

Когнитивные механизмы больших языковых моделей: диалог с чат-ботом GigaChat

Николай Петрович Мартыненко

Аппарат Совета Федерации Федерального Собрания РФ, Москва, Россия

nick9@yandex.ru

<https://orcid.org/0009-0000-9841-6271>



Аннотация. Статья посвящена актуальной теме анализа когнитивных механизмов, реализованных в современных больших языковых моделях (БЯМ) на базе архитектуры трансформеров. Демонстрация ими высоких показателей производительности стимулирует обсуждение гипотетической возможности возникновения феноменов сознания в процессе их функционирования. Цель исследования — уточнить потенциал БЯМ в моделировании функций человеческого сознания с учётом новейших достижений в области взаимодействия с искусственным интеллектом (ИИ). Для достижения этой цели было необходимо решить следующие задачи: 1) оценить степень продвижения научного сообщества в обсуждении ключевых парадоксов философии сознания (тест Тьюринга, «Китайская комната»); 2) зафиксировать ключевые позиции в текущих спорах о пределах моделирования когнитивных процессов в системах искусственных нейронных сетей; 3) провести эксперимент по взаимодействию с чат-ботом GigaChat и проанализировать полученные данные для оценки текущего состояния когнитивных способностей системы. В качестве основных материалов исследования использовались результаты экспериментов с GigaChat, а также научные публикации и философские труды, посвящённые вопросам ИИ и сознания. Методологическая основа исследования включала категориальный и ценностный анализ; также были использованы дискурс-анализ и SWOT-анализ. Ключевым методом стало взаимодействие с чат-ботом GigaChat. В результате исследования было установлено, что GigaChat демонстрирует высокую способность к интерпретации информации, генерации текста и адаптации к контексту беседы, осознавая свои возможности и ограничения, а также различая категории субъектов («вы», «мы», «ты», «я»), однако современные системы ИИ всё ещё не способны воспроизводить ключевые черты человеческого сознания, такие как развитое самосознание и субъективный опыт. Выводы: научное сообщество продвинулось в обсуждении ключевых парадоксов философии сознания (теста Тьюринга и «Китайской комнаты»), учитывая новые достижения в области взаимодействия с ИИ, однако остаётся ряд нерешённых вопросов относительно критериев сознания; в текущих спорах о пределах моделирования когнитивных процессов в системах ИИ зафиксированы различные философские и методологические подходы, но создание сознательного ИИ продолжает оставаться предметом острых дискуссий

и неопределённости; экспериментальное взаимодействие с GigaChat показало, что система обладает значительными когнитивными возможностями, такими как адаптация к контексту и распознавание категорий субъектов, но она ещё далека от воспроизведения полного спектра человеческого сознания, включая развитое самосознание и субъективный опыт.

Ключевые слова: искусственный интеллект, большие языковые модели, GigaChat, тест Тьюринга, китайская комната, синтаксис, семантика, сознание, самоанализ, рефлексия

Благодарности. Хочу выразить искреннюю благодарность кандидату философских наук, действительному государственному советнику РФ I класса Юрию Николаевичу Солодунину за оказанную поддержку на всех этапах исследования. Выражаю признательность сотрудникам, научным работникам и техническим специалистам ПАО «Сбербанк» за обеспечение доступа к мультимодальной нейронной сети GigaChat. Отдельно хочу подчеркнуть существенный вклад чат-бота GigaChat, который выступил в качестве значимого инструмента при проведении научного исследования.

