Перевод: английский - русский - www.onlinedoctranslator.com

2023-11-04

Уровни ОИИ: операционализация прогресса на пути к ОИИ

Мередит Рингел Моррис[1], **Яша Золь-Дикштейн**[1], **Ной Фидель**[1], **Трис Варкентин**[1], **Аллан Дефо**[1], **Александра Фауст**[1], **Клемент Фарабет**[1] **и Шейн Легг**[1]
[1]Гугл ДипМайнд

Мы предлагаем структуру для классификации возможностей и поведения моделей искусственного общего интеллекта (AGI) и их предшественников. Эта структура представляет уровни производительности, универсальности и автономности AGI. Мы надеемся, что эта структура будет полезна аналогично уровням автономного вождения, предоставляя общий язык для сравнения моделей, оценки рисков и измерения прогресса на пути к AGI. Чтобы разработать нашу структуру, мы анализируем существующие определения AGI и выделяем шесть принципов, которым должна удовлетворять полезная онтология AGI. Эти принципы включают сосредоточение внимания на возможностях, а не на механизмах; раздельная оценка общности и производительности; и определение этапов на пути к ОИИ, а не сосредоточение внимания на конечной точке. Учитывая эти принципы, мы предлагаем «Уровни AGI», основанные на глубине (производительности) и широте (общности) возможностей, и размышляем над тем, как существующие системы вписываются в эту онтологию. Мы обсуждаем сложные требования к будущим тестам, которые количественно оценивают поведение и возможности моделей AGI по сравнению с этими уровнями. Наконец, мы обсудим, как эти уровни AGI взаимодействуют с такими аспектами развертывания, как автономность и риск, и подчеркнем важность тщательного выбора парадигм взаимодействия человека и ИИ для ответственного и безопасного развертывания высокоэффективных систем ИИ.

Ключевые слова: ИИ, AGI, общий искусственный интеллект, общий ИИ, ИИ человеческого уровня, HLAI, ASI, передовые модели, бенчмаркинг, метрики, безопасность ИИ, риск ИИ, автономные системы, взаимодействие человека и ИИ.

arXiv:2311.02462v1 [cs.AI] 4 ноября 2023 г.

Введение

Общий искусственный интеллект (AGI)[1]— это важная и иногда противоречивая концепция в компьютерных исследованиях, используемая для описания системы искусственного интеллекта, которая, по крайней мере, так же способна, как и человек, решать большинство задач. Учитывая быстрое развитие моделей машинного обучения (ML), концепция AGI превратилась из предмета философских дебатов в предмет, имеющий практическое значение в ближайшем будущем. Некоторые эксперты полагают, что «искры» AGI (Бубек и др.,2023 год) уже присутствуют в последнем поколении больших языковых моделей (LLM); некоторые предсказывают, что ИИ в целом превзойдет людей примерно через десятилетие (Бенджио и др., 2023 год); некоторые даже утверждают, что нынешние программы LLM*являются*AGI (Агуэра-и-Аркас и Норвиг,2023 год). Однако если бы вы попросили 100 экспертов по искусственному интеллекту определить, что они подразумевают под «ИИО», вы, скорее всего, получили бы 100 связанных, но разных определений.

Концепция AGI важна, поскольку она отражает цели, прогнозы и риски ИИ:

Цели: Достижение «интеллекта» человеческого уровня является явной или неявной целью для многих специалистов в нашей области, начиная с Дартмутской конференции по искусственному интеллекту 1955 года (Маккарти и др.,1955 год), что положило начало

[1]Существуют разногласия по поводу использования термина «ИИО». Некоторые сообщества отдают предпочтение «Общему ИИ» или «ИИ человеческого уровня» (Грюцемахер и Парадиче,2019 год) в качестве альтернативы или даже просто «ИИ» как термин, который сейчас фактически охватывает ОИИ (или, согласно оптимистическим прогнозам, вскоре будет охватывать его). Однако AGI — это художественный термин, используемый как технологами, так и широкой публикой, и поэтому полезен для четкого общения. Аналогично, для ясности мы используем общепринятые термины, такие как «искусственный интеллект» и «машинное обучение», хотя мы с пониманием относимся к критике (Бигэм,2019 год), что эти термины антропоморфизируют вычислительные системы.

Соответствующий автор(ы): merrie@google.com ©
Google ДипМайнд, 2023. Все права защищены

современной области искусственного интеллекта некоторым ведущим сегодня исследовательским фирмам в области искусственного интеллекта, чьи заявления о миссии отсылают к таким концепциям, как «обеспечить преобразующий искусственный интеллект, помогающий людям и обществу» (антропный,2023а) или «обеспечить, чтобы общий искусственный интеллект принес пользу всему человечеству» (ОпенАИ,2023 год).

Прогнозы: Концепция AGI связана с предсказанием прогресса в области ИИ, а именно, что он движется к большей универсальности, приближаясь и превосходя человеческую общность. Кроме того, AGI обычно переплетается с понятием «эмерджентных» свойств (Вэй и др.,2022 год), то есть возможности, явно не предусмотренные разработчиком. Такие возможности многообещающие, возможно, включая способности, дополняющие типичные человеческие навыки, открывающие новые типы взаимодействия или новые отрасли. Такие прогнозы о возможностях AGI, в свою очередь, предсказывают вероятные социальные последствия; AGI может иметь значительные экономические последствия, т.е. достижение необходимых критериев для повсеместного замещения рабочей силы (Делл'Аква и др.,2023 год;Эллингруд и др.,2023 год), а также геополитические последствия, связанные не только с экономическими преимуществами, которые может дать ОИИ, но и с военными соображениями (Киссинджер и др., 2022 год).

Риски: Наконец, некоторые рассматривают AGI как концепцию определения момента, когда существуют экстремальные риски (Бенджио и др.,2023 год;Шевлан и др.,2023 год), поскольку некоторые предполагают, что системы AGI могут обманывать и манипулировать, накапливать ресурсы, продвигать цели, вести себя агентно, перехитрить людей в широких областях, вытеснять людей с ключевых ролей и/или рекурсивно самосовершенствоваться.

В этой статье мы утверждаем, что исследовательскому сообществу ИИ крайне важно четко задуматься над тем, что мы подразумеваем под «ИИО», и стремиться дать количественную оценку таких характеристик, как производительность, универсальность и автономность систем ИИ. Общие практические определения этих концепций будут способствовать: сравнению моделей; оценки рисков и стратегии их смягчения; четкие критерии со стороны политиков и регулирующих органов; определение целей, прогнозов и рисков для исследований и разработок; и способность понимать и сообщать, на каком этапе пути к ОИИ мы находимся.

Определение AGI: тематические исследования

Многие исследователи и организации в области искусственного интеллекта предложили определения AGI. В этом разделе мы рассматриваем девять ярких примеров и размышляем об их сильных сторонах и ограничениях. Этот анализ лег в основу нашего последующего введения двумерной уровневой онтологии ОИИ.

Пример 1: Тест Тьюринга.Тест Тьюринга (Тьюринг,1950 год), пожалуй, самая известная попытка реализовать концепцию, подобную AGI. «Игра в имитацию» Тьюринга была предложена как способ операционализировать вопрос о том, могут ли машины думать, и предлагала человеку в интерактивном режиме различать, создается ли текст другим человеком или машиной. Тест в первоначальной формулировке представляет собой мысленный эксперимент и является предметом множества критических замечаний (Википедия,2023б); на практике тест часто подчеркивает легкость обмана людей (Вайценбаум,1966 год; Википедия,2023а), а не «интеллект» машины. Учитывая, что современные LLM соответствуют некоторым критериям теста Тьюринга, кажется очевидным, что этого критерия недостаточно для внедрения или сравнительного анализа AGI. Мы согласны с Тьюрингом в том, что вопрос о том, может ли машина «думать», хотя и является интересным философским и научным вопросом, кажется ортогональным вопросу о том, что машина может делать; последнее гораздо проще измерить и оно более важно для оценки воздействия. Поэтому мы предлагаем определять AGI с точки зрения*возможности*скорее, чем*процессы*[2].

Пример 2: Сильный ИИ – системы, обладающие сознанием. Философ Джон Сирл размышлял: «Согласно сильному ИИ, компьютер — это не просто инструмент изучения разума; скорее, правильно запрограммированный компьютер действительно является разумом в том смысле, что компьютеры обладают

[2]Как исследование механистической интерпретируемости (Ройкер и др.,2023 год) развивается, это может позволить использовать метрики, ориентированные на процесс. Они могут иметь отношение к будущим определениям ОИИ.

можно буквально сказать, что правильные программы понимают и имеют другие когнитивные состояния» (Сирл,1980 год). Хотя сильный ИИ может быть одним из путей к достижению ИИ, не существует научного консенсуса относительно методов определения того, обладают ли машины сильными атрибутами ИИ, такими как сознание (Бутлин и др., 2023 год), что делает ориентированную на процесс направленность этой структуры непрактичной.

Пример 3: Аналогии с человеческим мозгом.Первоначально термин «искусственный общий интеллект» был использован в статье Марка Губруда о военных технологиях (1997 г.).Губруд,1997 год), который определил AGI как «системы искусственного интеллекта, которые конкурируют или превосходят человеческий мозг по сложности и скорости, которые могут приобретать, манипулировать и рассуждать на основе общих знаний и которые можно использовать практически на любом этапе промышленных или военных операций, где человеческий интеллект может быть использован». в противном случае понадобится». Это раннее определение подчеркивает*процессы*(соперничающий с человеческим мозгом по сложности) и возможностям; Хотя архитектуры нейронных сетей, лежащие в основе современных систем машинного обучения, во многом вдохновлены человеческим мозгом, успех архитектур на основе трансформаторов (Васвани и др.,2023 год), чья производительность не зависит от обучения, подобного человеческому, предполагает, что строгие процессы и тесты, основанные на мозге, не являются неотъемлемой частью AGI.

