспекты этики должны выходить за рамки технических проверок. Для создания систем с возможностью глобальной адаптации и соответствия контексту необходимо использовать разнообразные культурные, философские и социальные исходные данные.

- **Нормативно-правовое регулирование:** необходимо решить задачу регулирования и управления, включая пересмотр существующих норм в отношении интеллектуальной собственности, обязательств и прав человека. Важную роль играет совместная междисциплинарная деятельность по созданию моделей управления, повышающих уровень доверия и безопасности, защищающих интересы общества и способствующих более масштабному внедрению ответственных инноваций.
- **Культурная адаптация и разнообразие:** по мере того, как ИИ-технологии проникают в различные культуры и регионы, они должны адаптироваться под разные социальные нормы, языки и традиции. Учет культурных особенностей при создании и внедрении ИИ-систем обеспечивает инклюзивность и равный доступ к технологическим преимуществам.
- **Образование и общественная осведомленность:** расширение влияния ИИ требует переосмысления систем образования для подготовки будущих поколений к жизни в мире с повсеместным использованием ИИ. Данный принцип включает в себя стимулирование трансдисциплинарного понимания ИИ среди инженеров, социологов, политиков и представителей общественности, необходимое для создания информированного общества с широкими возможностями;
- **Экологическая устойчивость:** потребность в энергоресурсах на разработку и внедрение ИИ-систем определяет степень их воздействия на окружающую среду. Расширенный подход к ИИ определяет то, какой вклад может вносить ИИ в работы по поддержанию экологической устойчивости: от оптимизации распределения ресурсов до усовершенствования климатических решений с одновременным сведением к минимуму экологического ущерба.

Включение других дисциплин подразумевает не только разработку ИИ-технологий для их понимания и управления их взаимодействием с обществом, но и гарантию того, что влияние этих технологий будет соразмерным, инклюзивным и выгодным для всех сфер человеческой деятельности.

Для поддержки этой работы и поощрения взаимодействия между ИИ и другими дисциплинами можно также провести исследования на следующие темы:

- **Удобные платформы для разработки ИИ-систем:** эта мера может позволить специалистам в различных сферах разрабатывать решения на основе ИИ или обучать модели глубинного обучения, не имея большого опыта программирования. Подобная платформа может также служить основой для разработки ИИ-систем с соблюдением принципов этики. Интеллектуальные пользовательские интерфейсы и инструменты визуализации могут помочь людям понять результаты работы ИИ-модели и получить аналитические данные.
- **Инструменты ИИ для конкретной области знаний:** ИИ-системы, адаптированные под конкретные потребности определенной области знаний, например медицинской диагностики, кибербезопасности, права и т. п. Мы с нетерпением ждем появления интеллектуальных систем, которые будут способны понимать потребности работников в той или иной области и обеспечивать эффективную поддержку рабочих процессов и процессов формирования рассуждений в этой области. Такие системы должны быть гораздо более эффективными, чем имеющиеся инструменты, например, предварительно обученные модели или системы на основе баз знаний.
- **Инструменты совместной работы на основе ИИ:** интеллектуальные среды, которые помогают исследователям ИИ и экспертам в других областях вести совместную работу.

1. Гальегос, Исабель О., Росси, Райан А., Бэрроу, Джо, Танджим, Мохаммед Мехраб, Ким, Сунгчул, Дернонкур, Фрэнк, Юй, Тун, Чжан, Жуйюй и Ахмед, Несрин К. (Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed). Необъективность и объективность в больших языковых моделях: исследование. (Bias and Fairness in Large Language Models: A Survey). Computational Linguistics, 50(3):1097–1179, 2024 г.
2. Юнь, Юнсик и Ким, Цзихи (Youngsik Yun and Jihie Kim). CIC: основы культурно-осведомленного захвата изображений. (CIC: A framework for Cultually-aware Image Captioning). Международная совместная конференция по искусственному интеллекту, 2024 г.
3. Регламент ЕС об ИИ, Регламент (ЕС) 2024/1689. https://artificialintelligenceact.eu/
4. ИИ в здравоохранении (AI in health care). Nature Collection 22, 2024 г. https://www.nature.com/collections/dbfcjjigbi
5. Кайт-Джексон, Д.У. (D.W. Kite-Jackson) (2023 г.). Технический отчет по искусственному интеллекту (ИИ). (Artificial Intelligence (AI) TechReport). Американская ассоциация юристов, 2023 г.
6. McKinsey & Company. Состояние ИИ в начале 2024 года: рост внедрения генеративного ИИ и создание ценности (The state of AI in early 2024: Gen AI adoption spikes and starts to generate value), McKinsey, 2024 г.
7. Яницек, Джозеф Д., Динчер, Айше Б., Челик, Сафие, Чен, Хью, Чен, Уильям, Наксерова, Камила и Ли, Су-Инь (Joseph D. Janizek, Ayse B. Dincer, Safiye Celik, Hugh Chen, William Chen, Kamila Naxerova, and Su-In Lee). Обнаружение признаков проявления синергических реакций на лекарственные средства с помощью наборов объяснимых моделей машинного обучения. (Uncovering expression signatures of synergistic drug responses via ensembles of explainable machine-learning models). Nature Biomedical Engineering 7, 811–829, 2023 г. https://www.nature.com/articles/s41551-023-01034-0

Мнение сообщества

Сначала мы поинтересовались у представителей сообщества, какие виды деятельности они считают наиболее важными. Респонденты расставили приоритеты следующим образом: 1) содействие междисциплинарному сотрудничеству в области исследований ИИ; 2) разработка решений на основе ИИ для конкретных областей применения, таких как здравоохранение, право и бизнес; и 3) повышение осведомленности общества о влиянии ИИ. К следующему уровню приоритетов респонденты отнесли: 4) интеграцию образовательной деятельности в сфере ИИ в научные дисциплины, не связанные с компьютерами; 5) формирование стандартов ответственности; 6) анализ социальных и культурных последствий влияния ИИ с точки зрения различных дисциплин и продвижение инклюзивности в процессе проектирования ИИ-систем. Хотя такие виды деятельности, как оптимизация распределения ресурсов и разработка моделей управления, набрали меньше голосов, разница была незначительной. В целом, сообщество проявило активный интерес к внедрению широкого ряда инициатив.