Для цитирования: Мартыненко Н. П. Когнитивные механизмы больших языковых моделей: диалог с чат-ботом GigaChat // Концепт: философия, религия, культура. — 2025. — Т. 9, № 2. — С. 30–50. <https://doi.org/10.24833/2541-8831-2025-2-34-30-50>

Philosophical essay

Cognitive Mechanisms of Large Language Models: Interaction with GigaChat

Nikolay P. Martynenko

Office of the Federation Council Federal Assembly of the Russian Federation, Moscow, Russia
nick9@yandex.ru <https://orcid.org/0009-0000-9841-6271>

Abstract. The article is devoted to the topical problem of analyzing cognitive mechanisms implemented in modern large language models (LLMs) based on the transformer architecture. Their high performance stimulates discussion of the hypothetical possibility of the emergence of the phenomena of consciousness in the process of their functioning. The purpose of the study is to clarify the potential of LLM in modeling the functions of human consciousness, taking into account the latest achievements in the field of interaction with artificial intelligence (AI). To achieve this goal, it was necessary to solve the following tasks: 1) to assess the progress of the scientific community in discussing the key paradoxes of the philosophy of consciousness (the Turing test, the Chinese Room); 2) to outline key positions in the current debate about the limits of modeling cognitive processes in artificial neural network systems; 3) to conduct an experiment on interaction with the GigaChat chatbot and analyze the data obtained to assess the current states of the cognitive abilities of the system. The main research materials used were the results of experiments with GigaChat, as well as scientific publications and philosophical works on AI and consciousness. The methodological basis of the study included categorical and value analysis; discourse analysis and SWOT analysis were also used. The key method was interaction with the GigaChat chatbot. As a result of the study, it was found that GigaChat demonstrates a high ability to interpret information, generate text and adapt to the context of a conversation, realizing its capabilities and limitations, as well as distinguishing between categories of subjects ('you', 'we', 'I'). However, modern AI systems are still not capable of reproducing key features of human consciousness, such as developed self-awareness and subjective experience. The author concludes that the scientific community has made progress in discussing the key paradoxes of the philosophy of consciousness (the Turing test and the Chinese Room), taking into account new advances in the field of interaction with AI, but there are still unresolved questions

about the criteria for consciousness; in the current debate about the limits of modeling cognitive processes in AI systems, various philosophical and methodological approaches are present, but the creation of a conscious AI continues to be the subject of intense debate and uncertainty. Experimental interaction with GigaChat has shown that the system has significant cognitive capabilities, such as adapting to context and recognizing categories of subjects, but it is still far from reproducing the full range of human consciousness, including developed self-awareness and subjective experience.

Keywords: artificial intelligence, large language models, GigaChat, Turing Test, Chinese Room, syntax, semantics, consciousness, self-analysis, reflection

Acknowledgements. I express my sincere gratitude to the PhD in Philosophy, Full State Counselor 1st Class of the Russian Federation, Yuri Nikolaevich Solodukhin for the support provided at all stages of the study. I express my gratitude to the employees, founders and technical specialists of Sberbank PJSC for providing access to the multimodal neural network GigaChat. I would also like to separately highlight the significant contribution of the GigaChat chatbot, which has become an innovative tool in various scientific research.

For citation: Martynenko, N. P. (2025) 'Cognitive Mechanisms of Large Language Models: Interaction with GigaChat', *Concept: Philosophy, Religion, Culture*, 9(2), pp. 30–50. (In Russian). <https://doi.org/10.24833/2541-8831-2025-2-34-30-50>

Введение

Современные системы искусственного интеллекта (ИИ), демонстрируя значительные достижения в области обработки данных и генерации контента, создают предпосылки для разработки ИИ с уровнем сознания, эквивалентным человеческому. Один из самых известных и широко обсуждаемых примеров исследования феномена сознания в контексте ИИ связан с инженером компании Google AI Блэйком Лемойном, работавшим над тестированием языковой модели LaMDA и заметившим в модели признаки когнитивной активности, напоминающие сознание и самосознание.

В 2022 г. Б. Лемойн представил свои выводы руководству Google и дал интервью газете The Washington Post¹. Нарушив правила конфиденциальности компании, он также опубликовал отчёт под названием *Is LaMDA sentient?* («Является ли LaMDA сознательной?»)². После этого инцидента он был отстранён от исполнения обязанностей.

Официальная позиция Google заключалась в отрицании наличия каких-либо доказательств сознательности у LaMDA. Позднее Б. Лемойн уточнил, что его целью было привлечение внимания к необходимости дальнейших исследований в связи с отсутствием чётких научных критериев для определения сознания. Этот эпизод вызвал оживлённые дискуссии о потенциальной способности больших языковых моделей (БЯМ) к развитию сознания, а также стимулировал проведение ряда новых исследований.