Тематическое исследование 4: Эффективность выполнения когнитивных задач на уровне человека.Легг (Легг,2008 год) и Герцель (Герцель,2014 год) популяризировал термин AGI среди ученых-компьютерщиков в 2001 году (Легг,2022 год), описывающий AGI как машину, способную выполнять когнитивные задачи, которые обычно могут выполнять люди. Это определение, в частности, фокусируется на нефизических задачах (т.е. не требует воплощения роботов в качестве предшественника ОИИ). Как и многие другие определения AGI, эта формулировка представляет собой двусмысленность в отношении таких вариантов, как «какие задачи?» и «какие люди?».

Пример 5: Способность изучать задачи.В*Технологическая сингулярность*(Шанахан,2015 год), Шанахан предполагает, что AGI — это «искусственный интеллект, который не специализируется на выполнении конкретных задач, но может научиться выполнять такой же широкий спектр задач, как и человек». Важным свойством этой концепции является акцент на важности включения метакогнитивных задач (обучения) в число требований для достижения ОИИ.

Пример 6: Экономически ценная работа.Устав OpenAI определяет AGI как «высокоавтономные системы, которые превосходят людей в большинстве экономически ценных работ» (ОпенАИ,2018 год). Это определение имеет сильные стороны по критерию «возможности, а не процессы», поскольку оно фокусируется на производительности, независимой от основных механизмов; кроме того, это определение предлагает потенциальный критерий измерения, т.е. экономическую ценность. Недостатком этого определения является то, что оно не охватывает все критерии, которые могут быть частью «общего интеллекта». Существует множество задач, связанных с интеллектом, которые могут не иметь четко определенной экономической ценности (например, художественное творчество или эмоциональный интеллект). Такие свойства могут быть косвенно учтены в экономических показателях (например, художественное творчество может создавать книги или фильмы, эмоциональный интеллект может быть связан со способностью быть успешным генеральным директором), хотя остается неясным, охватывает ли экономическая ценность весь спектр «интеллекта». Еще одна проблема, связанная с определением ОИИ с точки зрения экономической ценности, заключается в том, что это подразумевает необходимость*развертывание* ОИИ, чтобы реализовать эту ценность, тогда как сосредоточение внимания на возможностях может потребовать только*потенциал* для AGI для выполнения задачи. У нас вполне могут быть системы, которые технически способны выполнять экономически важные задачи, но не осознают этой экономической ценности по разным причинам (правовым, этическим, социальным и т. д.).

Практический пример 7: Гибкий и общий – «Кофейный тест» и связанные с ним проблемы.Маркус предполагает, что AGI — это «сокращение от любого интеллекта (их может быть много), который является гибким и общим, с находчивостью и надежностью, сравнимыми с человеческим интеллектом (или превосходящими его)» (Маркус,2022б). Это определение охватывает оба*общности*и*производительность*(за счет включения надежности); упоминание «гибкости» примечательно, поскольку, как и формулировка Шанахана, это предполагает, что*метакогнитивный* такие задачи, как способность изучать новые навыки, должны быть включены в набор возможностей AGI, чтобы достичь достаточной общности. Далее Маркус реализует свое определение, предлагая пять конкретных

задачи (понимание фильма, понимание романа, приготовление пищи на произвольной кухне, написание безошибочной программы на 10 000 строк и преобразование математических доказательств на естественном языке в символическую форму) (Маркус,2022a). Сопровождение определения эталоном имеет ценность; однако потребуется дополнительная работа для создания достаточно всеобъемлющего эталонного показателя. Хотя мы согласны с тем, что невыполнение некоторых из этих задач указывает на то, что система*нет*AGI, неясно, достаточно ли их прохождения для получения статуса AGI. В разделе «Тестирование AGI» мы далее обсуждаем проблему разработки набора задач, который одновременно необходим и достаточен для охвата универсальности AGI. Отметим также, что одно из предложенных Маркусом заданий – «поработать грамотным поваром на произвольной кухне» (вариант «Кофейного теста» Стива Возняка (Возняк,2010 год)), требует роботизированного воплощения; это отличается от других определений, ориентированных на нефизические задачи.[3].

Практический пример 8: Искусственный интеллект.В*Грядущая волна*Сулейман предложил концепцию «Искусственного интеллекта (ACI)» (Мустафа Сулейман и Майкл Бхаскар,2023 год) для обозначения систем искусственного интеллекта, обладающих достаточной производительностью и универсальностью для выполнения сложных, многоэтапных задач в открытом мире. В частности, Сулейман предложил экономически обоснованное определение навыка ACI, которое он назвал «Современным тестом Тьюринга», согласно которому ИИ получит капитал в размере 100 000 долларов и поручит превратить его в 1 000 000 долларов в течение нескольких месяцев. Эта формулировка более узкая, чем определение экономически ценной работы OpenAI, и имеет дополнительный недостаток, заключающийся в потенциальном введении рисков согласования (Кентон и др.,2021 год), ориентируясь только на финансовую прибыль. Однако сильной стороной концепции Сулеймана является сосредоточенность на выполнении сложной, многоэтапной задачи, которую ценят люди. В более широком смысле, чем зарабатывание миллиона долларов, акцент ACI на сложных, реальных задачах заслуживает внимания, поскольку такие задачи могут иметь более*экологическая обоснованность*чем многие текущие тесты искусственного интеллекта; Вышеупомянутые пять тестов Маркуса на гибкость и общность (Маркус,2022a), похоже, также соответствует духу ACI.

Пример 9: Магистр права SOTA как специалисты широкого профиля.Агуэра-и-Аркас и Норвиг (Агуэра-и-Аркас и Норвиг,2023 год) предположил, что современные LLM (например, развертывание GPT-4, Bard, Llama 2 и Claude в середине 2023 г.) уже*являются*АГИ, утверждая, что*общность*является ключевым свойством AGI, и это происходит потому, что языковые модели могут обсуждать широкий круг тем, выполнять широкий спектр задач, обрабатывать мультимодальные входы и выходы, работать на нескольких языках и «обучаться» с нуля или с небольшим количеством шагов. примеры, они достигли достаточной общности. Хотя мы согласны с тем, что общность является важнейшей характеристикой AGI, мы полагаем, что он также должен сочетаться с мерой*производительность*(т. е. если LLM может писать код или выполнять математические операции, но не является надежно корректным, то его универсальность еще недостаточно производительна).

Определение AGI: шесть принципов

Размышляя над этими девятью примерами формулировок ОИИ (или смежных с ОИИ концепций), мы выявляем свойства и общие черты, которые, по нашему мнению, способствуют четкому, практическому определению ОИИ. Мы утверждаем, что любое определение AGI должно соответствовать следующим шести критериям:

1. Сосредоточьтесь на возможностях, а не на процессах.Большинство определений сосредоточены на том, что может сделать ОИИ. выполнить, а не на механизме, с помощью которого он выполняет задачи. Это важно для выявления характеристик, которые не обязательно являются предпосылкой для достижения ОИИ (но, тем не менее, могут представлять собой интересную тему для исследований). Такой акцент на возможностях позволяет нам исключить из наших требований к AGI следующее:

- Достижение ОИИ не означает, что системы*думать*или*понимать*по-человечески (поскольку основное внимание уделяется процессам, а не возможностям)

[3]Хотя робототехника также может подразумеваться под ориентацией устава OpenAI на «экономически ценную работу», тот факт, что OpenAI закрыла свое исследовательское подразделение робототехники в 2021 году (Виггеры,2021 год) предполагает, что это не их предполагаемая интерпретация.

- Достижение ОИИ не означает, что системы обладают такими качествами, как *сознание* (субъективное осознание) (Бутлин и др., 2023 год) или *разум* (способность испытывать чувства) (поскольку эти качества не только имеют процессуальную направленность, но и в настоящее время не поддаются измерению общепринятыми научными методами)

2. Сосредоточьтесь на общности *и* Производительность. Все приведенные выше определения подчеркивают общность. в разной степени, но некоторые исключают критерии эффективности. Мы утверждаем, что общность и производительность являются ключевыми компонентами AGI. В следующем разделе мы представляем уровневую таксономию, которая учитывает взаимодействие между этими измерениями.

3. Сосредоточьтесь на когнитивных и метакогнитивных задачах. Требуется ли роботизированное воплощение (Рой и другие., 2021 год) как критерий AGI является предметом споров. Большинство определений сосредоточено на когнитивных задачах, под которыми мы подразумеваем нефизические задачи. Несмотря на последние достижения в области робототехники (Брохан и др., 2023 год), физические возможности систем ИИ, похоже, отстают от нефизических возможностей. Возможно, что воплощение в физическом мире необходимо для построения знаний о мире и успешного решения некоторых познавательных задач (Шанахан, 2010 год), или, по крайней мере, может быть одним из путей к успеху в некоторых классах когнитивных задач; если это окажется правдой, то воплощение может иметь решающее значение для некоторых путей к AGI. Мы предполагаем, что способность выполнять физические задачи увеличивает общность системы, но не должна считаться необходимой предпосылкой для достижения ОИИ. С другой стороны, метакогнитивные способности (такие как способность изучать новые задачи или способность знать, когда просить разъяснений или помощи у человека) являются ключевыми предпосылками для того, чтобы системы достигли универсальности.

4. Сосредоточьтесь на потенциале, а не на развертывании. Демонстрация того, что система может выполнять необходимые набор задач при заданном уровне производительности должен быть достаточным для объявления системы AGI; развертывание такой системы в открытом мире не должно быть неотъемлемой частью определения ОИИ. Например, определение AGI с точки зрения достижения определенного уровня замещения рабочей силы потребует внедрения в реальном мире, в то время как определение AGI с точки зрения *способный* замены рабочей силы будет сосредоточено на потенциале. Требование развертывания как условия измерения AGI создает нетехнические препятствия, такие как юридические и социальные соображения, а также потенциальные проблемы этики и безопасности.

