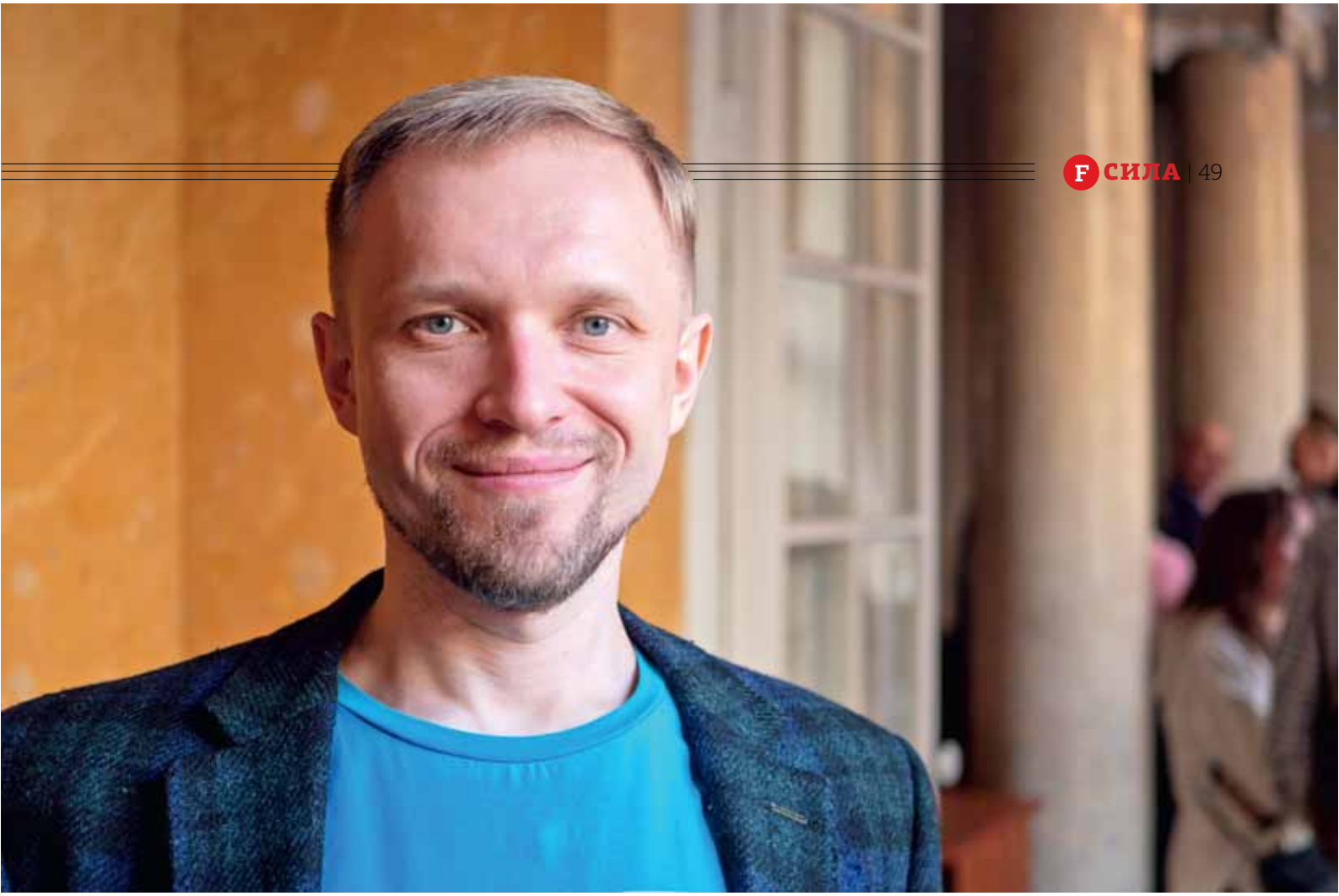


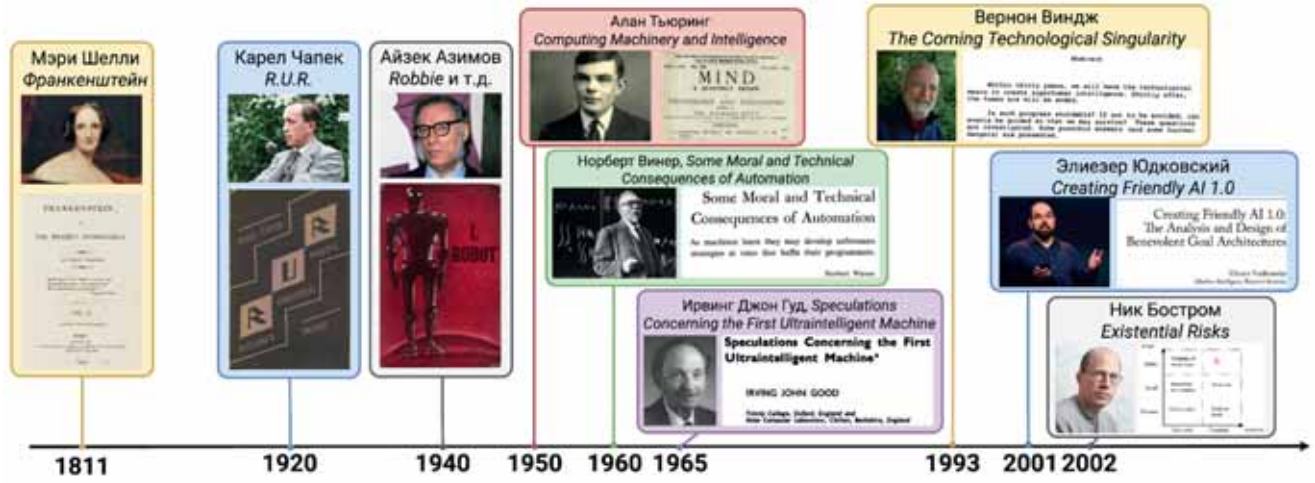
# ТОЧКА НЕВОЗВРАТА?

Создание сверхмощного ИИ ведёт общество к очередной промышленной революции и заодно ставит под вопрос место человека на планете. Чем ещё грозит появление человекоподобного ИИ, рассказал «ИП» доктор физико-математических наук, заведующий лабораторией искусственного интеллекта Санкт-Петербургского отделения Математического института им. В.А. Стеклова (ПОМИ РАН) Сергей Николенко.

**Обеспокоенные исследователи ИИ (в частности, Элиезер Юдковский) призывают бизнес бессрочно отключить нейросети для спасения человечества. На ваш взгляд, с точки зрения как защиты данных, так и общего развития технологий есть ли основания для беспокойства?**

— Создание искусственного интеллекта человеческого, а значит, и сверхчеловеческого уровня — это важная развилка для людей на Земле. Это может быть как огромный шаг вперёд для нашего процветания (об этом рекомендую прочесть текст основателя Anthropic Дарио Амодеи «Machines of Loving Grace»), так и последнее, что человечество сделает в своей истории (а вот об этом как раз очень убедительно пишет тот самый Элиезер Юдковский). Возможный сценарий такой развилки недавно представили Даниэль Кокотайло и другие известные прогнозисты в работе «AI 2027» — это, конечно, всего лишь прогноз, но хорошо обоснованный.