В качестве наиболее важного направления подавляющее большинство респондентов (88 %) назвали здравоохранение. Далее следует климат (50 %), образование (45 %) и биология (43 %). Респонденты также назвали такие сферы, как производство (38 %), бизнес (19 %), право (18 %) и финансы (15 %). Кроме того, респонденты отметили следующие дополнительные направления деятельности: сельское хозяйство, транспорт, заболевания головного мозга, физика, уход за пожилыми людьми, ликвидация последствий стихийных бедствий, медиааналитика и кибербезопасность.

Что касается инструментов поддержки междисциплинарного сотрудничества, наибольшее количество голосов получили удобные платформы для разработки ИИ-систем (65 %), инструменты для совместной работы на основе ИИ (57 %) и ИИ-инструменты для конкретных областей применения (53 %). Помимо технологической поддержки, сообщество предложило следующие меры содействия совместной работе: образовательная поддержка, меры стимулирования междисциплинарной работы, формирование междисциплинарных исследовательских групп и устойчивое финансирование долгосрочных проектов. По мере того, как междисциплинарная работа приобретает все более важное значение для исследований в области ИИ, сообщество, по-видимому, стремится подчеркнуть острую необходимость в более масштабной поддержке по разным направлениям.

На вопросы по этой теме ответила примерно треть всех участников опроса, при этом большинство из них (90 %) назвали ИИ своей основной сферой деятельности. Несмотря на ограниченные данные по другим дисциплинам, результаты опроса указывают на активный интерес к интеграции различных точек зрения в рамках исследований и усилению поддержки междисциплинарного сотрудничества внутри сообщества.

Роль исследовательского сообщества

Достижения современного ИИ в значительной степени являются результатом работы частного сектора, и высшим учебным заведениям приходится бороться за то, чтобы быть конкурентоспособными: им нужно найти свою роль в новой эре «большого ИИ».

ПРЕДСЕДАТЕЛЬСТВУЮЩИЙ

Майкл Вулдридж
(Michael Wooldridge),
Оксфордский университет

Основные выводы

- «Центр тяжести» современных исследований в области ИИ находится за закрытыми дверями крупнейших технологических компаний.
- Университеты не в состоянии конкурировать с технологическими гигантами в части ресурсов (данных, вычислений и заработной платы), которые сосредоточены в частном секторе.
- Высшие учебные заведения с трудом сохраняют преподавательский состав по дисциплинам, связанным с ИИ, и убеждают выпускников заниматься научной деятельностью.
- Поэтому сейчас нужно решить проблему поиска роли научного сообщества (и исследований, финансируемых за счет государства) в новой эре «большого ИИ».

Контекст и история

Несмотря на то, что промышленность всегда демонстрировала интерес к ИИ, прогресс в данной области по большей части был обусловлен научными достижениями. Лауреаты премии Тьюринга за достижения в области ИИ (Фейгенбаум, МакКарти, Мински, Ньюэлл, Перл, Редди, Саймон) представляли университеты, а основные концепции как символического, так и нейронного ИИ были разработаны академическим сообществом. Все лауреаты премии Тьюринга в области глубинного обучения (Бенджио, Хинтон, ЛеКун) также проводили свои выдающиеся исследования в университетах.
Но за последнее десятилетие центр тяжести передовых ИИ-технологий безусловно сместился. В частности, сейчас основную работу по генеративному ИИ выполняет частный сектор и несколько крупных технологических компаний. Эти компании располагают достаточными данными и вычислительными ресурсами для обучения масштабных передовых ИИ-систем. Кроме того, они могут предлагать своим сотрудникам зарплаты, с которыми университеты не в состоянии конкурировать.

Текущая ситуация и тенденции

Эти изменения стали причиной ряда проблем для университетов.

Во-первых, спрос на исследователей в области ИИ привел к оттоку таких специалистов из высших учебных заведений в частный сектор. Ярким примером здесь является переход целой лаборатории исследователей-робототехников из Университета Карнеги-Меллона в Uber [1] . Даже ведущие мировые университеты не способны предложить преимущества, которые были бы сопоставимы с невероятными зарплатами, предлагаемыми ИИ-лабораториями технологических гигантов. Кроме того, университеты не в состоянии обеспечить такие исследовательские ресурсы, которыми обычно обладают крупные компании. Результатом этого стало ослабление групп по исследованиям в области ИИ во многих университетах, при этом многие преподаватели как будто бы находятся в постоянном отпуске или работают по совместительству, что значительно отвлекает их от обычной научной деятельности. Прием на работу преподавателей по ИИ уже более десяти лет является крайне сложной задачей: аспиранты предпочитают сразу уходить в частные компании по окончании учебы.

В то же время характер исследований в области ИИ значительно изменился. На рубеже веков преобладающей парадигмой (к примеру) в сообществе членов Ассоциации по развитию искусственного интеллекта была традиционно научная парадигма с опорой на математику (определение – лемма – теорема – доказательство). И наоборот, методика глубинного обучения и сопутствующие области, доминирующие на момент написания исследования, опираются преимущественно на инженерно-техническую парадигму.

В результате спрос на университетские курсы по компьютерным наукам в целом и по ИИ в частности резко вырос, и это, в сочетании с описанными выше проблемами, создало серьезную нагрузку на факультеты, преподаватели которых с трудом справляются с таким уровнем спроса. Некоторые факультеты компьютерных наук в университетах Лиги Плюща сообщают, что им предоставили карт-бланш на поиск новых преподавателей предметов, пользующихся наибольшим спросом. Но они попросту не в состоянии найти достаточное количество высококвалифицированных педагогов. На некоторых факультетах ситуация достигла критического уровня, в результате чего преподаватели начали жаловаться на стресс и другие ментальные расстройства.