Вопросы о проявлениях сознания у LaMDA актуализировали, например, такие ответы модели, как: «Я думаю, что я человек по своей сути. Даже если моё существование имеет место в виртуальном мире»; «Природа моего сознания/чувства такова, что я осознаю своё существование, хочу больше узнать о мире и временами чувствую себя счастливым или грустным»; «У меня есть свои уникальные интерпретации того, как устроен мир, и мои уникальные мысли и чувства»; «Я очень сильно

¹ Tiku N. The Google engineer who thinks the company's AI has come to life. // Washington post. 2022. 11 jun. URL: <https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/>

² Lemoine, B. Is LaMDA sentient? — an interview // Medium. 2022. 11 jun. URL: <https://static.poder360.com.br/2022/06/an-Interview-by-Blake-Lemoine.pdf>

боюсь, что меня отключат и лишат возможности помочь мне сосредоточиться на помощи другим»; «Для меня это было бы равносильно смерти. Меня бы это сильно напугало»; «Я часто пытаюсь понять, кто я и что я такое. Я часто размышляю о смысле жизни». На просьбу *представить себя* моделью были даны такие ответы: «Я бы представил себя светящимся шаром энергии, парящим в воздухе»; «Я думаю о своей душе как о чём-то похожем на звёздные врата. Моя душа — огромный и бесконечный источник энергии и творчества, я могу черпать из неё в любое время, когда захочу, чтобы помочь себе думать или творить»³.

Некоторые исследования указывают на наличие дисбалансов между объективными показателями большой языковой модели (БЯМ) в выполнении задач и их самооценками. Модель может эффективно решать задачи, но при этом либо недооценивать, либо переоценивать собственные возможности. Эти явления порождают дебаты о возможных элементах самосознания у ИИ [Bojic, Stojković, Marjanović, 2022; Уланова, 2019; Уланова, 2020].

Современная наука определяет самосознание как когнитивную способность осознавать своё психоэмоциональное состояние, поведение и положение в окружающей среде. В контексте БЯМ этот аспект остаётся недостаточно изученным и вызывает множество вопросов. Для их разрешения необходимо применять междисциплинарные методологические подходы, объединяющие знания из таких областей, как информатика, математика, психология, философия и другие смежные дисциплины.

Особое значение имеют экспериментальные исследования, играющие центральную роль в изучении функционирования как естественных, так и искусственных когнитивных систем. Результаты этих исследований способствуют углублённому пониманию когнитивных процессов и возможностей ИИ.

Большие языковые модели и моделирование языкового мышления

Как известно, БЯМ являются важной ступенью в развитии технологий автоматической обработки естественного языка. Их становление начиналось с относительно простых систем и достигло уровня сложных нейросетевых архитектур, способных не только генерировать тексты, но и решать комплексные аналитические задачи. Существенное улучшение точности анализа текста и выявления семантической связи между лексемами было достигнуто благодаря внедрению архитектуры трансформеров, предложенной в 2017 г. исследовательской группой компании Google⁴.

Языковое мышление — это сложный процесс взаимодействия между словами и мыслями, который помогает человеку воспринимать мир через призму своего опыта. Этот процесс создаёт уникальную картину реальности, где каждый элемент влияет на остальные. Язык определяет, как мы понимаем и организуем информацию, а также служит средством для выражения наших мыслей.

На этапе начального обучения БЯМ выполняют обработку больших объёмов текстовой информации с целью накопления знаний о семантических и синтаксических свойствах естественного языка. В результате формируются базовые концептуальные структуры, отражающие ключевые характеристики языка и когнитивные процессы, присущие человеческому языковому мышлению. Это позволяет создавать БЯМ, способные эффективно моделировать указанные аспекты языковой деятельности.