5. Сосредоточьтесь на экологической обоснованности. Задачи, которые можно использовать для оценки прогресса на пути к ОИИ: имеет решающее значение для практической реализации любого предлагаемого определения. Обсуждая это далее в разделе «Тестирование на AGI», мы подчеркиваем здесь важность выбора задач, которые соответствуют реальным (т. е. экологически обоснованным) задачам, которые люди ценят (трактуя «ценность» в широком смысле, а не только как экономическую ценность). но также социальная ценность, художественная ценность и т. д.). Это может означать отказ от традиционных показателей ИИ, которые легко автоматизировать или количественно оценить (Раджи и др., 2021 год), но может не отражать те навыки, которые люди будут ценить в AGI.

6. Сосредоточьтесь на пути к AGI, а не на одной конечной точке. Примерно так же, как принятие стандартного набора Уровни автоматизации вождения (САЭ Интернешнл, 2021 год) позволили провести четкое обсуждение политики и прогресса в области беспилотных транспортных средств, мы полагаем, что определение «уровней AGI» имеет смысл. Как мы обсудим в последующих разделах, мы намерены, чтобы каждый уровень AGI был связан с четким набором показателей/контрольных показателей, а также с выявленными рисками, возникающими на каждом уровне, и соответствующими изменениями в парадигме взаимодействия человека и ИИ (Моррис и др., 2023 год). Этот основанный на уровне подход к определению AGI поддерживает сосуществование многих известных формулировок – например, определения Агеры-и-Аркаса и Норвига (Агуэра-и-Аркас и Норвиг, 2023 год) попадет в категорию «Новый AGI» нашей онтологии, тогда как порог замены рабочей силы OpenAI (ОпенАИ, 2018 год) лучше соответствует «Виртуозному AGI». Наш уровень «Компетентный AGI», вероятно, является лучшим обобщением для многих существующих определений AGI (например, Legg (Легг, 2008 год), Шанахан (Шанахан, 2015 год) и Сулейман (Мустафа Сулейман и Майкл Бхаскар, 2023 год) составы). В следующем разделе мы представляем онтологию AGI, основанную на уровнях.

Уровни ОИИ

Производительность(ряды) x **общность**(столбцы)	**Узкий** *четко обозначенная задача или набор задач*	**Общий** *широкий спектр нефизических задач, включая метакогнитивные способности, такие как изучение новых навыков*
Уровень 0: Нет ИИ	**Узкий** программа-калькулятор; компилятор	**Общие сведения** вычисления с участием человека, например, Amazon Mechanical Turk
Уровень 1: Развивающийся *равный или несколько лучше, чем неквалифицированный человек*	**Развивающийся узкий ИИ** ГОФАИ4; простые системы, основанные на правилах, например SHRDLU (Виноград, 1971 год)	**Развивающийся искусственный интеллект** ЧатGPT (ОпенАИ,2023 год), Бард (Анил и др.,2023 год), Лама 2 (Туврон и др.,2023 год)
Уровень 2: Компетентный *не менее 50-го процентиля квалифицированных взрослых*	**Компетентный узкий ИИ** детекторы токсичности, такие как Jigsaw (Дас и др.,2022 год); Умные колонки, такие как Siri (Яблоко), Алекса (Амазонка) или Google Ассистент (Google); Системы VQA, такие как PaLI (Чен и др.,2023 год); Ватсон (ИБМ); SOTA LLM для подмножества задач (например, написание короткого эссе, простое программирование)	**Компетентный искусственный интеллект** еще не достигнуто
Уровень 3: Эксперт *не менее 90-го процентиля квалифицированных взрослых*	**Экспертный узкий ИИ** средства проверки орфографии и грамматики, такие как Grammarly (Грамматика,2023 год); генеративные модели изображений, такие как Imagen (Сахария и др.,2022 год) или Далл-Э 2 (Рамеш и др.,2022 год)	**Экспертный АГИ** еще не достигнуто
Уровень 4: Виртуоз *не менее 99-го процентиля квалифицированных взрослых*	**Виртуозный узкий ИИ** Темно-синий (Кэмпбелл и др., 2002 г.), АльфаГо (Сильвер и др., 2016 год,2017 год)	**Виртуозный AGI** еще не достигнуто
Уровень 5: Сверхчеловек *превосходит 100% людей*	**Сверхчеловеческий узкий ИИ** АльфаФолд (Джампер и др.,2021 год; Варади и др.,2021 год), АльфаЗеро (Сильвер и др.,2018 год), СтокФиш (вяленая рыба,2023 год)	**Искусственный сверхинтеллект (ИСИ)** еще не достигнуто

Таблица 1 | Уровневый матричный подход к классификации систем на пути к ОИИ на основе глубины (производительности) и широты (общности) возможностей. Примеры систем в каждой ячейке представляют собой приближения, основанные на текущих описаниях в литературе или опыте взаимодействия с развернутыми системами. Однозначная классификация систем ИИ потребует стандартизированного эталона задач, как мы обсуждаем в разделе «Тестирование AGI».

В соответствии с принципом 2 («Фокус на общность и производительность») и принципом 6 («Фокус на пути к AGI, а не на единой конечной точке»), в таблице1мы вводим матричную систему прокачки, которая фокусируется на *производительности*и*общность*как два измерения, которые являются ключевыми для AGI:

- **Производительность**относится к*глубина*возможностей системы искусственного интеллекта, т. е. как она соотносится с производительностью человеческого уровня для выполнения данной задачи. Обратите внимание, что для всех уровней успеваемости выше «Нового» процентили относятся к выборке взрослых, обладающих соответствующими навыками (например, «Компетентный» или более высокая производительность при выполнении такой задачи, как умение писать на английском языке, будет измеряться только по набору взрослые, грамотные и свободно владеющие английским языком).
- **общность**относится к*широта*возможностей системы ИИ, т. е. диапазона задач, для которых система ИИ достигает целевого порога производительности.

Эта таксономия определяет минимальную производительность при выполнении большинства задач, необходимых для достижения определенного рейтинга — например, компетентный AGI должен иметь производительность не ниже 50-го процентиля для квалифицированных взрослых людей при выполнении большинства когнитивных задач, но может иметь экспертную, виртуозную или даже сверхчеловеческую производительность. по подмножеству задач. В качестве примера того, как отдельные системы могут охватывать разные точки нашей таксономии, мы утверждаем, что на момент написания этой статьи в сентябре 2023 года пограничные языковые модели (например, ChatGPT (ОпенАИ,2023 год), Бард (Анил и др.,2023 год), Лама2 (Туврон и др.,2023 год) и т. д.) демонстрируют «компетентный» уровень производительности для некоторых задач (например, написание коротких эссе, простое программирование), но все еще находятся на «развивающемся» уровне производительности для большинства задач (например, математические способности, задачи, связанные с фактами). В целом, текущие модели передового языка будут считаться общим ИИ уровня 1 («Новый AGI») до тех пор, пока уровень производительности не повысится для более широкого набора задач (на этом этапе критерии общего ИИ уровня 2, «Компетентный AGI», будут встретил). Мы предлагаем, чтобы документация для передовых моделей ИИ, например карты моделей (Митчелл и др.,2019 год), следует подробно описать эту смесь уровней производительности. Это поможет конечным пользователям, политикам и другим заинтересованным сторонам прийти к общему и детальному пониманию вероятной неравномерной производительности систем, идущих по пути к ОИИ.

Порядок, в котором приобретаются более сильные навыки в определенных когнитивных областях, может иметь серьезные последствия для безопасности ИИ (например, приобретение глубоких знаний в области химического машиностроения до приобретения сильных навыков этического рассуждения может быть опасной комбинацией). Также обратите внимание, что скорость прогресса между уровнями производительности и/или общности может быть нелинейной. Приобретение способности приобретать новые навыки может особенно ускорить переход на следующий уровень.

Хотя эта таксономия классифицирует системы в соответствии с их производительностью, системы, которые*способный* достижения определенного уровня производительности (например, по заданному эталону) может не соответствовать этому уровню *на практике*при развертывании. Например, ограничения пользовательского интерфейса могут снизить производительность развертывания. Рассмотрим пример ДАЛЛЕ-2 (Рамеш и др.,2022 год), который мы оцениваем как узкий ИИ уровня 3 («Узкий ИИ-эксперт») в нашей таксономии. Мы оцениваем уровень производительности «Эксперт», поскольку DALLE-2 выдает изображения более высокого качества, чем способно нарисовать большинство людей; однако у системы есть режимы сбоя (например, рисование рук с неправильным количеством цифр, отображение бессмысленного или неразборчивого текста), которые не позволяют ей достичь статуса «виртуозно». Хотя теоретически это система уровня «Эксперт»,*на практике*система может быть только «компетентной», поскольку интерфейсы подсказок слишком сложны для большинства конечных пользователей, чтобы обеспечить оптимальную производительность (о чем свидетельствует существование торговых площадок (например, (PromptBase)) в котором подсказки продают опытные инженеры-подсказки). Это наблюдение подчеркивает важность разработки экологически обоснованных тестов (которые будут измерять развернутую, а не идеализированную производительность), а также важность рассмотрения того, как парадигмы взаимодействия человека и ИИ взаимодействуют с понятием AGI (тема, к которой мы возвращаемся в разделе «Возможности и возможности»). раздел «Автономия»).