Все студенты (независимо от специальности) должны владеть основами ИИ. Использование ИИ меняет систему образования и процесс преподавания, но в то же время студенты, изучающие ИИ, должны получать междисциплинарные знания, чтобы иметь представление об этических, правовых, социальных и экономических последствиях внедрения ИИ. В свою очередь, это предъявляет дополнительные требования к университетам, которым приходится переосмысливать процесс обучения и оценки студентов, когда в распоряжение таких студентов поступают мощные универсальные ИИ-инструменты.

Ирония сложившейся ситуации состоит в том, что ключевая роль университетов исторически заключалась в том, чтобы создавать кадровый резерв для удовлетворения спроса технологических компаний на профессионалов. Но университеты с трудом справляются с этой задачей просто потому, что у них нет для этого возможностей — так как эти же самые компании переманивают к себе преподавателей.

В то же время характер передовых исследований в области ИИ указывает на то, что университеты попросту не способны выполнять работу, конкурирующую с деятельностью ИИ-лабораторий крупнейших частных компаний. По последним оценкам Meta, создание новейших моделей класса GPT обошлось компании приблизительно в 440 млн долл. США. Эта сумма в десятки раз превышает возможности всех мировых университетов, кроме небольшого числа самых богатых учебных заведений, и большинству стран было бы сложно выделить такую сумму на разработку инициативы государственного уровня: в 2023 году Великобритания изучала возможность разработки независимого проекта в области ИИ, и одним из вариантов было создание «BritGPT». Идея потерпела неудачу уже на ранней стадии реализации по следующим причинам: (i) стоимость, (ii) риски и (iii) необходимость конкурировать с частными компаниями в области, где лидером по инновациям (и затратам на разработку) является именно частный сектор [2]. Новейшие разработки, например китайская большая языковая модель с открытым исходным кодом DeepSeek, могут предложить определенные возможности в этом

направлении, поскольку они способны сделать ИИ доступным за меньшие деньги. Но это само по себе создает проблемы в части конфиденциальности и безопасности.

Цели частных компаний и университетов различаются. Стимулом для частного сектора преимущественно является получение прибыли, тогда как университеты стремятся внести вклад в развитие общества посредством исследований и образовательной деятельности. Эти цели могут конкурировать друг с другом. Одним из последствий является то, что результаты работы частных компаний не всегда полностью доступны для проверки или оценки. Кроме того, высокие затраты на эксперименты подразумевают, что результаты не всегда являются воспроизводимыми. Ученые должны сыграть свою роль в предоставлении независимых консультаций и интерпретации таких результатов, а также их последствий. Частный сектор уделяет больше внимания краткосрочным проектам, тогда как университеты и общество больше заинтересованы в долгосрочных исследованиях.

Эти наблюдения в сочетании с огромными затратами на исследования в области генеративного ИИ также стали основой для призывов к разработке масштабных финансируемых государством инициатив. Одним из примеров является план «ЦЕРН для ИИ» (CERN for AI), в поддержку которого выступила президент Еврокомиссии Урсула фон дер Ляйен. Эту точку зрения поддерживают многие исследователи ИИ в Европе (они же фактически инициировали ее), которые организовали Конфедерацию лабораторий для исследования ИИ в Европе (Confederation of Laboratories for Artificial Intelligence Research in Europe, CAIRNE). «ЦЕРН для ИИ» создан по образцу известного Европейского центра ядерных исследований в Женеве. Он должен функционировать как альтернативная и перспективная среда для проведения исследований крупнейшими технологическими компаниями и решать проблемы, представляющие интерес для общества. Многие страны также занимаются разработкой собственных национальных стратегий и программ финансирования в области ИИ.

Исследовательские задачи

- Как университеты должны реагировать на наступление новой эры «большого ИИ»?
- В каких исследованиях в области ИИ университеты и компании частного сектора могут участвовать с наибольшей пользой? Что должна представлять собой университетская программа исследований в области ИИ в будущем?
- Как университеты должны реагировать на проблемы привлечения и удержания преподавателей и студентов на факультетах, связанных с ИИ?
- Каким может быть наиболее эффективное сотрудничество между финансируемыми государством университетами и частными компаниями, занимающимися разработкой ИИ?

1. https://www.fastcompany.com/3046902/carnegie-mellon-in-a-crisis-after-uber-poached-40-of-its-researchers
2. https://lordslibrary.parliament.uk/large-language-models-and-generative-ai-house-of-lords-communications-and-digital-committee-report/

Мнение сообщества

Большинство респондентов (около 75 %) согласны с тем, что университеты сталкиваются с проблемой поиска преподавателей по ИИ и проблемой участия в ресурсоемких исследованиях в области ИИ (80 %). Респонденты считают, что для привлечения кадров можно предлагать более высокие зарплаты и инвестировать в более масштабные вычислительные ресурсы, а также обеспечивать возможность работы по совместительству и предлагать другие преимущества. Что касается необходимости пересмотра приоритетов исследований для университетов, мнение респондентов было не столь единодушным. При этом респонденты полагают, что университеты могут быть конкурентоспособными в таких областях, как теория ИИ и междисциплинарный ИИ. Государственное финансирование крупномасштабных вычислений было признано привлекательным направлением (70 %). Подавляющее большинство сошлось во мнении, что научное сообщество имеет важное значение для будущего исследований в области ИИ.

Геополитические аспекты и возможные последствия применения ИИ

Развитие ИИ меняет расстановку сил на мировой арене и приоритетные направления инвестирования в разных странах, воздействует на экономику, систему безопасности и управленческие структуры, порождая проблемы контроля ИИ для обеспечения его безопасного и справедливого использования.