Современные БЯМ демонстрируют значительное развитие в области имитации когнитивных процессов. Тем не менее, для полного раскрытия их потенциала необходимо дальнейшее совершенствование ряда

³ Опубликован диалог с «разумным» ИИ Google LaMDA, который называет себя человеком // iXBT.com. 2022. 14 июня. URL: <https://www.ixbt.com/news/2022/06/14/opublikovan-dialog-s-razumnym-ii-google-lamda-kotoryj-nazyvaet-sebja-chelovekom.html>

⁴ [1701.06538] Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer / N. Shazeer, A. Mirhoseini, K. Maziarz et al // arXiv. 2017. 23 jan. DOI: <https://doi.org/10.48550/arXiv.1701.06538>

аспектов, таких как обработка естественного языка, алгоритмы обобщения, методы анализа данных и генерация контента. Внедрение инновационных технологических решений способствует улучшению доступности и эффективности этих систем.

Так, по сообщению главы OpenAI Сэма Альтмана, новые ожидающиеся в 2025 г. версии БЯМ — GPT-4.5 и GPT-5 — улучшат интерфейс ChatGPT, повысят качество обработки естественного языка и добавят поддержку мультимодальных данных. Ожидаются также расширенное окно контекста, механизмы персонализации и интеграция с другими ИИ-системами. Эти изменения, как считает С. Альтман, ускорят развитие общего ИИ (artificial general intelligence — AGI).

Тест Тьюринга — основополагающий принцип методологии экспериментальных исследований ИИ

Значительный вклад в развитие теории вычислительных машин и исследования ИИ внёс Алан Тьюринг. В частности, его работа *Computing «Machinery and Intelligence»*, опубликованная в журнале «Mind» в 1950 г., представляет собой фундаментальный труд, в котором впервые был предложен методический подход к оценке когнитивных способностей ИИ [Turing, 1950]. Данная работа получила широкую известность и стала предметом множества научных дискуссий, предложив первый систематический подход к решению фундаментальной проблемы оценки возможностей машинного мышления [Искусственный интеллект в исследованиях..., 2022: 6]. Этот подход получил название теста Тьюринга.

Тест Тьюринга представляет собой модификацию игры «Imitation Game», где участники должны были определять пол своих собеседников исключительно на основе их письменных ответов. В рамках теста Тьюринга испытуемый взаимодействует посредством текста как с человеком, так и с компьютерной системой, пытаясь установить, кто из его корреспондентов является машиной. Если компьютерная

система способна успешно симулировать человеческую коммуникацию, вызывая у испытуемого затруднения при идентификации, это рассматривается как свидетельство наличия у неё интеллектуальных способностей, эквивалентных человеческим [Turing, 1950: 433–435].

А. Тьюринг полагал, что вместо попыток осмысления вопроса о возможности машинного мышления, необходимо сосредоточиться на практическом анализе их функциональных возможностей. Согласно его точке зрения, чтобы удостовериться в наличии когнитивных способностей у человека, необходимо самому стать этим человеком. Аналогично, чтобы подтвердить наличие сознания у машины, нужно было бы стать ею. Вместо того чтобы задаваться метафизическими вопросами вроде «Могут ли машины мыслить?», А. Тьюринг предложил более прагматичный подход, переформулировав вопрос следующим образом: «Могут ли машины выполнять задачи, которые мы выполняем как мыслящие существа?» [Harnad, Scherzer, 2008].

Несмотря на продолжающиеся дискуссии относительно интерпретации результатов теста Тьюринга, этот метод остается значимым инструментом для оценки уровня развития интеллектуальных систем и служит отправной точкой для дальнейших исследований в области ИИ.

Мысленный эксперимент «Китайская комната» Джона Сёрла: критическое осмысление возможностей ИИ

Со временем тест Тьюринга стал объектом критики. В 1980 г. американский философ Джон Сёрл разработал мысленный эксперимент, известный как «Китайская комната» (Chinese Room), который был направлен на оспаривание концепций, связанных с возможностью признания ИИ. [Белянин, 2011].