Самый высокий уровень в нашей матрице с точки зрения совокупной производительности и общности — это ИСИ (искусственный сверхинтеллект). Мы определяем «сверхчеловеческую» производительность как превосходящую 100% людей. Например, мы полагаем, что AlphaFold (Джампер и др.,2021 год;Варади и др.,2021 год) является узким ИИ 5-го уровня («Сверхчеловеческий узкий ИИ»), поскольку он выполняет одну задачу (предсказание трехмерной структуры белка на основе аминокислотной последовательности) выше уровня ведущих ученых мира. Это определение означает, что системы общего ИИ («ASI») уровня 5 смогут выполнять широкий спектр задач на уровне,

человек может соответствовать. Кроме того, эта формулировка также подразумевает, что сверхчеловеческие системы могут быть способны выполнять даже более широкий круг задач, чем более низкие уровни ОИИ, поскольку способность выполнять задачи, которые качественно отличаются от существующих человеческих навыков, по определению превзойдет всех людей (которые принципиально не могут этого сделать). такие задачи). Например, нечеловеческие навыки, которыми мог бы обладать ИСИ, могли бы включать в себя такие возможности, как нейронные интерфейсы (возможно, посредством таких механизмов, как анализ сигналов мозга для декодирования мыслей).Тан и др.,2023 год)), пророческие способности (возможно, с помощью таких механизмов, как анализ больших объемов данных для получения высококачественных прогнозов) или способность общаться с животными (возможно, с помощью таких механизмов, как анализ закономерностей в их вокализации, мозговых волнах или языке тела). .

Тестирование на AGI

Два из шести предложенных нами принципов определения AGI (принцип 2: общность*и*Производительность; Принцип 6: Сосредоточьтесь на пути к AGI) повлиял на наш выбор матричной, уровневой онтологии для облегчения детального обсуждения широты и глубины возможностей ИИ. Остальные четыре принципа (Принцип 1: Возможности, а не процессы; Принцип 3: Когнитивные и метакогнитивные задачи; Принцип 4: Потенциал, а не развертывание; и Принцип 5: Экологическая обоснованность) относятся к вопросу измерения.

В то время как наш*производительность*измерение определяет один аспект измерения (например, диапазоны процентилей выполнения задач по отношению к определенным подгруппам людей), наши*общность*измерение оставляет открытыми важные вопросы: каков набор задач, составляющих критерии общности? Какую долю таких задач должна освоить система ИИ, чтобы достичь заданного уровня общности в нашей схеме? Существуют ли какие-то задачи, которые всегда необходимо выполнять, чтобы соответствовать критериям определенных уровней общности, например метакогнитивные задачи?

Для практической реализации определения AGI необходимо ответить на эти вопросы, а также разработать конкретные разнообразные и сложные задачи. Из-за огромной сложности этого процесса, а также важности включения широкого спектра точек зрения (включая межорганизационные и междисциплинарные точки зрения), мы не предлагаем ориентир в этой статье. Вместо этого мы работаем над разъяснением онтологии, которую должен попытаться измерить тест. Мы также обсуждаем свойства, которыми должен обладать тест AGI.

Наша цель состоит в том, чтобы тест AGI включал широкий набор когнитивных и метакогнитивных задач (в соответствии с Принципом 3), измеряющих разнообразные свойства, включая (но не ограничиваясь) лингвистический интеллект, математическое и логическое мышление (Уэбб и др.,2023 год), пространственное мышление, межличностный и внутриличностный социальный интеллект, способность осваивать новые навыки и творческий потенциал. Эталон *мощь* включать тесты, охватывающие психометрические категории, предложенные теориями интеллекта в психологии, нейробиологии, когнитивной науке и образовании; однако такие «традиционные» тесты сначала необходимо оценить на предмет их пригодности для сравнительного тестирования вычислительных систем, поскольку многим из них может не хватать экологической и конструктивной валидности в этом контексте (Серапио-Гарсия и др.,2023 год).

Один из открытых вопросов для сравнительного анализа производительности заключается в том, разрешать ли использование инструментов, в том числе потенциально основанных на искусственном интеллекте, в качестве средства повышения производительности человека. Этот выбор может в конечном итоге зависеть от задачи и должен учитывать экологическую обоснованность выбора эталонных показателей (согласно Принципу 5). Например, при определении того, является ли беспилотный автомобиль достаточно безопасным, сравнение с человеком, управляющим автомобилем без использования каких-либо современных инструментов безопасности с помощью искусственного интеллекта, будет не самым информативным сравнением; поскольку соответствующий контрфакт включает в себя некоторую технологию помощи водителю, мы можем предпочесть сравнение с этим базовым уровнем.

Хотя тест AGI может основываться на некоторых существующих тестах AI (Линч,2023 год) (например, HELM (Лян и др.,2023 год), БОЛЬШАЯ-скамейка (Шривастава и др.,2023 год)) мы также предполагаем включение

открытые и/или интерактивные задачи, которые могут потребовать качественной оценки (Бубек и др.,2023 год; Папакириакопулос и др.,2021 год;Ян и др.,2023 год). Мы подозреваем, что эти последние классы сложных, открытых задач, хотя и трудно поддающиеся сравнительному анализу, будут иметь большую экологическую обоснованность, чем традиционные показатели ИИ или адаптированные традиционные меры человеческого интеллекта.

Невозможно перечислить весь набор задач, решаемых достаточно общим интеллектом. Таким образом, тест AGI должен быть*жизнь*эталон. Таким образом, такой ориентир должен включать в себя основу для постановки и согласования новых задач.

Определение того, что что-то есть*нет*AGI на данном уровне просто требует выявления нескольких[5]задачи, которые люди обычно могут выполнять, но система не может выполнять их адекватно. Можно предположить, что системы, которые проходят большую часть предполагаемого теста AGI на определенном уровне производительности («Новый», «Компетентный» и т. д.), включая новые задачи, добавленные тестировщиками, имеют соответствующий уровень общности для практических целей (то есть, хотя теоретически все еще может быть тест, который AGI потерпит неудачу, в какой-то момент неисследованные отказы настолько специализированы или нетипичны, что становятся практически нерелевантными).

Разработка эталонного показателя AGI будет сложным и итеративным процессом. Тем не менее, это ценная цель для исследовательского сообщества в области ИИ. Измерение сложных концепций может быть несовершенным, но само измерение помогает нам четко определить наши цели и является индикатором прогресса.

Риск в контексте: автономия и взаимодействие человека и искусственного интеллекта

Обсуждения ОИИ часто включают в себя обсуждение риска, в том числе «x-риска» – экзистенциального (для безопасности ИИ, 2023 год) или другие очень экстремальные риски (Шевлан и др.,2023 год). Уровневый подход к определению AGI позволяет более детально обсудить, как различные комбинации производительности и общности связаны с различными типами рисков ИИ. Несмотря на то, что рассмотрение сценариев экстремального риска имеет ценность, понимание ОИИ через предложенную нами онтологию, а не как единую конечную точку (согласно Принципу 6), может помочь гарантировать, что политики также выявляют и расставляют приоритеты рисков в краткосрочной перспективе и на пути к ОИИ.

Уровни AGI как основа оценки рисков

По мере того, как мы продвигаемся по уровням наших возможностей к ASI, появляются новые риски, включая риски неправильного использования, риски согласования и структурные риски (Цветслут и Дефо,2019 год). Например, уровень «Эксперт AGI», вероятно, будет включать в себя структурные риски, связанные с экономическими потрясениями и сокращением рабочих мест, поскольку все больше и больше отраслей достигают порога замены машинного интеллекта вместо человеческого труда. С другой стороны, достижение «Экспертного AGI», вероятно, снижает некоторые риски, связанные с «Новым AGI» и «Компетентным AGI», например, риск неправильного выполнения задачи. На уровнях «Виртуозный AGI» и «ASI» наиболее вероятно возникновение многих проблем, связанных с риском x (например, ИИ, который может превзойти своих людей-операторов в широком спектре задач, может обмануть их для достижения неправильно заданного результата). цель, как в мысленных экспериментах по смещению (христианин,2020 год)).

Системные риски, такие как дестабилизация международных отношений, могут вызывать беспокойство, если скорость перехода между уровнями опережает регулирование или дипломатию (например, первая страна, достигшая ASI, может иметь существенное геополитическое/военное преимущество, создавая сложные структурные риски). На уровнях ниже

[5]Мы не решаемся указать точное количество или процент задач, которые система должна выполнить на данном уровне производительности, чтобы ее можно было объявить общим ИИ на этом уровне (например, такое правило, как «система должна пройти не менее 90% тест AGI на заданном уровне производительности, чтобы получить этот рейтинг»). Хотя мы думаем, что это будет очень высокий процент, он, вероятно, не будет 100%, поскольку кажется очевидным, что широкая, но несовершенная общность имеет большое значение (отдельные люди также не имеют стабильной производительности во всех возможных задачах, но в целом остаются умными). Определение того, какая часть задач бенчмаркинга на данном уровне демонстрирует общность, остается открытым исследовательским вопросом.

«Экспертный ИИ» (например, «Новый ИИ», «Компетентный ИИ» и все «узкие» категории ИИ), скорее всего, связаны с действиями человека (например, риски неправильного использования ИИ, будь то случайного, случайного или злонамеренного). Более полный анализ профилей рисков, связанных с каждым уровнем, является важным шагом на пути к разработке таксономии AGI, которая может служить руководством для исследований в области безопасности/этики и разработки политики.

Мы признаем, что тест AGI должен включать тесты на потенциально опасные возможности (например, способность обманывать, убеждать (Вирабадран и др.,2023 год) или выполнить расширенную биохимию (Моррис,2023 год)) спорно. Мы склоняемся к включению таких возможностей в бенчмаркинг, поскольку большинство таких навыков, как правило, имеют двойное назначение (имеют действенное применение как в социально положительных сценариях, так и в гнусных сценариях). При сравнительном анализе опасных возможностей можно снизить риски с помощью Принципа 4 (Потенциал, а не развертывание), гарантируя, что тесты для любых опасных задач или задач двойного назначения должным образом изолированы и не определены с точки зрения развертывания. Однако включение таких тестов в общедоступные тесты может позволить злоумышленникам оптимизировать эти возможности; понимание того, как снизить риски, связанные со сравнительным анализом способностей двойного назначения, остается важной областью исследований экспертов по безопасности ИИ, этике ИИ и управлению ИИ.

