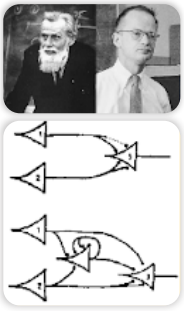
ЧЕТВЁРТАЯ РЕВОЛЮЦИЯ

В ГОЛОВОКРУЖИТЕЛЬНОМ СКАЧКЕ AI-ТЕХНОЛОГИЙ МОЖНО ВЫДЕЛИТЬ ТРИ РЕВОЛЮЦИИ, КАЖДАЯ ИЗ КОТОРЫХ МЕНЯЛА НАШИ ПРЕДСТАВЛЕНИЯ О ВОЗМОЖНОСТЯХ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА. СЕГОДНЯ МОЖНО ГОВОРИТЬ О КОНТУРАХ СЛЕДУЮЩЕЙ, КОТОРЫЕ УЖЕ НЕСЛОЖНО РАЗЛИЧИТЬ В ТУМАНЕ БУДУЩЕГО.

СЕРГЕЙ НИКОЛЕНКО, ДОКТОР ФИЗИКО-МАТЕМАТИЧЕСКИХ НАУК, ЗАВЕДУЮЩИЙ ЛАБОРАТОРИЕЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА САНКТ-ПЕТЕРБУРГСКОГО ОТДЕЛЕНИЯ МАТЕМАТИЧЕСКОГО ИНСТИТУТА ИМ. В.А. СТЕКЛОВА (ПОМИ РАН)