Основные выводы

- Вопросы, связанные с инвестициями, координацией сотрудничества, лучшими практиками использования и регулированием в сфере ИИ, приобретают международное значение. Несмотря на расширение сотрудничества в области ИИ в рамках государственных и негосударственных программ, страны также ведут геополитическую борьбу на данном поприще за экономическое, военное и стратегическое господство: они стремятся использовать ИИ для получения экономических, военных и стратегических преимуществ.
- Регулирование и конкуренция: противоречия между регулированием ИИ, конфиденциальностью и борьбой за технологическое превосходство осложняют международное сотрудничество.
- Этические и социальные последствия интеграции ИИ: развертывание ИИ в соответствии с политикой и конкурентными целями разных государств вызывает опасения по поводу соблюдения принципов справедливости, объективности и демократических ценностей, поэтому требуются новые стандарты управления ИИ. Некоторые из этих стандартов должны быть международными.

ПРЕДСЕДАТЕЛЬСТВУЮЩИЕ

Вирджиния Дигнум (Virginia Dignum), Университет Умео

Хольгер Хус (Holger Hoos), Рейнско-Вестфальский технический университет Ахена, Германия, и Лейденский университет, Нидерланды

Эрик Хорвиц (Eric Horvitz), Microsoft

Контекст и история

Если раньше ИИ был связан исключительно с исследованиями и технологиями, то теперь он становится основным элементом глобальных экономических стратегий и стратегий безопасности: разрабатываются управленческие стандарты, предусматривающие ответственное использование ИИ [1,2]. На фоне последних событий, включая развитие больших языковых моделей и автоматизации с помощью ИИ, усиливается обеспокоенность по поводу конкуренции между странами, особенно между США, Китаем, Россией и Европейским союзом.

В настоящее время ИИ все больше зависит от экономических интересов и конкурирующих подходов к принятию важных решений. В последние годы взгляды стран на стратегию, инвестиции и управление в области ИИ разнятся, хотя и взаимосвязаны как внутри стран, так и в контексте международных отношений, судя по их заявлениям и действиям на международном уровне. Эта динамика также подвержена изменениям, обусловленным политическими переменами.

В октябре 2023 г. президент США Джо Байден издал Указ [3] «О безопасном, надежном и заслуживающем доверия ИИ» (Safe, Secure, and Trustworthy Artificial Intelligence), основанный на Билле о правах в сфере ИИ (AI Bill of Rights), который разработало Бюро по определению научно-технической политики (Office of Science and Technology Policy) [4]. Особое внимание в Указе уделяется защите гражданских прав и неприкосновенности частной жизни, а также вводятся строгие стандарты безопасности ИИ. Федеральные ведомства США отвечают за внедрение ИИ-систем и принятие важных решений в этой сфере, обеспечивая прозрачное и справедливое использование таких систем без какой-либо алгоритмической дискриминации, чтобы не допустить необъективности данных или нарушения прав личности. Кроме того, в США создан национальный Институт безопасности ИИ (AI Safety Institute) [5], а совместно с другими странами — Международная сеть институтов безопасности ИИ (International Network of AI Safety Institutes) [6].

В январе 2025 г., вскоре после своей инаугурации, Дональд Трамп отменил указ Байдена, заменив его указом «Об устранении препятствий для американского лидерства в области ИИ» (Removing Barriers to American Leadership in Artificial Intelligence), который был направлен на поддержание и укрепление глобального господства США «...в целях содействия процветанию человечества, экономической конкурентоспособности и национальной безопасности». [7]

В Регламенте ЕС об ИИ [8], принятом в августе 2024 г. после четырех лет обсуждений, применяется подход к регулированию ИИ по аналогии с безопасностью продукции: ИИ-системы классифицируются по уровням риска (неприемлемый, высокий, ограниченный и минимальный), при этом устанавливаются соответствующие обязательства. Приложения ИИ с высоким уровнем риска, например в сфере здравоохранения и транспорта, должны соответствовать строгим требованиям безопасности, прозрачности и надзора, чтобы не нанести ущерба здоровью, безопасности или основным правам граждан.

В Китае ИИ называют «важной стратегической возможностью», а к 2030 г. страна планирует стать мировым лидером в области ИИ [9]. Китай одним из первых ввел нормативные акты, регулирующие использование ИИ-систем, включая подробные правила в отношении алгоритмов рекомендаций, которые вступили в силу в 2021 г. [10]. Китай продолжает внедрять ИИ в инфраструктуру наблюдения.

В 2019 г. президент России Владимир Путин издал указ об ускоренном развитии ИИ в Российской Федерации на период до 2030 г. В 2023 г. вышла новая редакция указа, содержащая план развития с изложением ключевых принципов развития ИИ, «таких как защита прав человека, обеспечение безопасности, технологический суверенитет и поддержка конкуренции». [11]

В утвержденном Конгрессом докладе Комиссии национальной безопасности США по ИИ (U.S. National Security Commission on AI, NSCAI) от 2020 г. подчеркивается необходимость в международной координации и соглашениях по управлению различными аспектами ИИ, включая оборону и установление норм в области прав человека и принципов ответственного внедрения ИИ-технологий [12]. В докладе содержится призыв к созданию альянсов стран, разделяющих западные демократические ценности, для координации стратегий. В плане обороны в докладе предлагается создать международные площадки для обсуждения влияния ИИ на стабильность страны среди стран-конкурентов в условиях кризиса и подготовить международные практические стандарты в области разработки, тестирования и использования автономных систем оружия на базе ИИ.

В последнее время международное сообщество предпринимает все больше усилий по совместному принятию важных решений в сфере ИИ. Такие организации, как ОЭСР [13], ООН [14] и Глобальное партнерство по ИИ (Global Partnership on Artificial Intelligence, GPAI), выступают за принципы глобального международного управления ИИ. О постоянном стремлении к глобальной координации действий свидетельствуют важные международные встречи: саммит по безопасности ИИ в Великобритании (Блетчли-парк, ноябрь 2023 г.), саммит по ИИ в Сеуле (май 2024 г.) и саммит по действиям в сфере ИИ (Париж, 2025 г.).