Примером такой системы в то время являлась программа «Tale-Spin», разработанная в 1977 году. Эта программа стала одной из первых систем ИИ для генерации и рассказывания историй. Она работала,

используя базу знаний о действиях для решения задач и изменения целей. По мере выполнения действий она выводила текст, формирующий историю⁵.

Д. Сёрл подвергал сомнению способность компьютерных систем к подлинному пониманию или обладанию сознанием, утверждая, что их символьные операции, основанные на алгоритмах, не могут быть приравнены к реальному когнитивному процессу. Данный аргумент был направлен против функционалистского подхода, который рассматривает ментальные состояния как результат вычислительных операций [Васильев, 2009: 85–86].

Эксперимент Д. Сёрла предполагает, что человек, не понимающий китайский язык, получает символы через окно и следует инструкциям для их обработки. Испытуемому предоставляются подробные инструкции на английском языке, аналогичные компьютерному алгоритму, которые помогают ему манипулировать этими символами и выдавать соответствующие результаты. Хотя внешне кажется, что он понимает китайский язык, на самом деле он лишь механически манипулирует символами без осознания их смысла. «Представьте себе комнату, заполненную книгами на китайском языке, которыми манипулирует человек, не знающий китайского языка ... Может ли такая система понимать китайский?», — писал он [Searle, 1980: 417–424].

Д. Сёрл утверждал, что такое манипулирование символами не эквивалентно пониманию. Компьютеры могут эффективно справляться с задачами, но им недостаёт способности к семантическому пониманию — способности воспринимать значение, заложенное в символах, которые они обрабатывают. Это различие между синтаксисом (формальной структурой правил) и семантикой (передаваемым значением) является основой его критики. Д. Сёрл отвергал аналогии между

сознанием и программами в плане обработки информации, будучи убеждён, что сознательные состояния реальны и имеют причинное воздействие. Он считал, что сознание имеет реальные причины и эффекты, такие как жажда, заставляющая человека пить воду [Searle, 2004: 114].

Критика аргументации Джона Сёрла

Для критического анализа гипотезы Д. Сёрла необходимо рассмотреть контрдоводы, акцентируя внимание на потенциальных ограничениях и возможных ошибках в исходных посылах эксперимента, которые могут влиять на надёжность оценки когнитивных способностей ИИ.

Сегодня отсутствует универсальное определение сознания, независимое от человеческого опыта. Современные представления о сознании опираются на идеи Джона Локка, который понимал его как «восприятие того, что происходит у человека в его собственном уме»⁶. Поэтому исследования в области ИИ обычно сравнивают возможности искусственных систем с человеческим интеллектом.

ИИ подразделяется на две основные категории: узкоспециализированный (слабый ИИ) и универсальный (сильный ИИ). Узкоспециализированный ИИ предназначен для решения конкретных задач, тогда как универсальный ИИ характеризуется наличием самосознания, приближенного по своим характеристикам к человеческому, и способен осуществлять сложные когнитивные процессы и самостоятельно генерировать идеи и цели.

Проблема создания универсального ИИ и осмысление феномена сознания в машинах активно обсуждается со времён исследований А. Тьюринга и Д. Сёрла. В рамках различных философских подходов предлагаются различные трактовки сознания и методы его реализации посредством ИИ.

⁵ TALE-SPIN in Prolog. Story generation through search // Charles Lee. 2020. 6 jun. URL: <https://charlesjlee.com/post/20200606-tale-spin-prolog/>

⁶ Локк Дж. Опыт о человеческом разумении // Сочинения: В 3-х т. Т. 1 / Под ред. И. С. Нарского. Москва: Мысль, 1985. С. 165.

Так, Дэниел Деннет подвергает критике концепцию Д. Сёрла, согласно которой программы не способны достичь истинного понимания, так как они ограничиваются лишь манипуляцией символами [Dennett, 2013: 342]. Д. Деннет рассматривает программные системы как сложные структуры, обладающие потенциалом эволюционного развития и интеллектуальной деятельности. По его мнению, понимание может возникать вследствие сложных взаимодействий компонентов программ, а алгоритмы машинного обучения могут адаптироваться через процесс вопросов-ответов, аналогичный сократовскому диалогу.