1943
Математическая модель нейрона Маккаллоха и Питтса

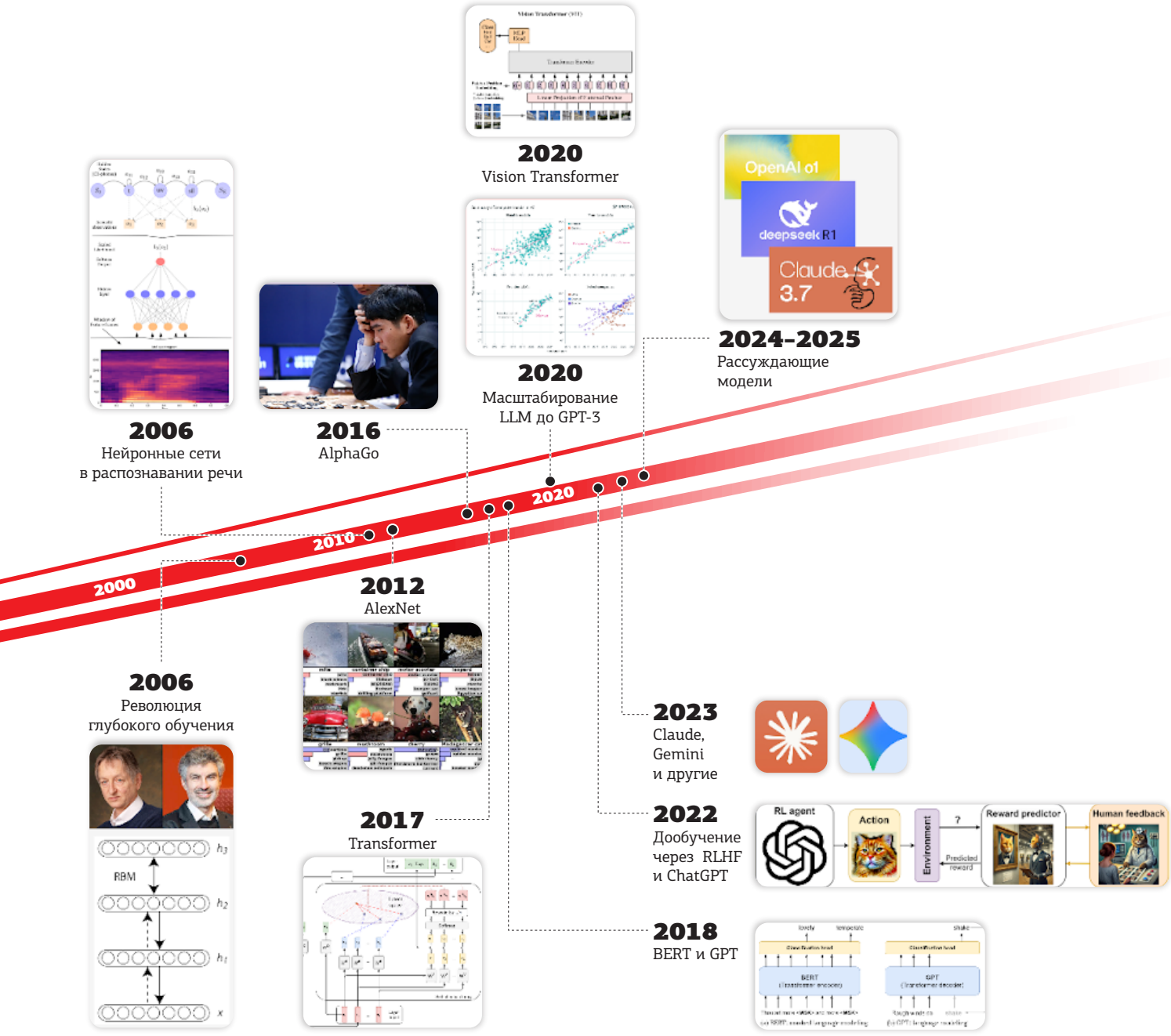
1940 1950 1960 1970 1980 1990

1958
Перцептрон Розенблатта

История развития искусственного интеллекта

Двадцать пять лет назад, на пороге нового тысячелетия, сильный искусственный интеллект казался далёкой, а то и несбыточной мечтой. Нейронные сети существовали только в академической среде, передним краем искусственного интеллекта были рекомендательные системы и поиск в интернете, а идея машины, способной поддержать осмысленный разговор или создать произведение искусства, оставалась уделом фантастов.

Сегодня AI-системы пишут код, ставят медицинские диагнозы, создают музыку и изображения, ведут переговоры и даже помогают в научных исследованиях. Более того, они уже начинают участвовать в разработке следующего поколения AI-систем. В этой ста-



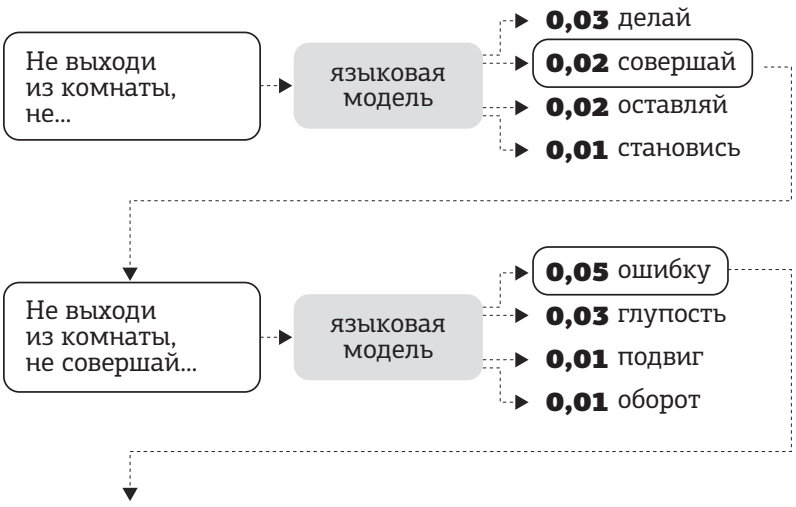
тье мы рассмотрим этапы развития технологии и заглянем в будущее.

Революция глубокого обучения: когда нейросети наконец заработали

Начало пути. История искусственных нейронных сетей началась ещё до того, как AI оформился как научная дисциплина. Первые математические модели нейронов и их взаимодействий появились уже в 1940–х годах, а перцептрон Розенблатта, который в 1958 году стал одной из первых реализованных на практике моделей машинного обучения, был по сути моделью одного нейрона. Метод обратного распространения ошибки, которым обучаются глубокие нейросети,

представляет собой просто дифференцирование сложной функции и к нейросетям был успешно применён уже в 1970-х.

Но в XX веке нейросети оставались скорее предметом академических исследований, чем практическим инструментом. Они работали на игрушечных задачах и демонстрировали принципиальную возможность своего обучения, но неизменно проигрывали более простым методам. «Нейросети — это второй лучший способ сделать всё что угодно», — говорил в начале 1990-х Джон Денкер.



Революция глубокого обучения. Всё изменилось в середине 2000-х. С математической, идейной стороны Джеффри Хинтон и его коллеги представили новый способ, который позволял обучать глубокие нейронные сети; аналогичный прорыв произошёл и в группе Йошуа Бенджи.