Одновременно с этой работой Anthropic выпустила версию 1.0 своей Политики ответственного масштабирования (RSP) (антропный,20236). В этой политике используется подход, основанный на уровнях (вдохновленный стандартами уровня биобезопасности), для определения уровня риска, связанного с системой ИИ, определения того, какие опасные возможности могут быть связаны с каждым уровнем безопасности ИИ (ASL), и какие меры по сдерживанию или развертыванию следует предпринять. приниматься на каждом уровне. Текущие генеративные ИИ SOTA классифицируются как риск ASL-2. Включение элементов, соответствующих возможностям ASL, в любой тест AGI свяжет точки в нашей таксономии AGI с конкретными рисками и мерами по снижению рисков.

Возможности против автономности

Хотя возможности создают предпосылки для рисков ИИ, системы ИИ (включая системы AGI) не работают и не будут работать в вакууме. Скорее, системы искусственного интеллекта развертываются с определенными интерфейсами и используются для решения определенных задач в определенных сценариях. Эти контекстуальные атрибуты (интерфейс, задача, сценарий, конечный пользователь) оказывают существенное влияние на профили рисков. Возможности AGI сами по себе не определяют судьбу риска, но их следует рассматривать в сочетании с контекстными деталями.

Рассмотрим, например, возможности пользовательских интерфейсов для систем AGI. Расширение возможностей открывает новые парадигмы взаимодействия, но*не определяй их*. Скорее, разработчики систем и конечные пользователи остановятся на режиме взаимодействия человека и ИИ (Моррис и др.,2023 год), который уравновешивает множество соображений, включая безопасность. Мы предлагаем охарактеризовать парадигмы взаимодействия человека и ИИ с помощью шести **Уровни автономии**, описанный в таблице2.

Эти уровни автономии коррелируют с уровнями AGI. Более высокие уровни автономии «разблокируются» при развитии возможностей AGI, хотя более низкие уровни автономии могут быть желательны для определенных задач и контекстов (в том числе по соображениям безопасности), даже когда мы достигаем более высоких уровней AGI. Тщательно продуманный выбор в отношении взаимодействия человека и ИИ имеет жизненно важное значение для безопасного и ответственного развертывания передовых моделей ИИ.

Мы подчеркиваем важность парадигмы «Без ИИ». Может быть много ситуаций, когда это желательно, в том числе по соображениям образования, удовольствия, оценки или безопасности. Например, в сфере беспилотных транспортных средств, когда технология беспилотного управления 5-го уровня широко доступна, могут быть причины для использования транспортного средства 0-го уровня (без автоматизации). К ним относятся инструктаж нового водителя (обучение), удовольствие от вождения любителями вождения (удовольствие), экзамены на получение водительских прав (оценка) или условия, когда на датчики нельзя положиться, например, технологические сбои или экстремальные погодные явления (безопасность). Во время самостоятельного вождения 5-го уровня (САЭ Интернешнл,2021 год) транспортные средства, скорее всего, будут уровня

Уровень автономности	Примеры систем	Разблокировка Уровень(и) AGI	Примеры рисков Представлено
Уровень автономности 0: Нет ИИ *человек делает все-вещь*	Аналоговые подходы (например, рисование карандашом на бумаге) Не-ИИ цифровой рабочие процессы (например, ввод текста в текстовом редакторе; рисование в программе рисования)	Нет ИИ	н/д (риски статус-кво)
Уровень автономности 1: ИИ как инструмент *человек полностью контролирует задачу и использует ИИ для автоматизации рутинных подзадачи*	Поиск информации с помощью поисковой системы Проверка написанного с помощью программы проверки грамматики. Чтение вывески с помощью приложения машинного перевода	Возможный: Развивающийся узкий ИИ Вероятный: Компетентный Узкий ИИ	деквалификация (например, чрезмерная уверенность) нарушение учредил отрасли промышленности
Уровень автономности 2: ИИ в качестве консультанта *ИИ берет на себя существенную роль, но только при вызове человек*	Использование языковой модели для обобщения набора документов Ускорение компьютерного программирования с помощью модели генерации кода Использование большей части развлечений с помощью сложной системы рекомендаций.	Возможный: Компетентный Узкий ИИ Вероятный: Экспертный узкий ИИ; Развивающийся искусственный интеллект	чрезмерное доверие радикализация целевой манипуляция
Уровень автономности 3: ИИ как Соавтор *равный человек-ИИ сотрудничество; меж-активная координация целей и задач*	Обучение шахматисту посредством взаимодействия и анализа ИИ, играющего в шахматы. Развлечение с помощью Социальное взаимодействие с личностями, созданными ИИ	Возможный: Развивающийся искусственный интеллект Вероятный: Экспертный узкий ИИ; Компетентный искусственный интеллект	антропоморфизация (например, парасоциальный отношения) быстрые социальные изменения
Уровень автономности 4: ИИ как эксперт *ИИ управляет взаимодействием; человек обеспечивает руководство и обратная связь или выполняет подзадачи*	Использование системы искусственного интеллекта для продвижения научных открытий (например, сворачивания белков)	Возможный: Виртуозный узкий ИИ Вероятный: Экспертный АГИ	скука общественного масштаба массовый труд перемещение упадок человечества исключительность
Уровень автономности 5: ИИ как агент *полностью автономный ИИ*	Автономный на базе искусственного интеллекта личные помощники *(пока не разблокирован)*	Вероятный: Виртуозный AGI; КАК и Я	перекос концентрация власти

Таблица 2 | Более мощные системы ИИ открывают новые парадигмы взаимодействия человека и ИИ (включая полностью автономный ИИ). Выбор соответствующего уровня автономии не обязательно должен быть максимально достижимым с учетом возможностей базовой модели. Одним из соображений при выборе уровня автономии являются связанные с этим риски. Примеры в этой таблице иллюстрируют важность тщательного рассмотрения решений по проектированию взаимодействия человека и ИИ.

5 Узкий ИИ («Сверхчеловеческий Узкий ИИ») согласно нашей таксономии[6], те же соображения относительно автономности человека и компьютера применимы и к AGI. Мы можем разработать AGI, но предпочитаем не развертывать его автономно (или предпочитаем развертывать его с дифференцированными уровнями автономии в различных обстоятельствах, которые диктуются контекстуальными соображениями).

Определенные аспекты общности могут потребоваться для того, чтобы сделать желательными конкретные парадигмы взаимодействия. Например, уровни автономии 3, 4 и 5 («Сотрудник», «Эксперт» и «Агент») могут работать хорошо только в том случае, если система ИИ также демонстрирует высокие показатели определенных метакогнитивных способностей (научившись, когда спрашивать человека о помощь, теория моделирования разума, социально-эмоциональные навыки). В нашем определении уровня автономии 5 («ИИ как агент») подразумевается, что такой полностью автономный ИИ может действовать согласованно без постоянного человеческого надзора, но знает, когда следует проконсультироваться с людьми (Шах и др., 2021 год). Интерфейсы, которые поддерживают согласованность действий человека и искусственного интеллекта посредством более точной спецификации задач, преодоления разрывов в процессах и оценки результатов (Терри и др.,2023 год) являются жизненно важной областью исследований, позволяющей гарантировать, что область взаимодействия человека и компьютера идет в ногу с проблемами и возможностями взаимодействия с системами AGI.

Парадигма взаимодействия человека и искусственного интеллекта как основа оценки рисков

Стол2иллюстрирует взаимодействие между уровнем AGI, уровнем автономии и риском. Достижения в производительности и универсальности моделей открывают дополнительные возможности выбора парадигмы взаимодействия (включая потенциально полностью автономный ИИ). Эти парадигмы взаимодействия, в свою очередь, привносят новые классы рисков. Взаимодействие возможностей модели и проектирования взаимодействия позволит проводить более детальную оценку рисков и принимать ответственные решения по развертыванию, чем рассмотрение только возможностей модели.

Стол2также предоставляет конкретные примеры каждого из шести предлагаемых уровней автономии. Для каждого уровня автономии мы указываем соответствующие уровни производительности и общности, которые «открывают» эту парадигму взаимодействия (т. е. уровни AGI, на которых возможно или вероятно успешное развертывание и принятие этой парадигмы).

Наши прогнозы относительно уровней «разблокировки», как правило, требуют более высоких уровней производительности для узких систем ИИ, чем для общих; например, мы полагаем, что использование ИИ в качестве консультанта вероятно либо с экспертным узким ИИ, либо с развивающимся AGI. Это несоответствие отражает тот факт, что для общих систем развитие возможностей, вероятно, будет неравномерным; например, общий ИИ уровня 1 («Новый AGI»), вероятно, будет иметь производительность уровня 2 или, возможно, даже уровня 3 при выполнении некоторого подмножества задач. Такая неравномерность возможностей обычных ИИ может открыть более высокие уровни автономии для выполнения конкретных задач, соответствующих их конкретным сильным сторонам.

Рассмотрение систем AGI в контексте использования людьми позволяет нам задуматься о взаимодействии между достижениями в моделях и достижениями в парадигмах взаимодействия человека и ИИ. Роль исследований по построению моделей можно рассматривать как содействие развитию возможностей систем на пути к ОИИ в их производительности и общности, так что способности системы ИИ будут перекрывать все большую часть человеческих способностей. И наоборот, роль исследования взаимодействия человека и ИИ можно рассматривать как обеспечение того, чтобы новые системы ИИ были*годный к употреблению*и*полезный*людям так, что системы искусственного интеллекта успешно расширяют возможности людей (т.е. «увеличение интеллекта» (Бриньольфссон,2022 год)).

[6]Согласно нашей таксономии, полностью автономные транспортные средства можно отнести к узкому ИИ уровня 4 («Виртуозный узкий ИИ»); однако мы подозреваем, что на практике автономным транспортным средствам, возможно, потребуется достичь сверхчеловеческого стандарта производительности, чтобы добиться широкого общественного признания в отношении восприятия безопасности, что иллюстрирует важность контекстуальных соображений.