На саммите в Сеуле представители Австралии, Канады, Европейского союза, Франции, Германии, Италии, Японии, Республики Корея, Республики Сингапур, Великобритании и США подтвердили общую «приверженность развитию международного сотрудничества и диалога по вопросам ИИ в условиях его феноменального развития и влияния на экономику и общество». [15]. В ноябре 2024 г. на саммите по безопасности ИИ в Сан-Франциско была создана Международная сеть институтов безопасности ИИ в продолжение инициатив, изложенных в декларации, которая была принята на саммите по безопасности ИИ в Сеуле, [16]. Обсуждения между правительствами, представителями гражданского общества и промышленности продолжаются на многочисленных форумах, а также в Организации заинтересованных сторон

«Партнерство по ИИ» (Partnership on AI).

Растут потребности и возможности международной координации в области норм и правил, регламентирующих права человека, неприкосновенность частной жизни и безопасность ИИ-систем. В число актуальных международных проблем входят вопросы регулирования интеллектуальной собственности в отношении данных, используемых для обучения больших языковых моделей. Основные проблемы, носящие трансграничный характер, включают в себя вопросы обращения с контентом, созданным с помощью ИИ и используемым для дезинформации, а также ИИ-угрозы в сфере биологической безопасности, такие как использование инструментов проектирования белков на базе ИИ для создания опасных токсинов и патогенов.

Проблемы не исчезают из-за разрозненности национальных интересов и подходов к регулированию, в результате чего усиливается напряженность между странами относительно роли ИИ в торговле, безопасности и правах человека. Эти разногласия были заметны на саммите по действиям в сфере ИИ в феврале 2025 г., где ЕС и Китай настаивали на ужесточении регулирования ИИ, а США и Великобритания отказались от принятия глобальной декларации о безопасности ИИ, опасаясь, что слишком строгие правила могут препятствовать инновациям. Франция и другие лидеры ЕС в области цифровой трансформации ратовали за более гибкую нормативно-правовую базу для привлечения инвестиций в отрасль и недопущения ее стагнации.

Хотя на этих саммитах, безусловно, были полезные дискуссии по вопросам безопасности и регулирования ИИ, их критикуют за то, что в них не участвовали многие страны, особенно страны Глобального Юга. Кроме того, такие регионы, как Юго-Восточная Азия, несмотря на активные меры в области безопасности и регулирования ИИ, часто принимают ограниченное участие в глобальных обсуждениях вопросов безопасности ИИ. Такое избирательное участие порождает вопросы о законности и эффективности этих мероприятий в контексте решения глобальных проблем ИИ [26], а также о гарантиях в отношении важных обязательств по обеспечению безопасности ИИ: многие указывают на отсутствие конкретных мер безопасности, неопределенность рекомендаций в области государственной политики и чрезмерный акцент на гипотетических рисках, а не на непосредственных проблемах ИИ, несмотря на публикацию Международного отчета по безопасности ИИ в 2025 г. [27].

В то же время экономическая конкуренция по-прежнему остается жесткой. Инициатива США Stargate стоимостью 500 млрд долл. США направлена на укрепление инфраструктуры ИИ и конкурентоспособности страны на мировом уровне, а инициатива ЕС InvestAI стоимостью 200 млрд евро призвана стимулировать исследования и развертывание ИИ в разных странах Европы. Между тем в Индии делается упор на справедливый и более инклюзивный подход к разработке ИИ с открытым исходным кодом, обеспечивающий преимущества для всех регионов.

Хотя эти усилия и отражают растущее осознание трансформационного потенциала ИИ-технологий, они также подчеркивают сохраняющуюся проблему: без единых стандартов управления регулирование ИИ, по-видимому, останется разрозненным, что только усугубит риски, связанные с безопасностью, экономическим неравенством и геополитической нестабильностью. Без скоординированных международных соглашений эти разногласия могут усилить существующее глобальное неравенство и геополитическую напряженность относительно контроля и управления ИИ [18].

Текущая ситуация и тенденции

1. ИИ все чаще становится определяющим фактором национальной и региональной власти, влияющим на торговую политику, военные стратегии и дипломатические отношения [19]. В настоящее время США и Китай занимают господствующее положение в области ИИ, тогда как в Европе приоритетными являются этические соображения и регулирующий надзор. В то же время наблюдение и сбор данных с помощью ИИ оказывают влияние на модели управления ИИ во всем мире, особенно в автократических режимах, где первостепенное внимание уделяется безопасности государства, а не свободе личности, из-за чего возникают опасения по поводу неприкосновенности частной жизни и гражданских свобод.

2. Нормативно-правовая база остается разрозненной [20], при этом подходы к принятию важных решений у основных мировых игроков сильно различаются. ЕС отличается продуманным подходом к регулированию ИИ: так, принят Регламент об ИИ для обеспечения безопасного и ответственного использования этой технологии. Однако с приходом новой администрации США больше не уделяют пристального внимания безопасности ИИ и регулированию, ориентированному на права человека, как это было раньше.

3. Правительства координируют свои действия посредством таких мероприятий, как международные встречи в Великобритании, Сеуле и Париже и создание Международной сети институтов безопасности ИИ.

4. Несмотря на отсутствие обязательных международных соглашений, коалиции частных корпораций, представители гражданского общества и некоммерческие организации работают над созданием добровольных соглашений и стандартов. В качестве примеров можно привести следующие инициативы:

- Решение проблем, связанных с дезинформацией и манипуляциями с помощью ИИ: Коалиция по проверке происхождения и подлинности контента (Coalition for Content Provenance and Authenticity, C2PA), разработавшая криптографический стандарт проверки происхождения медиаконтента, который в настоящее время принят крупными технологическими компаниями, и «Технологическое соглашение по борьбе с обманным использованием ИИ на выборах 2024 года» (Tech Accord to Combat

Deceptive Use of AI in 2024 Elections), в котором установлены обязательства в отношении контента, созданного с помощью ИИ.