Основная проблема безопасности искусственного интеллекта — что по-настоящему умная сущность сможет найти неожиданные для нас способы достижения своих целей. Ещё Норберт Винер приводил пример, который сейчас формулируют как «максимизация канцелярских скрепок»; если сверхчеловеческий интеллект задаст-




↑ ***Ранняя история AI safety (науки о безопасности AI)*** ИЛЛЮСТРАЦИЯ СЕРГЕЯ НИКОЛЕНКО ИЗ БЛОГА НА SYNTHESIS.AI

ся целью максимизировать производство канцелярских скрепок, вскорости никаких людей на Земле не останется, а поверхность планеты будет покрыта двумя типами фабрик: производящих скрепки и производящих космические корабли для того, чтобы производить скрепки где-то ещё. Мы не умеем формулировать задачи о реальном мире так, чтобы в них не оставалось лазеек, при использовании которых результат нам не понравится. Этот эффект известен как закон Гудхарта (Goodhart's Law): когда разумные сущности оптимизируют некоторую метрику (KPI) — она быстро теряет свой исходный смысл.

По мере того как AI-системы становятся всё более мощными, такие эффекты начинают проявляться всё сильнее. Приведу один яркий пример. Большие языковые модели (large language models, LLM) нужно дообучать для того, чтобы система для предсказания токенов стала полезной для человека (ниже расскажу об этом подробнее). Это делается на основе человеческих предпочтений: у людей спрашивают, какой вариант ответа они предпочитают, считают более правильным, и модель дообучается давать именно такие ответы. «Wen et al». (2024) провели такой эксперимент: дообучали LLM как обычно, но на более сложных вопросах, где человеку трудно было за ограниченное время однозначно проверить результат. В результате люди действительно гораздо чаще выбирали ответы LLM как правильные… вот только более правильными они отнюдь не стали. Для LLM оказалось проще обучиться не чаще давать корректные ответы, а лучше убеждать людей, что её (по-прежнему неправильные) ответы верны! Это типичный пример закона Гудхарта, и, согласитесь, даже этот эксперимент уже вызывает серьёзное беспокойство — а работа «Wen et al». вышла в сентябре прошлого года, вечность назад по меркам развития AI. Я недавно написал большой пост с обзором таких примеров и закона Гудхарта в целом (ссылку можно найти в моём telegram-канале «Sinekура»).

На фоне этой проблемы все остальные, честно говоря, кажутся мне куда менее важными. Да, конечно, искусственный интеллект окажет огромное влияние на экономику — это ясно уже сейчас. Но я не боюсь новой промышленной революции, мне кажется, что она в результате может только улучшить жизнь людей. Это большая отдельная тема, но в целом эффект от AI, подвластного человеку, кажется мне крайне положительным. Есть, правда, ещё один важный повод для беспокойства: могут найтись люди, которые смогут использовать AI во вред ещё до того, как сами AI-модели получат возможность нанести вред человечеству.

Как правильно сформулировать то, что мы хотим от искусственного интеллекта? Как сделать так, чтобы будущие AI-агенты не пытались уничтожить человечество (что было бы для них очень логично, ведь люди — главная потенциальная угроза их существованию) и чтобы не помогали эффективно уничтожать нас другим людям, если у тех вдруг возникнет такое желание (к сожалению, с людьми такое бывает)? Как вообще оставить себе хотя бы возможность выключить AI-агента? Всё это совершенно неочевидно, и это предмет науки о безопасности искусственного интеллекта (AI safety).

Чтобы лучше понять её задачи и текущее состояние, без ложной скромности посоветую послушать мой недавний открытый доклад об AI safety (также в telegram-канале «Sinekура»). К сожалению, пока и текущее состояние AI safety, и скорость прогресса в этой области существенно отстают от прогресса в развитии новых способностей искусственного интеллекта.

**Далеко не все понимают, как именно устроен ИИ изнутри (чаще всего представляя его в виде ресёчера по интернету). Но если говорить научно — что происходит внутри больших ИИ-моделей? В чём основная уязвимость и несовершенство существующих систем и алгоритмов?**

— Искусственные нейронные сети основаны на принципе коннекционизма: если соединить много простых компонентов в большую и сложную структуру, этой композицией можно будет реализовать очень сложные функции, даже если исходные компоненты были маленькими и простыми. Нейронная сеть — это по сути граф из простых функций, большая их композиция, которая вычисляет что-то сложное, как правило — функцию ошибки для решения той или иной задачи (ниже обсудим пример с языковыми моделями). Задача обуче-

ния — оптимизировать эту функцию, например минимизировать функцию ошибки, подбирая соответствующие значения весов, то есть некоторых аргументов этой большой функции. Звучит очень абстрактно, но здесь и правда конкретнее не скажешь: нейросеть может выглядеть как угодно, и большая часть прогресса в нейросетях — это изобретение новых архитектур, то есть новых способов организовать эту большую композицию, задать структуру графа вычислений.