Но даже важнее, чем новые алгоритмические идеи, было то, что технологическая база к этому времени тоже созрела для успеха нейросетей. Графические процессоры (GPU), изначально созданные для трёхмерной графики в видеоиграх, оказались идеальным инструментом для обучения нейросетей. Матричные операции, составляющие основу вычислений в нейронных сетях, выполнялись на GPU в десятки и сотни раз быстрее, чем на обычных процессорах. Одновременно развитие интернета породило огромный поток данных: миллионы изображений, тысячи часов видео, терабайты текста. У нейросетей наконец-то появилось и достаточно мощное «железо», и пища для обучения.

Нейросети шагают по планете. Первой практически важной областью применения нейросетей стало тогда распознавание речи: появившиеся в начале 2010-х голосовые ассистенты были бы невозможны без обработки речевых сигналов теми самыми глубокими сетями Хинтона.

Символическим моментом революции стал 2012 год, когда на главном соревновании по распознаванию изображений (на датасете ImageNet) команда Джеффри Хинтона представила свёрточную нейронную сеть AlexNet. Она не просто победила, она уничтожила конкурентов, снизив лучший показатель ошибки с 26% примерно до 14%. Это был огромный качественный скачок, и с тех пор каждый год победителями этого соревнования становились исключительно нейронные сети (архитектуры которых, конечно, менялись и улучшались со временем).

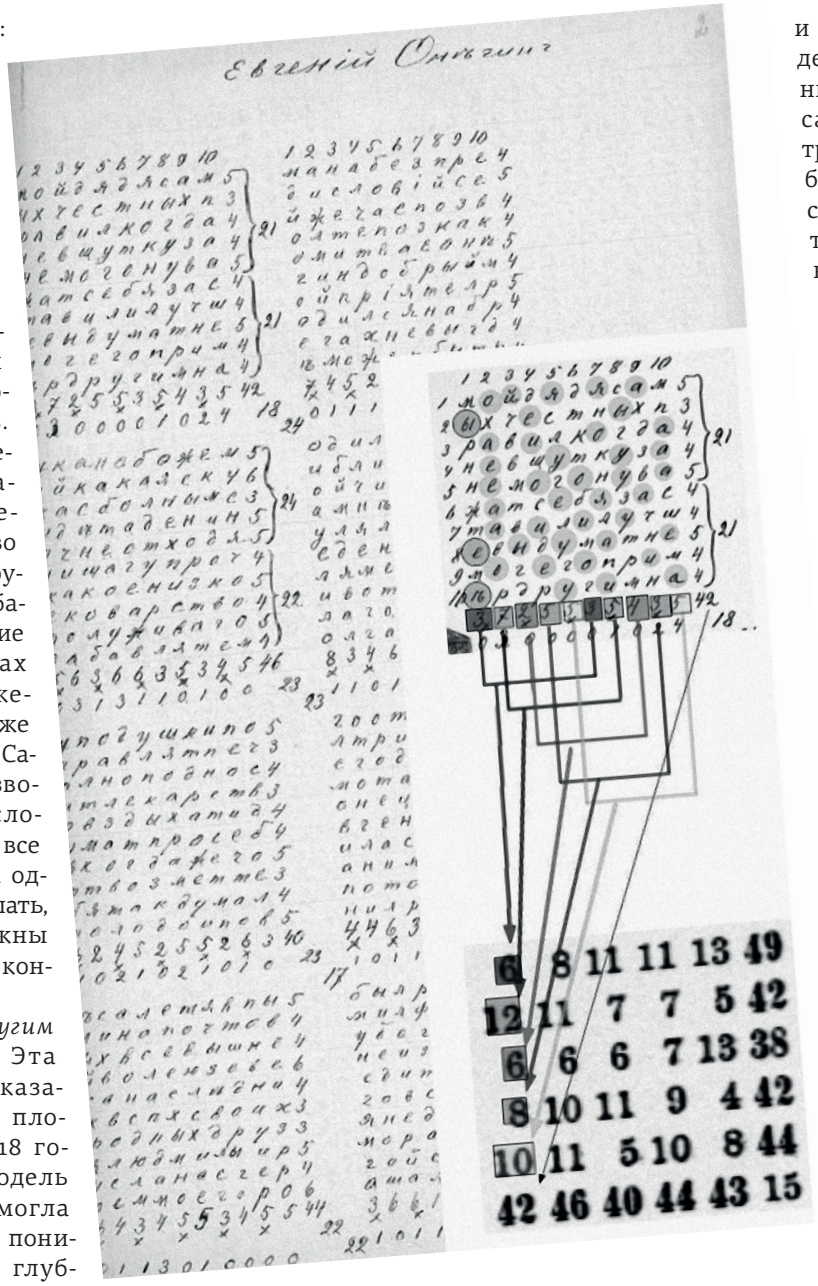
А в 2016 году AlphaGo, основанная на глубоких нейронных сетях, победила Ли Седоля, одного из ведущих профессионалов в игре го. Ранее эта игра всегда считалась слишком сложной для компьютеров из-за астрономического числа возможных позиций (поиск по дереву в го не работает совсем), и победы AlphaGo не ожидал практически никто — ни профессионалы го, ни специалисты по искусственному интеллекту.