Заключение

Общий искусственный интеллект (AGI) — это концепция, имеющая как желательные, так и практические последствия. В этой статье мы проанализировали девять известных определений AGI, выявив сильные и слабые стороны. На основе этого анализа мы представляем шесть принципов, которые, по нашему мнению, необходимы для четкого и практичного определения ОИИ: сосредоточение внимания на возможностях, а не на процессах; ориентируясь на общность*и*производительность; сосредоточение внимания на когнитивных и метакогнитивных (а не физических) задачах; сосредоточение внимания на потенциале, а не на развертывании; сосредоточение внимания на экологической обоснованности задач сравнительного анализа; и сосредоточить внимание на пути к созданию AGI, а не на одной конечной точке.

Помня об этих принципах, мы представили онтологию «Уровни AGI», которая предлагает более детальный способ определения нашего продвижения к AGI, рассматривая общность (узкую или общую) в тандеме с пятью уровнями производительности (Новые, Компетентные, Экспертные, Виртуозные). и сверхчеловек). Мы поразмышляли над тем, как существующие системы искусственного интеллекта и определения AGI вписываются в эту структуру. Кроме того, мы обсудили последствия наших принципов для разработки живого, экологически обоснованного эталона AGI и утверждаем, что такое начинание (хотя оно наверняка будет сложным) является жизненно важным для нашего сообщества.

Наконец, мы рассмотрели, как наши принципы и онтология могут изменить дискуссию о рисках, связанных с ОИИ. Примечательно, что мы заметили, что AGI не обязательно является синонимом автономии. Мы ввели уровни автономии, которые открываются, но не определяются при прохождении уровней AGI. Мы проиллюстрировали, как рассмотрение уровня AGI совместно с уровнем автономии может дать более детальное представление о вероятных рисках, связанных с системами ИИ, подчеркнув важность инвестиций в исследования взаимодействия человека и ИИ в тандеме с усовершенствованием моделей.

Благодарности

Благодарим членов команды Google DeepMind PAGI за поддержку этих усилий, а также Мартина Ваттенберга, Майкла Терри, Джеффри Ирвинга, Мюррея Шанахана, Дилипа Джорджа и Блеза Агуэра-и-Аркаса за полезные обсуждения этой темы.

Рекомендации

Блез Агуэра-и-Аркас и Питер Норвиг. Общий искусственный интеллект — это готов Здесь. Ноэма, октябрь 2023 г. URL-адресhttps://www.noemamag.com/общий-искусственный интеллект-уже-здесь/.

Амазонка. Амазонка Алекса. URL-адресhttps://alexa.amazon.com/. доступ 20 октября 2023 г.

Рохан Анил, Эндрю М. Дай, Орхан Фират и др. Технический отчет PaLM 2.*КОРР*, абс/2305.10403, 2023. doi: 10.48550/arXiv.2305.10403. URL-адресhttps://arxiv.org/abs/2305.10403.

Антропный. Компания: Антропик, 2023а. URL-адресhttps://www.anthropic.com/company. Доступ 12 октября 2023 г.

Антропный. Политика ответственного масштабирования Anthropic, сентябрь 2023 г.b. URL-адресhttps://www-файлы.anthropic.com/production/files/responsible-scaling-policy-1.0.pdf. доступ 20 октября 2023 г.

Яблоко. Сири. URL-адресhttps://www.apple.com/siri/. доступ 20 октября 2023 г.

Йошуа Бенджио, Джеффри Хинтон, Эндрю Яо, Дон Сонг, Питер Аббил, Юваль Ноа Харари, Я-Цин Чжан, Лан Сюэ, Шай Шалев-Шварц, Джиллиан Хэдфилд, Джефф Клун, Теган Махарадж, Фрэнк Хаттер,

Атылым Гюнеш Байдин, Шейла Макилрайт, Цици Гао, Ашвин Ачарья, Дэвид Крюгер, Анка Драган, Филип Торр, Стюарт Рассел, Дэниел Канеман, Ян Браунер и Серен Миндерманн. Управление рисками ИИ в эпоху быстрого прогресса.*КОРР*, abs/2310.17688, 2023. doi: 10.48550/arXiv.2310.17688. URL-адрес https://arxiv.org/abs/2310.17688.

Джеффри Бигэм. Грядущая осень искусственного интеллекта. Сообщение в блоге, 2019 г. URL https://jeffreybigham.com/blog/2019/the-coming-ai-autumnn.html.

Энтони Броэн, Ноа Браун, Джастис Карбахал, Евгений Чеботар, Си Чен, Кшиштоф Хоромански Тианли Дин, Дэнни Дрисс, Авинава Дубей, Челси Финн, Пит Флоренс, Чуюань Фу, Монтсе Гонсалес Аренас, Киртана Гопалакришнан, Кехан Хан, Кароль Хаусман, Александр Херцог, Жасмин Сюй, Брайан Ихтер, Алекс Ирпан, Никхил Джоши, Райан Джулиан, Дмитрий Калашников, Юхэн Куанг, Изабель Леал, Лиза Ли, Цанг-Вей Эдвард Ли, Сергей Левин, Яо Лу, Генрик Михалевский, Игорь Мордач, Карл Перч, Канишка Рао, Криста Рейманн, Майкл Рю, Гресия Салазар, Паннаг Санкети, Пьер Сермане, Джаспиар Сингх, Аникаит Сингх, Раду Сорикут, Хуонг Чан, Винсент Ванхук, Куан Выонг, Айзаан Вахид, Стефан Велкер, Пол Уолхарт, Цзялин Ву, Фей Ся, Тед Сяо, Пэн Сюй, Сычунь Сюй, Тяньхэ Ю и Брианна Зиткович. RT-2: Модели «видение-язык-действие» передают веб-знания роботизированному управлению.*КОРР*, abs/2307.15818, 2023. doi: 10.48550/arXiv.2307.15818. URL-адрес https://arxiv.org/abs/2307.15818.

Эрик Бриньольфссон. Ловушка Тьюринга: перспективы и опасности искусственного интеллекта, подобного человеческому.*КОРР*, abs/2201.04200, 2022. doi: 10.48550/arXiv.2201.04200. URL-адрес https://arxiv.org/abs/2201.04200.

Себастьен Бюбек, Варун Чандрасекаран, Ронен Эльдан, Йоханнес Герке, Эрик Хорвиц, Эсе Камар, Питер Ли, Инь Тат Ли, Юаньчжи Ли, Скотт Лундберг, Харша Нори, Хамид Паланги, Марко Тулио Рибейро и И Чжан. Искры общего искусственного интеллекта: ранние эксперименты с GPT-4. *КОРР*, abs/2303.12712, 2023. doi: 10.48550/arXiv.2303.12712. URL-адрес https://arxiv.org/абс/2303.12712.

Патрик Батлин, Роберт Лонг, Эрик Эльмознино, Йошуа Бенджио, Джонатан Бёрч, Аксель Констант, Джордж Дин, Стивен М. Флеминг, Крис Фрит, Сюй Цзи, Рёта Канаи, Колин Кляйн, Грейс Линдси, Матиас Мишель, Лиад Мудрик, Меган А.К. Питерс, Эрик Швицгебель, Джонатан Саймон и Руфин Ван-Руллен. Сознание в искусственном интеллекте: идеи науки о сознании.*КОРР*, abs/2308.08708, 2023. doi: 10.48550/arXiv.2308.08708. URL-адрес https://arxiv.org/abs/2308.08708.

Мюррей Кэмпбелл, А. Джозеф Хоан и Фэн-сюн Сюй. Темно-синий.*Артиф. Интел.*, 134(1–2):57–83, январь 2002 г. ISSN 0004-3702. дои: 10.1016/S0004-3702(01)00129-1. URL-адрес https://doi.org/10.1016/C0004-3702(01)00129-1.

Си Чен, Сяо Ван, Соравит Чанпиньо и др. PaLI: многоязычный язык совместного масштабирования. Модель изображения.*КОРР*, abs/2209.06794, 2023. doi: 10.48550/arXiv.2209.06794. URL-адрес https://arxiv.org/abs/2209.06794.

Брайан Кристиан.*Проблема выравнивания*. WW Нортон и компания, 2020.

Миллон Мадхур Дас, Пуньяджой Саха и Митхун Дас. Какой из них более токсичен? Выводы из Головоломки Оцените серьезность токсичных комментариев.*КОРР*, abs/2206.13284, 2022. doi: 10.48550/arXiv.2206.13284. URL-адрес https://arxiv.org/abs/2206.13284.

Фабрицио Делл'Аква, Эдвард МакФоулэнд, Итан Р. Моллик, Хила Лифшиц-Ассаф, Кэтрин Келлогг, Саран Раджендран, Лиза Крайер, Франсуа Канделон и Карим Р. Лакхани. Навигация по Jagged

Технологический рубеж: полевые экспериментальные данные о влиянии искусственного интеллекта на производительность и качество работников умственного труда.*Рабочий документ Отдела технологий и управления операциями Гарвардской школы бизнеса, номер 24-013*, сентябрь 2023 г.

Квейлин Эллингруд, Саураб Сангви, Гурнит Сингх Дандона, Ану Мадгавкар, Майкл Чуи, Оливия Уайт и Пейдж Хасебе. Генеративный искусственный интеллект и будущее работы в Америке. McKinsey In-Глобальный отчет Института, июль 2023 г. URLhttps://www.mckinsey.com/mgi/our-research/generative-ai-and-the-future-of-work-in-america.

Центр безопасности искусственного интеллекта. Заявление о рисках ИИ, 2023 г. URL-адресhttps://www.safe.ai/заявление о риске.

Бен Герцель. Общий искусственный интеллект: концепция, современное состояние и перспективы.*Журнал общего искусственного интеллекта*, 01 2014. doi: 10.2478/jagi-2014-0001.