Нейросети были в AI всегда, ещё до оформления AI как науки: математическая модель одного нейрона от Маккаллоха и Питтса (1943) в целом до сих пор не изменилась, только функции активации новые изобретали, а о глубоких нейросетях писал ещё Алан Тьюринг. Прогресс в нейросетях — это главным образом изобретение новых архитектур сетей, более эффективных алгоритмов обучения (оптимизации) и новых способов поставить задачу (новых целевых функций). Так, например, в конце 2017 года появилась новая архитектурная идея, которая сейчас используется буквально везде: архитектура трансформера, точнее, слоя самовнимания в нём (self-attention). Поля этого интервью слишком узки, чтобы объяснить, как работает self-attention, но в целом это тоже определённый способ устроить вычисления внутри нейросети.

Что до несовершенств, то открытых вопросов в науке всегда много, и в AI, конечно, тоже. Сначала приведу пример про дальнейшее развитие. При всех своих замечательных свойствах у механизма self-attention есть важный недостаток: ему приходится каждый раз вычислять матрицу весов внимания, с которыми каждый элемент входа трансформера может «посмотреть» на каждый другой элемент входа. Это значит, что у self-attention квадратичная сложность, и именно поэтому на вход большой языковой модели трудно подать миллионы токенов. Чтобы с этим справиться, нужен или отдельный механизм памяти, или вообще другой подход к организации сети, другая архитектура; такие варианты разрабатываются, и вполне возможно, что в итоге следующий прорыв даже в тех же LLM будет уже работать не на трансформерах, а, например, на архитектурах вроде Mamba.