За эти 10 лет глубокие нейросети стали доминирующей парадигмой в машинном обучении. Но у них были свои ограничения. Свёрточные сети хорошо работали с изображениями, рекуррентные — с последовательностями вроде текста или временных рядов, но каждая архитектура была заточена под свой тип данных, обучение оставалось медленным, а масштабирование — проблематичным.

Революция трансформеров

Что такое трансформер. В 2017 году группа исследователей из Google опубликовала статью с провокационным названием Attention is All You Need («Внимание — это всё, что вам нужно»). В ней описывалась новая архитектура нейронных сетей — трансформер (Transformer). На первый взгляд это была просто ещё одна архитектура для обработки последовательностей, конкурент для классических рекуррентных нейронных сетей. Но быстро стало ясно, что это нечто большее.

Ключевая идея трансформеров — механизм самовнимания (self-attention). Представьте, что вы читае-



те предложение: «Кошка, которая жила у соседей и которую я часто видел во дворе, убежала». Чтобы понять, кто именно убежал, вам нужно связать слово «убежала» со словом «кошка», хотя между ними много других слов. Рекуррентные сети обрабатывали текст последовательно, слово за словом, и с трудом могли обрабатывать даже такие связи в пределах одного предложения, не говоря уже о более далёких. Самовнимание позволяет каждому слову «смотреть» на все остальные слова одновременно и решать, какие из них важны для понимания контекста.

От языка к другим модальностям. Эта простая идея оказалась невероятно плодотворной. В 2018 году появилась модель BERT, которая могла читать тексты и понимать их гораздо глубже, чем все предыдущие; здесь «понимать» означает преобразовывать в семантически богатые представления, при помощи которых потом можно решать разные задачи. В пару к BERT появилось и семейство моделей GPT, о которых мы поговорим ниже.

Но революция трансформеров оказалась шире, чем просто обработка текста. Исследователи быстро поняли, что та же архитектура работает и для изображений (в 2020 году вышел Vision Transformer, который стал основой для очень многих архитектур), и для звука, и для видео. Идея самовнимания оказалась универсальной. Более того, трансформеры можно было комбинировать с другими типами сетей, создавая гибридные архитектуры.

Масштабирование. Но, пожалуй, самое важное свойство трансформеров — это их способность к масштабированию. Трансформеры можно разделить на тысячи параллельных вычислений и обучать на сотнях и тысячах GPU одновременно. Исследователи обнаружили удивительный эмпирический закон: качество работы трансформеров предсказуемо растёт с увеличением размера модели, объёма данных и вычислительных ресурсов.

Эти законы масштабирования (scaling laws) перевернули индустрию. Раньше прогресс в машинном обучении достигался в основном за счёт новых архитектурных решений, новых моделей. А теперь появилась простая, почти механическая формула успеха: больше параметров, больше данных, больше вычислений — лучше результат, причём предсказуемо лучше. Это породило «гонку вооружений», результаты которой мы видим сегодня.

Революция языковых моделей

Откуда взять данные? Но и это ещё не всё. То самое масштабирование привело к тому, что размеченных

данных для обучения стало категорически не хватать. Когда модели стали содержать миллиарды параметров, большие датасеты «обычного» глубокого обучения, вроде ImageNet, перестали казаться большими.

Решение пришло из неожиданной области. Вместо того чтобы размечать данные вручную, можно использовать задачи с саморазметкой (self-supervision), где правильные ответы получаются автоматически, без участия людей. Самая естественная такая задача для текста — языковое моделирование (language modeling): предсказание следующего слова по предыдущим. Возьмите любой текст из интернета, оборвите в случайном месте и попросите модель предсказать следующее слово — и вот у вас уже есть обучающий пример. А в интернете триллионы слов.