Google. Google Assistant, ваш личный Google. URL-адресhttps://assistant.google.com/. доступ 20 октября 2023 г.

Грамматика, 2023. URLhttps://www.grammarly.com/.

Росс Грюцемахер и Дэвид Б. Парадис. Альтернативные методы сопоставления путей с HLAI.*КОРР*, abs/1905.00614, 2019. doi: 10.48550/arXiv.1905.00614. URL-адресhttp://arxiv.org/abs/1905.00614.

Марк Губруд. Нанотехнологии и международная безопасность.*Пятая Форсайт-конференция по молекулярной Нанотехнологии*, ноябрь 1997 г.

ИБМ. IBM Ватсон. URL-адресhttps://www.ibm.com/watson. доступ 20 октября 2023 г.

Джон Джампер, Ричард Эванс, Александр Притцель, Тим Грин, Майкл Фигурнов, Олаф Роннебергер, Кэтрин Туньясувунакул, Расс Бейтс, Огюстин Жидек, Анна Потапенко, Алекс Бриджланд, Клеменс Мейер, Саймон А.А. Коль, Эндрю Дж. Баллард, Эндрю Коуи, Бернардино Ромера-Паредес, Станислав Николов, Ришуб Джайн, Йонас Адлер, Тревор Бэк, Стиг Петерсен, Дэвид Рейман, Эллен Клэнси, Михал Зелински, Мартин Штайнеггер, Михалина Пачольска, Тамаш Бергхаммер, Себастьян Боденштейн, Дэвид Сильвер, Ориол Виньялс, Эндрю В. Сениор, Корай Кавукчуоглу, Пушмит Кохли и Демис Хассабис. Высокоточное предсказание структуры белка с помощью AlphaFold.*Природа*, 596:583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Закари Кентон, Том Эверитт, Лаура Вайдингер, Ясон Гэбриэл, Владимир Микулик и Джеффри Ирвинг. Согласование языковых агентов.*КОРР*, abs/2103.14659, 2021. doi: 10.48550/arXiv.2103.14659. URL-адресhttps://arxiv.org/abs/2103.14659.

Генри Киссинджер, Эрик Шмидт и Дэниел Хуттенлохер.*Эпоха искусственного интеллекта*. Книги Бэк-Бэй, ноябрь 2022.

Шейн Легг. Машина Супер Интеллект. Докторская диссертация представлена на факультете информатики Университета Лугано, июнь 2008 г.

Шейн Легг. Twitter (теперь «X»), май 2022 г. URLhttps://twitter.com/ShaneLegg/status/1529483168134451201. Доступ осуществлен 12 октября 2023 г.

Перси Лян, Риши Боммасани, Тони Ли и др. Целостная оценка языковых моделей.*КОРР*, abs/2211.09110, 2023. doi: 10.48550/arXiv.2211.09110. URL-адресhttps://arxiv.org/abs/2211.09110.

Шана Линч. Тесты AI достигли насыщения. Стэнфордский блог об искусственном интеллекте, ориентированном на человека, Апрель 2023 г. URLhttps://hai.stanford.edu/news/ai-benchmarks-hit-saturation.

Гэри Маркус. Дорогой Илон Маск, вот пять вещей, о которых вы, возможно, захотите подумать. АГИ. Подстек «Маркус об искусственном интеллекте», май 2022 г.a. URL-адресhttps://garymarcus.substack.com/p/Дорогой-Илон-Маск-здесь-пять-вещей?s=r.

Гэри Маркус. Twitter (теперь «X»), май 2022 г.b. URL-адресhttps://twitter.com/GaryMarcus/status/1529457162811936768. Доступ осуществлен 12 октября 2023 г.

Дж. Маккарти, М.Л. Мински, Н. Рочестер и К.Э. Шеннон. Предложение для Дартмутского летнего исследовательского проекта по искусственному интеллекту. Дартмутский семинар, 1955 год.

Маргарет Митчелл, Симона Ву, Эндрю Залдивар, Паркер Барнс, Люси Вассерман, Бен Хатчинсон, Елена Спитцер, Иниолува Дебора Раджи и Тимнит Гебру. Карты моделей для отчетов по моделям. В *Материалы конференции по справедливости, подотчетности и прозрачности*. ACM, январь 2019 г. doi: 10.1145/3287560.3287596. URL-адресhttps://doi.org/10.1145%2F3287560.3287596.

Мередит Рингел Моррис. Взгляды ученых на потенциал генеративного искусственного интеллекта в своих областях. *КОРР*, abs/2304.01420, 2023. doi: 10.48550/arXiv.2304.01420. URL-адресhttps://arxiv.org/абс/2304.01420.

Мередит Рингел Моррис, Кэрри Дж. Кай, Джесс Холбрук, Чинмей Кулкарни и Майкл Терри. Дизайн Пространство порождающих моделей.*КОРР*, abs/2304.10547, 2023. doi: 10.48550/arXiv.2304.10547. URL-адрес https://arxiv.org/abs/2304.10547.

Мустафа Сулейман и Майкл Бхаскар.*Грядущая волна: технологии, власть и XXI век Величайшая дилемма*. Корона, сентябрь 2023 г.

ОпенАИ. Хартия OpenAI, 2018. URLhttps://openai.com/charter. По состоянию на 12 октября 2023 г.

ОпенАИ. OpenAI: О проекте, 2023. URLhttps://openai.com/about. По состоянию на 12 октября 2023 г.

ОпенАИ. Технический отчет GPT-4.*КОРР*, abs/2303.08774, 2023. doi: 10.48550/arXiv.2303.08774. URL-адресhttps://arxiv.org/abs/2303.08774.

Орестис Папакириакопулос, Элизабет Энн Уоткинс, Эми Вайнкофф, Клаудия Язвиньска и Тити Чаттопадхай. Качественный анализ ИИ, ориентированного на человека.*КОРР*, абс/2112.03784, 2021. дои: 10.48550/arXiv.2112.03784. URL-адресhttps://arxiv.org/abs/2112.03784.

ПромптБаза. PromptBase: Торговая площадка Prompt. URL-адресhttps://promptbase.com/. доступ на 20 октября 2023 г.

Иниолува Дебора Раджи, Эмили М. Бендер, Амандалин Пауллада, Эмили Дентон и Алекс Ханна. ИИ и все в целом мире.*КОРР*, абс/2111.15366, 2021. дои: 10.48550/arXiv.2111.15366. URL-адресhttps://arxiv.org/abs/2111.15366.

Адитья Рамеш, Прафулла Дхаривал, Алекс Никол, Кейси Чу и Марк Чен. Иерархический текст-Генерация условного изображения с помощью CLIP Latents. Апрель 2022 г. URLhttps://cdn.openai.com/документы/dall-e-2.pdf.

Тилман Ройкер, Энсон Хо, Стивен Каспер и Дилан Хэдфилд-Менелл. На пути к прозрачному ИИ: А Обзор по интерпретации внутренних структур глубоких нейронных сетей.*КОРР*, abs/2207.13243, 2023. doi: 10.48550/arXiv.2207.13243. URL-адресhttps://arxiv.org/abs/2207.13243.

Николас Рой, Ингмар Познер, Тим Барфут, Филипп Бодуан, Йошуа Бенджио, Жаннетт Бог, Оливер Брок, Изабель Депати, Дитер Фокс, Дэн Кодичек, Томас Лозано-Перес, Викаш Мансингка, Кристофер Пал, Блейк Ричардс, Дорса Садиг, Стефан Шаал, Гаурав Сукхатме, Денис Териен, Марк Туссен и Мишель Ван де Панн. От машинного обучения к робототехнике: проблемы и возможности воплощенного интеллекта.*КОРР*, abs/2110.15245, 2021. doi: 10.48550/arXiv.2110.15245. URLhttps://arxiv.org/abs/2110.15245.

САЭ Интернешнл. Таксономия и определения терминов, связанных с системами автоматизации вождения для дорожных транспортных средств, апрель 2021 г. URLhttps://www.sae.org/standards/content/j3016_202104. По состоянию на 12 октября 2023 г.

Читван Сахария, Уильям Чан, Сауараб Саксена, Лала Ли, Джей Ванг, Эмили Дентон, Сейед Кам-Яр Сейед Гасемипур, Бурку Карагол Аян, С. Сара Махдави, Рафа Гонтихо Лопес, Тим Салиманс, Джонатан Хо, Дэвид Дж. Флит и Мохаммад Норузи. Фотореалистичные модели диффузии текста в изображение с глубоким пониманием языка.*КОРР*, абс/2205.11487, 2022. дои: 10.48550/arXiv.2205.11487. URL-адресhttps://arxiv.org/abs/2205.11487.

Джон Р. Сирл. Разум, мозг и программы.*Поведенческие и мозговые науки*, 3:417–424, 1980. дои: 10.1017/S0140525X00005756.

Грег Серапио-Гарсиа, Мустафа Сафдари, Клеман Крепи, Лунинг Сан, Стивен Фитц, Питер Ромеро, Марва Абдулхай, Александра Фауст и Майя Матарич. Личностные черты в больших языковых моделях. *КОРР*, abs/2307.00184, 2023. doi: 10.48550/arXiv.2307.00184. URL-адресhttps://arxiv.org/абс/2307.00184.

Рохин Шах, Педро Фрейре, Нил Алекс, Рэйчел Фридман, Дмитрий Крашенинников, Лоуренс Чан, Майкл Д. Деннис, Питер Аббил, Анка Драган и Стюарт Рассел. Преимущества помощи по сравнению с Вознаграждение за обучение, 2021 г. URLhttps://openreview.net/forum?id=DFIOGDZejIB.

Мюррей Шанахан.*Воплощение и внутренняя жизнь*. Издательство Оксфордского университета, 2010.

Мюррей Шанахан.*Технологическая сингулярность*. MIT Press, август 2015 г.