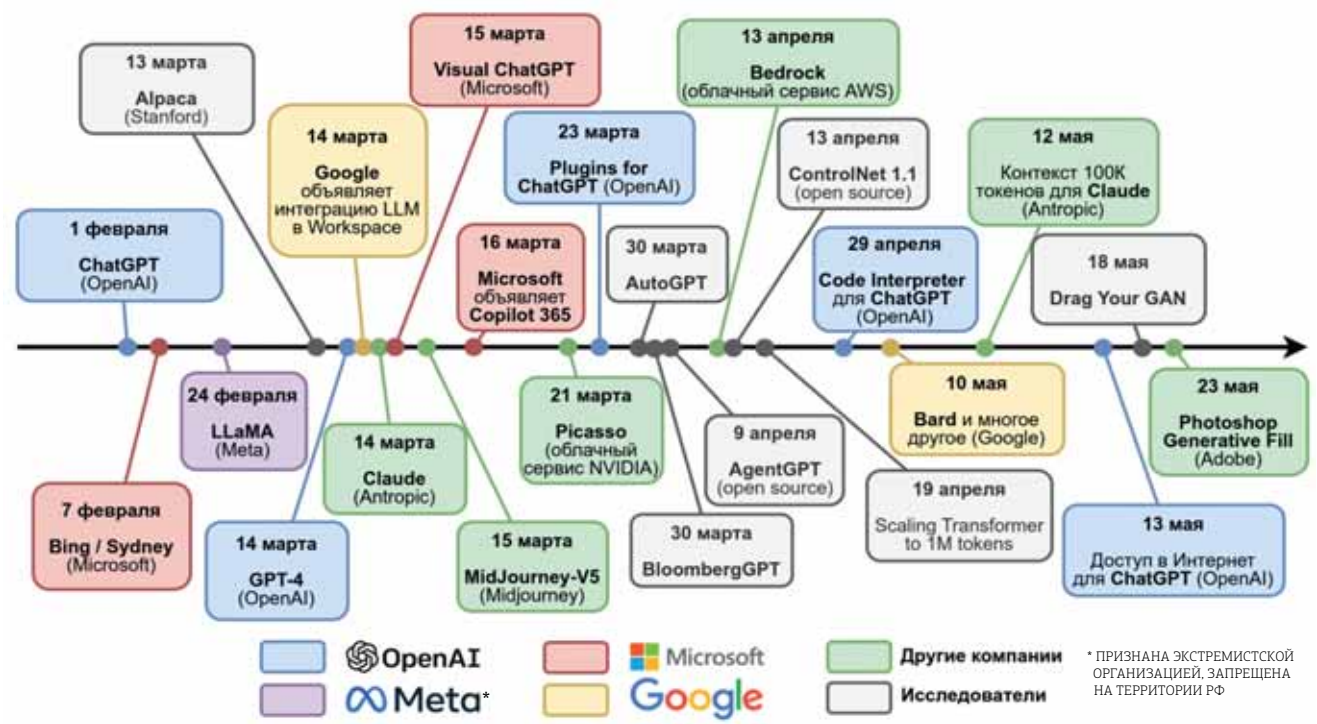

Ну а в «сдерживании», то есть в обеспечении безопасности AI, пока сплошные вопросы без ответов. Есть некоторый прогресс, например, в интерпретируемости: учёные из Anthropic в цикле недавних работ лучше научились понимать, «о чём модель сейчас думает», и сам Дарио Амодеи считает интерпретируемость ключевым направлением. Но на «проклятый вопрос» искусственного интеллекта — как мы собираемся управлять сущностями умнее нас самих — пока никаких хороших ответов нет. Очень надеюсь, что скоро появятся!

**Известно, что ИИ обучается на языковых структурах, чтобы воспроизводить смыслы. Как проходит обучение моделей, где находится граница между имитацией и рождением мысли и что «оживит» нейросеть? Может ли вообще ИИ приобрести когнитивные функции человеческого уровня, а не просто предсказывать? Возможно ли создать такой автономный сверхинтеллект только математическими алгоритмами или не обойдётся без нейробиологии? И главное — если учёные реализуют сильный ИИ, то что останется человеку и где хранится та красная кнопка, которая сможет деактивировать такой AGI?**

— Первую часть этого вопроса я, признаться, совсем не понял. По случаю посоветую читателю не пользоваться настолько туманным языком: что такое «смысл», как мы проверим, что мы его «воспроизвели», что такое «рождение мысли»... Для того чтобы понять, что происходит в искусственном (а может, и естественном!) интеллекте, отвечать на эти вопросы совершенно не обязательно.

Языковые модели обучаются предсказывать следующее слово (формально — «токен», там своя схема кодирования) в тексте. Такая постановка задачи оказалась крайне плодотворной именно потому, что она проста и не требует ручной разметки. Не нужно спрашивать людей, какой правильный ответ на вопрос, можно взять любой написанный людьми текст, и он тут же превратится в размеченный набор данных: вы обрезаете текст в любой момент, просите предсказывать следующее слово, et voila — правильный ответ у вас уже есть, он там же, в тексте дальше написан. Современные большие языковые модели — LLM обучаются таким образом на всём, что можно скачать (и на кое-чём из того, что нельзя), и это составляет самый продолжительный, вычислительно сложный и дорогой этап обучения новой LLM (предобучение, pretraining), хотя дальше нужны и другие этапы (дообучение, fine-tuning).