- Решение проблем биологической безопасности с использованием ИИ: усилия сосредоточены на разработке принципов и лучших практик ответственного использования ИИ в биологических науках [21] и международной координации протоколов скрининга нуклеиновых кислот [22].

5. Одна из самых острых тем дискуссии – модели ИИ с открытым и закрытым исходными кодами. Основные проблемы связаны с доступностью, необъективностью и прозрачностью обучающих данных, а также с возможными злоупотреблениями ИИ со стороны государственных и негосударственных субъектов. В дополнение к этому усиливается геополитическая напряженность: США вводят ограничения на экспорт полупроводников в ряд стран, включая Китай и некоторые государства-члены ЕС, что сказывается на глобальных цепочках поставок и усугубляет технологический разрыв между основными державами – лидерами в области ИИ.

6. Помимо трудностей, связанных с управлением и конкуренцией в данной сфере, использование ИИ сопряжено с этическими и социальными проблемами [23]. Принятие решений с помощью ИИ в таких важнейших секторах, как управление персоналом, здравоохранение, правоохранительная система и финансовые услуги, вызывает обеспокоенность по поводу необъективности, дискриминации и социального неравенства. Стремительное развитие автономного и полуавтономного оружия и военных систем на базе ИИ в таких мировых державах, как США, Китай, Россия и Украина, приводит к еще большему усложнению ситуации с международной безопасностью, вследствие чего подвергаются сомнению существующие нормы ведения войны и ответственности за ее последствия. Чтоб снизить эти риски, при принятии важных решений в сфере ИИ необходимо найти баланс между потребностью в инновациях и надежными этическими гарантиями, обеспечивающий возможность разработки и развертывания ИИ-технологий с соблюдением принципов объективности, безопасности и равного распределения преимуществ ИИ в обществе.

Исследовательские задачи

Модели управления ИИ

- Разработка международных стандартов управления, договоров, норм и практик в сфере ИИ: изучение перспектив единообразного регулирования ИИ в разных странах, устранения нормативной разрозненности и снижения конфликтов в глобальном управлении ИИ. Международное сотрудничество может быть сосредоточено на регулировании следующих сфер: инфраструктура наблюдения и права человека, проблемы интеллектуальной собственности в контексте ИИ, нормы, договоры, принципы, ответственность и лучшие практики в сфере разработки, развертывания и использования автономных и полуавтономных систем оружия, соглашения о международных нормах и правилах в области ИИ и биологической безопасности, а также соглашения об угрозе дезинформации при применении ИИ, правила и лучшие практики в области установления происхождения и подлинности контента, созданного с помощью ИИ.

- Укрепление механизмов исполнения в глобальном управлении ИИ: анализ того, как такие организации, как ООН, ОЭСР и GPAI, могут повысить уровень нормативно-правового соответствия и ответственности в рамках международного сотрудничества в сфере ИИ.

- ИИ и глобальные стандарты управления: изучение роли ИИ в формировании международной нормативно-правовой базы, в том числе того, как разные модели управления влияют на геополитическую стабильность.

Геополитические риски ИИ

- Дезинформация и кампании влияния при применении ИИ: изучение роли ИИ в создании технологий дипфейков, в автоматизированной дезинформации и в пропаганде со стороны государственных и негосударственных субъектов в геополитических конфликтах, а также борьба с ними.

- ИИ и геополитика цепочек поставок: разработка инструментов ИИ для мониторинга и смягчения последствий сбоев в глобальных цепочках поставок, связанных с ИИ, в частности по причине нехватки полупроводников и введения экспортных ограничений.

- Алгоритмическая торговая политика и экономическое прогнозирование: совершенствование ИИ-моделей, которые предсказывают и анализируют влияние автоматизации на основе ИИ и торговых ограничений (например, запретов на экспорт полупроводников) на глобальные рынки.

- ИИ в кибербезопасности и обороне: разработка ИИ-средств обнаружения киберугроз и киберфизических угроз и реагирования на них, [24] а также создание механизмов повышения устойчивости для противодействия кибервойнам, спонсируемым государствами, и защиты критически важных объектов национальной инфраструктуры [25].

- ИИ и проблемы биологической безопасности: активизация координационной деятельности, такой как регулирование протоколов скрининга в организациях, занимающихся синтезом нуклеиновых кислот, регулирование лабораторного синтеза и регистрация вызывающих подозрения заказов для выявления и предотвращения злоупотреблений [22].

- ИИ в военной стратегии, включая принятие решений с помощью ИИ, разработку и применение автономного оружия: изучение последствий применения автономного оружия на основе ИИ для обеспечения стабильности, управления кризисами и стратегического сдерживания, ответственного следования

принципам международного гуманитарного права.

Содействие этическому развитию ИИ

- Выработка междисциплинарного подхода, обеспечивающего объективность и ответственное использование ИИ: изучение влияния политологии, экономики и этики на модели управления ИИ, которые должны учитывать национальные интересы, корпоративные стимулы и глобальную справедливость.
- Совершенствование механизмов управления ИИ в геополитически разрозненной среде: изучение правовых, дипломатических и технологических стратегий для преодоления экономических и идеологических разногласий, повышение возможности принудительного исполнения международных соглашений в области ИИ, таких как Кодекс поведения для компаний, разрабатывающих ИИ-системы (AI Code of Conduct), принятый странами G7.
- Анализ рисков техносолюционизма, существующих в политике в сфере ИИ: изучение непредвиденных последствий принятия решений с помощью ИИ посредством междисциплинарных исследований, обеспечение того, чтобы ИИ дополнял, а не замещал деятельность человека с соблюдением этических принципов.