Языковые модели. Задача языкового моделирования, кстати, тоже была всегда. Ещё в 1913 году Андрей Андреевич Марков построил вероятностную модель последовательностей букв в «Евгении Онегине» — первую языковую модель в истории. Простые языковые модели десятилетиями использовались в распознавании речи и машинном переводе, помогая выбрать более вероятный вариант интерпретации. Но, конечно, никто не ожидал, что они смогут писать связный текст или отвечать на сложные вопросы.

И здесь сработало масштабирование трансформеров. В 2018 году OpenAI выпустила GPT — первую большую языковую модель на основе трансформеров. В 2019-м появилась GPT-2 с 1,5 млрд параметров, которая уже могла порождать довольно убедительные тексты. Исследователи из OpenAI даже побоялись выкладывать её в открытый доступ. В 2020-м вышла GPT-3 со 175 млрд параметров — и тут стало окончательно ясно, что происходит что-то экстраординарное.

GPT-3 могла не просто порождать убедительный текст. Она могла переводить, резюмировать, отвечать на вопросы, писать код, сочинять стихи — и всё это без дополнительного обучения на конкретных задачах, просто на основе нескольких примеров в запросе. Модель могла обобщаться на новые задачи прямо во время использования.

Всё это уже произвело революцию в академических кругах, но на публику она вышла только в ноябре 2022 года, когда OpenAI выпустила ChatGPT. Это была та же GPT-3, но дообученная на диалогах и с использованием обратной связи от людей (reinforcement learning from human feedback, RLHF). ChatGPT мог поддерживать связный разговор, помнить контекст, признавать ошибки, отвечать на поставленные вопросы и отказываться от неподходящих запросов. И им могли пользоваться все — через простой веб-интерфейс.

Скорость прогресса. За 5 дней ChatGPT набрал миллион пользователей. За 2 месяца — 100 млн. Это была самая быстрорастущая потребительская технология в истории. Дальше были GPT-4 и GPT-5 от OpenAI, семейства Claude от Anthropic и Gemini от Google, семейства открытых моделей вроде Llama или DeepSeek и многое другое. Началась гонка больших языковых моделей (large language models, LLM).

Сегодня LLM помогают программистам писать код, юристам — анализировать договоры, врачам — формулировать диагнозы, студентам — учиться, писателям — бороться с творческим кризисом. Они встроены в поисковики, текстовые редакторы, системы разработки. Большими языковыми моделями так или иначе пользуются сотни миллионов людей ежедневно.

И прогресс не останавливается. В 2022 году GPT-3 было нелегко справиться с задачами для третьеклассников вроде «У Васи было три теннисных мячика, и он купил ещё две упаковки по четыре; сколько у него теперь мячиков?» А в 2025-м GPT-5 и Gemini 2.5 Pro уже способны самостоятельно решать сложные математические задачи, как олимпиадные, так и исследовательские. Важным прорывом здесь стали рассуждающие модели (reasoning models), которые сначала «обдумывают» задачу «на черновике», а только потом начинают выдавать ответ. На основе современных LLM уже создаются системы, которые способны производить новые научные результаты, — и это только начало пути.

Пожалуй, самое поразительное здесь не конкретные достижения, а как раз скорость прогресса. Закон Мура для AI работает с удвоением производительности не каждые 2 года, а каждые несколько месяцев. Задачи, которые казались серьёзным вызовом год назад, сегодня решаются почти идеально. Количество вычислений, требующееся для обучения передовых моделей, удваивается примерно каждые 6 месяцев. Мы живём в эпоху языковых моделей, которые прямо сейчас меняют мир в самых разных областях, и экспоненциальный прогресс никак не хочет останавливаться...

Какой будет четвёртая революция?

Заглянуть в будущее всегда сложно, но кое-что мы уже видим. Четвёртая революция в AI, похоже, будет обеспечена не одним прорывом, а несколькими параллельными направлениями, которые могут сойтись неожиданным образом.

Новые архитектуры. Несмотря на доминирование трансформеров, у них есть фундаментальные ограничения. Главное — квадратичная сложность механизма внимания: каждый токен должен «посмотреть»

на все остальные токены, что означает, что вычисления растут пропорционально квадрату длины текста. Для контекста в миллионы токенов это становится вычислительно невозможным. Кроме того, у трансформеров нет настоящей памяти — они всегда обрабатывают весь контекст заново.