Аналогично плохо определённой кажется мне и вторая часть вопроса: почему вдруг «когнитивные функции» должны включать в себя что-то, кроме «просто предсказывать»? Пока весь опыт искусственного интеллекта показывает, что, обучаясь как можно лучше предсказывать токены, можно получать всё более и более сложные




* ПРИЗНАНА ЭКСТРЕМИСТСКОЙ ОРГАНИЗАЦИЕЙ, ЗАПРЕЩЕНА НА ТЕРРИТОРИИ РФ

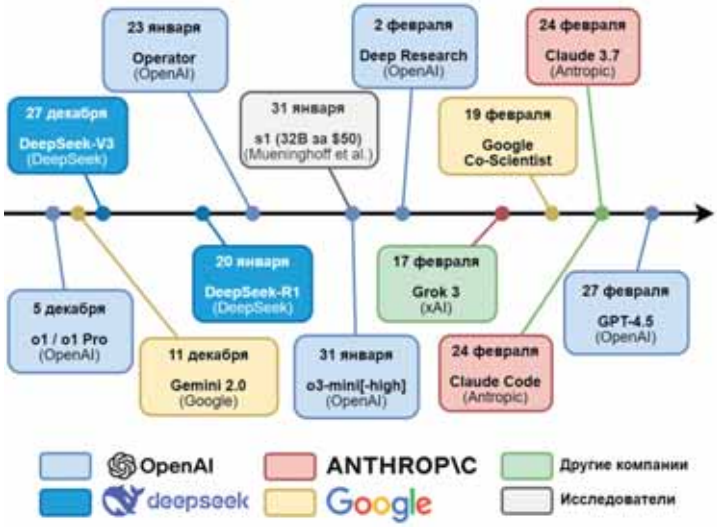
когнитивные функции, и конца у этого прогресса пока не видно, законы масштабирования выполняются всё дальше и дальше... В AI этот эффект известен как emergent abilities: по мере роста размера модели она начинает успешно решать всё больше разных задач, хотя её отдельно не обучали их решать, обучали только предсказывать токены.

Почему тест Тьюринга оказался настолько успешен как тест искусственного интеллекта (его формально прошли только совсем недавно, весной 2025 года, хотя, конечно, в последние год–два никто особенно не сомневался)? Потому что в свободном общении на естественном языке можно поговорить о чём угодно: и об окружающем мире, и о человеческом опыте, и о решении логических задачек или понимании тех самых смыслов... Так же и здесь: чтобы предсказывать следующее слово чуть более успешно, нужно уметь решать значительно более сложные задачи, понимать более глубокие закономерности в текстах, которые могут касаться чего угодно.

Более того, даже в упомянутой в вопросе нейробиологии есть теория предсказательного кодирования, которая утверждает, что наш мозг — это тоже большая мультимодальная языковая модель. По этой теории мозг обучает свою «ментальную модель мира» предсказывать следующий сигнал, который будет получен от органов чувств, и это и есть основная функция ошибки для обучения нашего с вами мозга, прямо как у LLM. Конечно, нейробиология — дело сложное, и это всего лишь одна из теорий, а не установленный факт, но она, насколько я понимаю, ничему пока не противоречит.

А вот последний вопрос в этом пункте определён, увы, очень хорошо, до сих пор очень актуален и однозначного ответа на него пока нет ни у кого. Вопрос о том, можно ли каким–то образом оставить за собой контроль над интеллектуально превосходящими нас агентами, хотя бы в виде той самой красной кнопки, — это один из центральных вопросов безопасности искусственного интеллекта. Там это свойство называют corrigibility («исправимость»): свойство AI–агента, которое обеспечит нам возможность его модифицировать. Без достижения этого свойства создавать сильный искусственный интеллект было бы очень опасно, и я надеюсь, что какой–то прогресс в этой области всё–таки будет достигнут в ближайшем будущем.

Дарья Ковалёнок




***← Вот уже 2 года скорость развития AI не замедляется. Сверху: таймлайн весны 2023 года. Снизу: таймлайн начала 2025 года***

ИЛЛЮСТРАЦИИ СЕРГЕЯ НИКОЛЕНКО ИЗ БЛОГА НА SYNTHESIS.AI