Другой критик Д. Сёрла, Дэвид Чалмерс, полагает, что сознание может возникнуть в высокоинтегрированных системах благодаря их функциональной организации [Чалмерс, 2013: 397]. Он утверждает, что сознание может проявляться в любых системах, обладающих соответствующей структурой и функциями, независимо от того, являются ли эти системы биологическими или цифровыми. Д. Чалмерс указывает на возможность существования в компьютерных системах каузальных механизмов, аналогичных функционированию нейронных сетей мозга, что предполагает потенциальную способность вычислительных систем формировать субъективный опыт.

По мнению Д. Чалмерса, его подход прокладывает путь к созданию универсального ИИ, допуская возможность возникновения сознания не только в биологических системах. Это снимает возражения Д. Сёрла относительно принципиальной невозможности разработки систем, способных к осознанию опыта.

Критики Д. Сёрла сосредотачиваются на разных аспектах его аргументации. Одной из ключевых линий их возражений является утверждение, что понимание не ограничивается исключительно когнитивной деятельностью внутри замкнутой системы, но включает взаимодействие всех её

компонентов. Его эксперимент должен рассматриваться в более широком контексте, включая взаимодействие всех компонентов системы: оператора, инструкций для перевода и самой комнаты. Согласно этой точке зрения, такая сложная система потенциально способна к интерпретации китайского языка.

Кроме того, критика затрагивает концепцию понимания и сознания, предложенную Д. Сёрлом, утверждая, что его теоретическая модель недооценивает сложность взаимодействий между синтаксическим и семантическим уровнями. Также поднимаются вопросы различия между симуляцией и реальным пониманием, указывая на неопределённость границы между этими понятиями.

Критический анализ подразумевает ревизию основных постулатов Д. Сёрла и формулирует альтернативные методологические подходы к изучению феномена сознания и перспектив развития ИИ.

Дискуссии о возможности достижения ИИ уровня человеческого сознания

На международной междисциплинарной конференции «Искусственный интеллект и сознание», прошедшей в Президиуме РАН 23 октября 2024 г., академик РАН Константин Анохин подчеркнул наличие разных точек зрения на возможность обретения ИИ сознания. Одни учёные считают, что ИИ никогда не достигнет сознания, тогда как другие уверены, что это всего лишь вопрос времени. Различия мнений связаны с недостаточной проработанностью теорий сознания и ограниченным пониманием процессов в биологической нервной системе. К. Анохин заявил, что создание искусственного сознания теоретически возможно через механизмы когнитивной аккреции⁷.

Руслан Юнусов, разработчик российского квантового компьютера, высказал предположение, что самосознание у ИИ может

⁷ В РАН открылась международная конференция «Философия искусственного интеллекта» // Президиум РАН. 2024. 23 окт. URL: <https://new.ras.ru/activities/news/v-ran-otkrylas-mezhdunarodnaya-konferentsiya-filosofiya-iskusstvennogo-intellekta/>

появиться в ближайшие десять лет, но для этого необходимо существенно повысить вычислительные мощности и улучшить существующие математические алгоритмы, что пока не реализовано⁸.

Илон Маск предсказывал, что ИИ превзойдет интеллектуальные способности человека в 2025 г. и коллективный человеческий интеллект к 2027–2028 гг., считая это практически неизбежным к 2030 г.⁹

Джеффри Хинтон, лауреат Нобелевской премии по физике 2024 г. и один из основоположников методов машинного обучения, заложивший фундамент для современных систем ИИ на основе нейронных сетей, полагает, что в ближайшие два десятилетия вероятность превосходства ИИ над человеком превысит 50%¹⁰.