В последние годы появляются альтернативы: SSM (State Space Models) вроде Mamba с линейной сложностью и встроенной памятью, архитектуры с разреженным вниманием, семейство JEPA (Joint Embedding Predictive Architecture) от Яна Лекуна и так далее. Пока неясно, какая из этих идей «выстрелит», но поиск архитектуры следующего поколения уже идёт полным ходом.

Мультимодальность и воплощённый AI. Сегодняшние модели всё ещё в основном работают с текстом и изображениями. Но человеческий интеллект развивался во взаимодействии с физическим миром — через прикосновения, движение, манипуляцию объектами. Есть гипотеза, что для создания по-настоящему общего интеллекта нужен воплощённый AI (embodied AI) — искусственный интеллект, который учится через непосредственный опыт некоего физического агента.

Уже появляются модели мира (world models), которые учатся предсказывать последствия действий в визуальной или тактильной среде. Роботы с AI-управлением начинают справляться со сложными задачами манипуляции. Многие компании сейчас работают над человекоподобными роботами, управляемыми большими мультимодальными моделями. Возможно, следующий прорыв придёт именно отсюда, когда AI научится не просто рассуждать о мире, но и действовать в нём.

Агентные системы. Современные LLM отвечают на запросы, но в основном пассивны. Агентные системы должны быть способны ставить себе цели, планировать, использовать инструменты, взаимодействовать с окружающей средой и другими агентами для достижения долгосрочных целей. Уже существуют прототипы, которые могут пользоваться компьютером и браузером, выполнять последовательности действий, реализовывать целые программистские проекты.

Но настоящие агенты потребуют решения проблем надёжности, безопасности и согласования (alignment) целей AI с человеческими ценностями. Агент, который может действовать автономно, потенциально гораздо опаснее, чем пассивный помощник.

AI для науки. И здесь мы подходим к самому головокружительному сценарию. В 2024 году появились системы вроде FunSearch от Google DeepMind, открывшей новые математические результаты, или AI Scientist от Sakana AI, способной проводить полный цикл научного исследования, от гипотез через эксперименты до готовой статьи. LLM уже помогают доказывать теоремы, предсказывать структуры белков, искать новые материалы.

Что будет, когда AI станет не просто помощником учёного, а самостоятельным исследователем? А что будет, когда AI-системы начнут проводить исследования в области самого искусственного интеллекта?

Многие слышали о технологической сингулярности, моменте, когда прогресс становится настолько быстрым, что люди уже не могут уследить за ним. До недавних пор эти рассуждения были чистой фантастикой. Но сейчас кажется, что если AI сможет лучше людей проводить исследования в области AI, то такая система сможет улучшать сама себя, создавая следующее поколение ещё более мощных систем, и этот процесс сможет развиваться экспоненциально без участия людей — та самая сингулярность.

Что будет в таком случае, не знает никто. Есть и утопические варианты прогнозов (решение всех научных и технологических проблем человечества, достижение изобилия, даже потенциального бессмертия), и экзистенциальные риски: если мы создадим системы умнее нас самих и их цели вдруг окажутся несовместимы с человеческим выживанием, человечество может и не сохранить контроль за будущим. Но важно, что все эти прогнозы и варианты очень, очень близки; в сфере искусственного интеллекта пессимистами считаются те, кто откладывает свой прогноз достижения сверхчеловеческого интеллекта на середину 2030-х, а оптимисты предсказывают это уже в нашем десятилетии...

Мы живём в уникальное время — возможно, самое важное в истории человечества. За четверть века искусственный интеллект прошёл путь от набора разрозненных методов с достаточно узкой сферой применимости до технологий, которые трансформируют все аспекты нашей жизни. Три революции — глубокого обучения, трансформеров и языковых моделей — изменили не только наши технологии, но и наше представление о возможном.

Следующие несколько лет будут определяющими. Четвёртая революция уже началась, но мы ещё не знаем её имени. Выборы, которые мы, как исследователи, разработчики, регуляторы, пользователи, сделаем сейчас, могут определить траекторию развития не только AI как науки, но и человечества в целом. У нас есть уникальная возможность сознательно направить развитие самой мощной технологии в истории. Будем ли мы достаточно мудры, чтобы воспользоваться ею?