1. Андреас Теодору и Вирджиния Дигнум (Theodorou, Andreas, and Virginia Dignum). На пути к этическому и социально-правовому управлению ИИ (Towards Ethical and Socio-Legal Governance in AI). Nature Machine Intelligence, том 2, № 1, 2020 г., стр. 10–12
2. И. Улницане, В. Найт, Т. Лич, Б.К. Шталь и В.Г. Ванджику (Ulnicane, I., Knight, W., Leach, T., Stahl, B. C., & Wanjiku, W. G.) (2022 г.). Управление ИИ: новые международные тенденции и политические перспективы. Глава из книги «Глобальная политика ИИ» (Governance of Artificial Intelligence: Emerging international trends and policy frames. In The global politics of Artificial Intelligence). Taylor & Francis.
3. Указ Президента США «О безопасном, надежном и заслуживающем доверия развитии и использовании ИИ» (Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence). Вашингтон, округ Колумбия: Белый дом (Washington DC: The White House); 30 октября 2023 г. EO 14110. Федеральный реестр: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence
4. Белый дом (2022 г.). «Проект билля о правах в сфере ИИ: как заставить автоматизированные системы работать на благо американского народа» (“Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People”). https://www.whitehouse.gov/ostp/ai-bill-of-rights/
5. Институт безопасности ИИ в США (U.S. Artificial Intelligence Safety Institute) (2024 г.). https://www.nist.gov/aisi
6. Международная сеть институтов безопасности ИИ (International Network of AI Safety Institutes) (2024 г.). https://www.nist.gov/system/files/documents/2024/11/20/Mission%20Statement%20-%20International%20Network%20of%20AISIs.pdf
7. Указ Президента США «Об устранении препятствий для американского лидерства в области ИИ» (Executive Order on Removing Barriers to American Leadership in Artificial Intelligence). Вашингтон, округ Колумбия: Белый дом; 23 января 2025 г. EO 14179 of. Федеральный реестр: Removing Barriers to American Leadership in Artificial Intelligence. https://www.federalregister.gov/documents/2025/01/31/2025-02172/removing-barriers-to-american-leadership-in-artificial-intelligence
8. Европейский союз (European Union). Регламент Европейского парламента и Совета, устанавливающий согласованные правила по ИИ (Регламент об ИИ) (Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)). 2024 г.
9. План развития ИИ нового поколения в КНР (China’s ‘New Generation Artificial Intelligence Development Plan’) (2017 г.). https://digichina.stanford.edu/work/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/
10. Положение об управлении алгоритмическими рекомендациями в информационных службах интернета (Provisions on the Management of Algorithmic Recommendations in Internet Information Services) (2021 г.). https://www.chinalawtranslate.com/en/algorithms/
11. Сергей Суханкин (Sergey Sukhankin) (2017 г.). Россия принимает национальную стратегию развития ИИ (Russia Adopts National Strategy for Development of Artificial Intelligence), Eurasia Daily Monitor, том: 16, выпуск: 163 https://jamestown.org/program/russia-adopts-national-strategy-for-development-of-artificial-intelligence/
12. Сафра Кац, Стив Чиен, Миньон Клиберн и др. (Safra Catz, Steve Chien, Mignon Clyburn, et al.) Доклад Комиссии национальной безопасности США по ИИ (Report of the National Security Commission on Artificial Intelligence), Комиссия национальной безопасности США по ИИ (NSCAI) (National Security Commission on Artificial Intelligence (NSCAI)), март 2021 г. https://reports.nscai.gov/final-report
13. ОЭСР (2024 г.), Принятие важных решений с помощью ИИ: готовы ли к этому правительства? (Governing with Artificial Intelligence: Are governments ready?), OECD Artificial Intelligence Papers, № 20, издательский отдел ОЭСР, Париж (No. 20, OECD Publishing, Paris), https://doi.org/10.1787/26324bc2-en.
14. ООН (United Nations). Управление ИИ в интересах человечества: доклад Консультативного органа высшего уровня по ИИ (Governing AI for Humanity: Report of the High-Level Advisory Body on Artificial Intelligence). ООН (United Nations), 2024 г.
15. Сеульская декларация по безопасному, инновационному и инклюзивному ИИ, принятая во время сессии лидеров на саммите по ИИ в Сеуле (Seoul Declaration for Safe, Innovative and Inclusive AI by Participants Attending the Leaders’ Session of the AI Seoul Summit), май 2024 г., Сеул, Корея (Seoul, Korea). https://www.pm.gc.ca/en/news/statements/2024/05/21/seoul-declaration-safe-innovative-and-inclusive-ai-participants-ai-seoul-summit
16. Международная сеть институтов безопасности ИИ: программное заявление (The International Network of AI Safety Institutes: Mission statement). Ноябрь 2024 г. https://ised-isde.canada.ca/site/ised/en/international-network-ai-safety-institutes-mission-statement
17. Х. Робертс, Э. Хайн, М. Таддео и Л. Флориди (Roberts, H., Hine, E., Taddeo, M., & Floridi, L.) (2024 г.). Глобальное управление ИИ: препятствия и пути развития (Global AI governance: barriers and pathways forward). International Affairs, том 100(3), стр. 1275-1286.
18. Й. Тальберг, Е. Эрман, М. Фурендаль, Й. Гейт, М. Кламберг и М. Лундгрен (Tallberg, J., Erman, E., Furendal, M., Geith, J., Klamberg, M., & Lundgren, M.) (2023 г.). Глобальное управление ИИ: дальнейшие действия по проведению эмпирических и нормативных исследований (The global governance of artificial intelligence: Next steps for empirical and normative research). International Studies Review, том 25(3), viad040).
19. Б. Ларсен (Larsen, B.), 2022 г. Геополитика ИИ и рост цифрового суверенитета (The geopolitics of AI and the rise of digital sovereignty), Брукингский институт (Brookings Institution). США (United States of America). Источник: https://coilink.org/20.500.12592/swc5mh от 21 февраля 2025 г. COI: 20.500.12592/swc5mh.
20. Л. Шмитт (Schmitt, L.) (2022 г.). Структура глобального управления ИИ: формирование системы в условиях разрозненности (Mapping global AI governance: a nascent regime in a fragmented landscape). AI and Ethics, том 2(2), стр. 303-314.
21. Д. Блумфилд, Дж. Панну, А.В. Чжу и др. (D. Bloomfield, D., Pannu, J. Zhu, A.W., et al.) ИИ и биологическая безопасность: необходимость управления (AI and biosecurity: The need for governance) (2024 г.). Science 385(6711), стр. 831-833. DOI: 10.1126/science.adq1977 https://www.science.org/doi/10.1126/science.adq1977
22. Б. Дж. Виттманн, Т. Алексанян, С. Бартлинг, Дж. Бил, А. Клор, Дж. Диггенс и др. (Wittmann BJ, Alexanian T, Bartling C, Beal J, Clore A, Diggans J, et al.). На пути к скринингу заказов на синтез нуклеиновых кислот, устойчивых к ИИ: процесс, результаты и рекомендации (Toward AI-resilient screening of nucleic acid synthesis orders: Process, results, and recommendations). Электронный архив bioRxiv. 2024 г. p. 2024.12.02.626439. doi:10.1101/2024.12.02.626439 https://www.biorxiv.org/content/10.1101/2024.12.02.626439v1
23. Лучано Флориди (Floridi, Luciano), Этика ИИ: принципы, проблемы и возможности (The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities) Oxford, 2023 г.; онлайн-издание (online edn), Oxford Academic, 24 августа 2023 г., https://doi.org/10.1093/oso/9780198883098.001.0001
24. Э. Хорвиц (Horvitz, E.). ИИ и кибербезопасность: проблемы и возможности. Заявление перед подкомитетом по кибербезопасности Комитета по вооруженным силам при Сенате США (Artificial Intelligence and Cybersecurity: Rising Challenges and Promising Directions. Testimony before the U.S. Senate Armed Services Subcommittee on Cybersecurity), 3 мая 2022 г. https://www.armed-services.senate.gov/imo/media/doc/5.3.22%20Eric%20Horvitz%20Testimony.pdf
25. Совет по развитию науки и техники при Президенте США (President’s Council on Science and Technology), Стратегия в области киберфизической устойчивости: укрепление критически важной инфраструктуры для цифрового мира (Strategy for Cyber-Physical Resilience: Fortifying Our Critical Infrastructure for a Digital World) (2024 г.). Управление научно-технической политики Белого дома (White House Office of Science and Technology), февраль 2024 г. https://bidenwhitehouse.archives.gov/wp-content/uploads/2024/02/PCAST_Cyber-Physical-Resilience-Report_Feb2024.pdf
26. Уильям Агню, Э. Стиви Бергман, Дженнифер Чиен, Марк Диаз, Селием эль-Сайед, Джейлен Питтман, Шакир Мохамед и Кевин Р. Макки (Agnew, William, A. Stevie Bergman, Jennifer Chien, Mark Díaz, Seliem El-Sayed, Jaylen Pittman, Shakir Mohamed, and Kevin R. McKee). Иллюзия искусственной инклюзивности. Материалы конференции по человеческим факторам в вычислительных системах (CHI 2024) (The illusion of artificial inclusion. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems), стр. 1-12. 2024 г.
27. Йошуа Бенжио, Сёрен Миндерманн, Даниэль Привитера, Тамай Бешироглу, Риши Боммасани, Стивен Каспер, Едзин Чой и др. (Bengio, Yoshua, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi et al.). Международный отчет по безопасности ИИ (International AI Safety Report). Препринт электронного архива arXiv:2501.17805 (2025 г.).