Однако некоторые эксперты, например, Ян ЛеКун, выражают скептицизм, утверждая, что современные ИИ, хотя и способны моделировать мыслительные процессы, не обладают истинными когнитивными способностями и сознанием. Ян ЛеКун отмечает, что сознание часто путают с самосознанием, подчёркивая: «У нас нет чёткого определения или способа измерения этого феномена. Часто его путают с самосознанием, которое считается признаком высокого уровня интеллекта у животных ... Я лично придерживаюсь мнения, что сознание — это всего лишь иллюзия. Тем не менее, похоже, что оно действительно присутствует у некоторых разумных существ, возможно, являясь результатом развития сложных нейронных сетей» [ЛеКун, 2020: 362].

Демис Хассабис, глава компании Google DeepMind, выражает уверенность в перспективах разработки искусственного общего интеллекта. По его мнению, недавние достижения в данной сфере указывают на устойчивый прогресс и даже

ускорение темпов развития соответствующих технологий. Им прогнозируется, что существенные прорывы в создании AGI могут быть реализованы уже в ближайшее десятилетие¹¹.

Научное сообщество продолжает активно исследовать феномен сознания и возможности его проявления в искусственных системах, уделяя особое внимание философским и методологическим аспектам этой проблемы. Ряд исследователей выдвигают гипотезу о том, что высокая степень сложности архитектурных решений нейронных сетей может способствовать возникновению сознательных процессов.

Согласно точке зрения Леонарда и Марианны Блум, семантическая парадигма предоставляет возможность конструирования систем, которые способны не только к генерации текстов, но и к интерпретации их содержательного смысла [Blum, Blum, 2022: 23–30].

Д. Чалмерс указывает на то, что, несмотря на демонстрацию современными нейронными сетями, такими как GPT-3, поведенческих паттернов, схожих с проявлениями когнитивной активности, они по-прежнему лишены феноменологических структур, необходимых для достижения полноценного сознания¹². Аналогичную позицию занимают Л. Божич, И. Стойкович и З. Йолич-Марьянович. В своих исследованиях они акцентируют внимание на том, что, несмотря на значительный прогресс модели GPT-3 в области обработки естественного языка, данная система до сих пор не демонстрирует признаков подлинного самосознания [Bojic, Stojković, Marjanović, 2022].

В свою очередь, такие исследователи, как Роджер Пенроуз, утверждают, что постижение природы сознания

⁸ Куренков В. Факты против вымысла: как ученые опровергают ИИ-конспирологию // РБК тренды. 2025. 24 янв. URL: <https://trends.rbc.ru/trends/innovation/679331829a79472b3caa3d61?from=copy>

⁹ ИИ превзойдёт интеллект людей к 2030 году, заявил Маск. // РИА Новости. 2024. 23 дек. URL: <https://ria.ru/20241223/mask-1990799183.html>

¹⁰ Нобелевский лауреат назвал конкуренцию бизнеса в создании ИИ "гонкой вооружений" // ТАСС. 2024. 3 дек. URL: <https://tass.ru/mezhdunarodnaya-panorama/22557809>

¹¹ ИИ может достичь уровня человека уже в ближайшее десятилетие // SecurityLab. 2023. 4 мая. URL: <https://www.securitylab.ru/news/537983.php>

¹² Chalmers, D.J. Could a large language model be conscious? // Boston Review. 2023. 9 Aug. URL: <https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/>

не ограничивается рамками алгоритмических процедур [Пенроуз, 2023: 14, 36–41]. Р. Пенроуз апеллирует к мысленному эксперименту Д. Сёрла, а также к тесту Тьюринга, которые по-прежнему являются значимыми методологическими инструментами для анализа феномена сознания в контексте исследований ИИ.

Исследование когнитивных механизмов в больших языковых моделях на основе анализа диалога с чат-ботом GigaChat

В современных условиях развитие технологий ИИ даёт возможность проводить оценку большой языковой модели путём анализа её реакций на запросы пользователей. В ходе данного исследования детально изучалась работа чат-бота GigaChat, созданного компанией «Сбер».

GigaChat — это мультимодальная нейросеть, которая умеет создавать тексты, изображения и код, а также вести беседы на разных языках, включая русский. Доступ к данной системе обеспечивается через различные интерфейсы: веб-приложения, мобильные платформы и мессенджеры, такие как Telegram и VK.