Тоби Шевлейн, Себастьян Фаркуар, Бен Гарфинкел, Мэри Фуонг, Джесс Уиттлстоун, Джейд Люн, Дэниел Кокотайло, Нахема Маршал, Маркус Андерлюнг, Ноам Кольт, Льюис Хо, Дивья Сиддарт, Шахар Авин, Уилл Хокинс, Бин Ким, Ясон Габриэль, Виджай Болина, Джек Кларк, Йошуа Бенджио, Пол Кристиано и Аллан Дефо. Оценка модели для экстремальных рисков.*КОРР*, abs/2305.15324, 2023. doi: 10.48550/arXiv.2305.15324. URL-адресhttps://arxiv.org/abs/2305.15324.

Дэвид Сильвер, Аджа Хуанг, Крис Дж. Мэддисон, Артур Гез, Лоран Сифре, Джордж ван ден Дриссе, Джулиан Шритвизер, Иоаннис Антоноглу, Веда Паннеершельвам, Марк Ланкто, Сандер Дилеман, Доминик Греве, Джон Нэм, Нал Кальхбреннер, Илья Суцкевер, Тимоти Лилликрап, Мадлен Лич, Корай Кавукчуоглу, Торе Грепель и Демис Хассабис. Освоение игры в го с помощью глубоких нейронных сетей и поиска по дереву.*Природа*, 529:484–489, 2016. doi: 10.1038/nature16961.

Дэвид Сильвер, Джулиан Шритвизер, Карен Симонян, Иоаннис Антоноглу, Аджа Хуанг, Артур Гез, Томас Юбер, Лукас Бейкер, Мэтью Лай, Адриан Болтон, Ютьян Чен, Тимоти Лилликрап, Фань Хуэй, Лоран Сифре, Джордж ван ден Дриссше, Торе Грапель и Демис Хассабис. Освоение игры Го без человеческого знания.*Природа*, 550:354–359, 2017. doi: 10.1038/nature24270.

Дэвид Сильвер, Томас Хьюберт, Джулиан Шритвизер, Иоаннис Антоноглу, Мэттью Лай, Артур Гез Марк Ланкто, Лоран Сифре, Даршан Кумаран, Торе Грепель, Тимоти Лилликрап, Карен Симонян и Демис Хассабис. Общий алгоритм обучения с подкреплением, позволяющий освоить шахматы, сёги и

Пройдите Самостоятельную игру.*Наука*, 362(6419):1140–1144, 2018. doi: 10.1126/science.aar6404. URL-адрес https://www.science.org/doi/abs/10.1126/science.aar6404.

Аарохи Шривастава, Абхинав Растоги, Абхишек Рао и др. За пределами игры в имитацию: количественная оценка и экстраполяция возможностей языковых моделей.*КОРР*, абс/2206.04615, 2023. дои: 10.48550/arXiv.2206.04615. URL-адресhttps://arxiv.org/abs/2206.04615.

Вяленая рыба. Stockfish — шахматный движок с открытым исходным кодом, 2023 г. URLhttps://stockfishchess.org/.

Джерри Танг, Аманда ЛеБел, Шейли Джайн и Александр Г. Хут. Семантическая реконструкция Непрерывный язык на основе неинвазивных записей мозга.*Природа Нейронауки*, 26:858–866, 2023. doi: 10.1038/s41593-023-01304-9.

Майкл Терри, Чинмей Кулкарни, Мартин Ваттенберг, Лукас Диксон и Мередит Рингел Моррис. Согласование ИИ при разработке интерактивного ИИ: согласование спецификаций, согласование процессов и поддержка оценки.*КОРР*, abs/2311.00710, 2023. doi: 10.48550/arXiv.2311.00710. URL-адрес https://arxiv.org/abs/2311.00710.

Уго Туврон, Луи Мартен, Кевин Стоун, Питер Альберт, Амджад Альмахаири, Ясмин Бабаи, Николай Башлыков, Сумья Батра, Праджвал Бхаргава, Шрути Бхосале, Дэн Бикел, Лукас Блечер, Кристиан Кантон Феррер, Моя Чен, Гиллем Кукурулл, Дэвид Эсиобу, Джуд Фернандес, Джереми Фу, Венинь Фу, Брайан Фуллер, Синтия Гао, Веданудж Госвами, Наман Гоял , Энтони Хартсхорн, Сагар Хоссейни, Руй Хоу, Хакан Инан, Марцин Кардас, Виктор Керкез, Мадиан Хабса, Изабель Клуманн, Артем Коренев, Пунит Сингх Кура, Мари-Анн Лашо, Тибо Лавриль, Женя Ли, Диана Лискович, Инхай Лу, Юнинг Мао, Ксавье Мартине, Тодор Михайлов, Пушкарь Мишра, Игорь Молибог, Исинь Не, Эндрю Поултон, Джереми Райзенштейн, Раши Рунгта, Калян Салади, Алан Шелтен, Руан Силва, Эрик Майкл Смит, Ранжан Субраманиан, Сяоцин Эллен Тан, Бинь Тан, Росс Тейлор, Адина Уильямс, Цзянь Сян Куан, Пусинь Сюй, Чжэн Янь, Илиян Заров, Ючен Чжан, Анджела Фан, Мелани Камбадур, Шаран Наранг, Орельен Родригес, Роберт Стойник, Сергей Эдунов и Томас Сьялом. Лама 2: Открытый фонд и точно настроенные модели чата, 2023 г.

А. М. Тьюринг. Вычислительная техника и интеллект.*Разум*, LIX: 433–460, октябрь 1950 г. URL. https://doi.org/10.1093/mind/LIX.236.433.

Михали Варади, Стивен Аньянго, Мандар Дешпанде, Шринат Наир, Синди Натассиа, Галабина Йорданова, Дэвид Юан, Оана Строу, Джемма Вуд, Агата Лэйдон, Огюстин Жидек, Тим Грин, Кэтрин Туньясувунакул, Стиг Петерсен, Джон Джампер, Эллен Клэнси, Ричард Грин, Анкур Вора, Мира Лютфи, Майкл Фигурнов, Эндрю Коуи, Николь Хоббс, Пушмит Кохли, Джерард Клейвегт, Юэн Бирни, Демис Хассабис и Самир Веланкар. База данных структуры белков AlphaFold: массовое расширение структурного охвата пространства белковых последовательностей с помощью моделей высокой точности.*Исследования нуклеиновых кислот*, 50: D439–D444, 11 2021 г. ISSN 0305–1048. doi: 10.1093/nar/gkab1061. URL-адрес https://doi.org/10.1093/nar/gkab1061.

Ашиш Васвани, Ноам Шазир, Ники Пармар, Якоб Ушкорейт, Лайон Джонс, Эйдан Н. Гомес, Лукаш Кайзер и Илья Полосухин. Внимание — это все, что вам нужно.*КОРР*, абс/1706.03762, 2023. дои: 10.48550/arXiv.1706.03762. URL-адресhttps://arxiv.org/abs/1706.03762.

В. Вирабадран, Дж. Голдман, С. Шанкар и др. Влиять как на человеческое, так и на машинное восприятие. 10.1038/c41467-023-40499-0. Тонкие состязательные манипуляции с изображениями *Природные коммуникации*, 14, 2023. дои:

Тейлор Уэбб, Кейт Дж. Холиок и Хунцзин Лу. Возникновение аналогичных рассуждений в больших масштабах Модели датчиков.*Природа Человеческое поведение*, 7:1526–1541, 2023. URL.https://doi.org/10.1038/s41562-023-01659-w.

Джейсон Вэй, Йи Тай, Риши Боммасани, Колин Раффел, Баррет Зоф, Себастьян Боржо, Дэни Йогатама Маартен Босма, Денни Чжоу, Дональд Мецлер, Эд Х. Чи, Тацунори Хашимото, Ориол Виньялс, Перси Лян, Джефф Дин и Уильям Федус. Новые возможности больших языковых моделей.*KOPP*, abs/2206.07682, 2022. doi: 10.48550/arXiv.2206.07682. URL-адресhttps://arxiv.org/abs/2206.07682.

Йозеф Вайценбаум. ELIZA — компьютерная программа для изучения общения на естественном языке. между Человеком и Машиной.*Коммун. ACM*, 9 (1): 36–45, январь 1966 г. ISSN 0001-0782. дои: 10.1145/365153.365168. URL-адресhttps://doi.org/10.1145/365153.365168.

Кайл Виггерс. OpenAI распускает исследовательскую группу в области робототехники. VentureBeat, июль 2021 г. URLhttps://venturebeat.com/business/openai-disbands-its-robotics-research-team/.

Википедия. Евгений Густман - Википедия, The Бесплатно Энциклопедия. https://en.wikipedia.org/wiki/Eugene_Goostman, 2023a. По состоянию на 12 октября 2023 г.

Википедия. Тест Тьюринга: Слабые стороны — Википедия, Свободная энциклопедия. https://en.wikipedia.org/wiki/Turing_test, 2023b. По состоянию на 12 октября 2023 г.

Терри Виноград. Процедуры как представление данных в компьютерной программе для понимания Естественный язык.*Технические отчеты MIT AI*, 1971.

Стив Возняк. Может ли компьютер приготовить чашку кофе? Интервью компании Fast: https://www.youtube.com/watch?v=MowergwQR5Y, 2010 г.

Чжэнъюань Ян, Линьцзе Ли, Кевин Линь, Цзяньфэн Ван, Чун-Цзин Линь, Цзычэн Лю и Лицзюань Ван. Рассвет LMM: предварительные исследования с GPT-4V(ision).*KOPP*, abs/2309.17421, 2023. doi: 10.48550/arXiv.2309.17421. URL-адресhttps://arxiv.org/abs/2309.17421.

Ремко Цветслут и Аллан Дефо. Размышляя о рисках, связанных с ИИ: несчастные случаи, неправильное использование и Состав.*Законность*, 11:2019, 2019. URLhttps://www.lawfaremedia.org/article/ think-about-risks-ai-accidents-misuse-and-structure.