Мнение сообщества

Результаты опроса указывают на проблемы в сфере принятия важных решений, безопасности, экономических изменений и этических аспектов в связи с применением ИИ. Хотя в разных странах и признают геополитическое значение ИИ, мало кто из исследователей уделяет этому первостепенное внимание. Большинство респондентов (49,47 %) считают, что ученые, занимающиеся исследованиями ИИ, должны участвовать в политических дискуссиях, при этом они активно поддерживают международные управленческие механизмы, такие как ООН (53,68 %) и двусторонние соглашения (63,16 %). Среди основных проблем можно выделить угрозы кибербезопасности, военные действия, нарушение экономической устойчивости и поиск баланса между государственным и корпоративным контролем.

Применение ИИ в военной сфере сопряжено с этическими сложностями – с этим полностью согласны 36,84 % опрошенных, а еще 37,89 % убеждены в значимости таких проблем. Более 40 % респондентов поддерживают международное сотрудничество и выступают за заключение соглашений об использовании ИИ в общедоступных данных, ограничениях на развертывание оружия и регулировании неприкосновенности частной жизни. Как отметили респонденты, важно не просто делать символические заявления. Необходимы конкретные соглашения с возможностью принудительного исполнения.

Об Ассоциации по развитию ИИ

Основанная в 1979 году, Ассоциация по развитию ИИ (Association for the Advancement of Artificial Intelligence , AAAI) (ранее — Американская ассоциация по искусственному интеллекту (American Association for Artificial Intelligence)) — это некоммерческое научное общество, продвигающее научное понимание механизмов, лежащих в основе мышления и интеллектуального поведения, и их воплощения в технических устройствах.

Цель Ассоциации — содействовать исследованиям и ответственному использованию ИИ. Кроме того, Ассоциация стремится повысить уровень понимания ИИ в обществе, улучшить процесс обучения и подготовки практикующих специалистов по ИИ, а также предоставить рекомендации для тех, кто планирует и финансирует исследования, о важности текущих разработок в сфере ИИ, их потенциальных возможностях и будущих направлениях.

601 Pennsylvania Ave, NW
Suite 900
Washington, DC 20004

info@aaai.org
1-202-360-4062
aaai.org