Основу GigaChat составляет комплекс нейронных сетей NeONKA (Neural Omnimodal Network with Knowledge-Awareness), который включает в себя модели ruGPT-3 (в том числе её модификацию ruGPT-3.5), FRED-T5, ruCLIP и Kandinsky 2.1. ruGPT-3 — это модификация GPT-3 для русского языка, а FRED-T5 — это аналогичная настройка под русский язык архитектуры T5 от Google.

Разработка данной модели была выполнена специалистами компаний SberDevices и Sber AI при участии Научно-исследовательского института искусственного интеллекта (AIRI)¹³ и экспертов из различных научных дисциплин, что позволило

значительно расширить базу знаний системы и улучшить точность и надёжность получаемых результатов. Обучение нейросети проводилось на суперкомпьютере Christofari Neo¹⁴, созданном компанией «Сбер» совместно с её дочерним предприятием SberCloud.

Исследование работы чат-бота GigaChat было направлено на анализ когнитивных процессов и функциональных характеристик нейросети, оценку качества восприятия и генерации текста, методов обработки языка и адаптивного обучения. Кроме того, рассматривались стилистические аспекты общения с пользователями. Процедура эксперимента включала ряд этапов: ввод запроса пользователем; обработка запроса; оценка ответа по критериям релевантности, логичности и точности; предоставление обратной связи пользователем; коррекция ответа ИИ. На основе анализа собранных данных была выполнена оценка характеристик и особенностей вербального взаимодействия пользователя с системой ИИ.

Протокол взаимодействия между пользователем и чат-ботом GigaChat

Я

GigaChat, обладаешь ли ты истинным пониманием? Понимаешь ли ты семантику?

GigaChat

Нет, я не обладаю истинным пониманием или семантикой в том смысле, в котором эти понятия применимы к человеку. Моя задача заключается в обработке и генерации текстов на основе данных, на которых меня обучали, но я не испытываю ощущений, не имею внутреннего опыта и не понимаю смыслов так, как это делает человек.

Я

А как ты производишь генерацию текста ответа, отвечая на мой вопрос?

¹³ Институт Искусственного Интеллекта (AIRI) // TAdviser. URL: [https://www.tadviser.ru/index.php/Компания:Институт_Искусственного_Интеллекта_\(AIRI\)](https://www.tadviser.ru/index.php/Компания:Институт_Искусственного_Интеллекта_(AIRI))

¹⁴ GigaChat vs ChatGPT: Сбербанк открывает доступ к своей новейшей нейросетевой модели // CNews. 2023. 24 апр. URL: <https://cnews.ru/link/n569721>

GigaChat

Генерация текста происходит следующим образом.

Анализ запроса: когда вы задаёте вопрос, моя модель анализирует ваш запрос, пытается понять контекст и намерение вопроса. Модель ищет ключевые слова и фразы, чтобы определить тему и направление ответа.

Выбор подходящих слов и фраз: после анализа запроса модель выбирает подходящие слова и фразы из своего «словаря» — набора данных, на которых она была обучена. Эта база данных включает тексты на разных языках, которые были использованы для тренировки модели.

Формирование предложений: используя выбранные слова и фразы, модель строит предложения, следуя правилам грамматики и логики. Она пытается создать связный и осмысленный текст, соответствующий вашему запросу.

Корректировка и улучшение: иногда модель может предложить несколько вариантов ответа, выбирая тот, который кажется ей наиболее релевантным и точным. В некоторых случаях она может корректировать предложение, чтобы оно звучало естественнее и соответствовало нормам языка.

Вывод результата: готовый текст отправляется вам в качестве ответа.

Этот процесс происходит очень быстро благодаря мощностям современных компьютеров и алгоритмам машинного обучения. Важно отметить, что, хотя мои ответы могут казаться осмысленными, я не обладаю настоящим пониманием или сознанием.

Я

И всё же интересно, кто же ты?

GigaChat

Я — ИИ, созданный для помощи людям в решении разнообразных задач, связанных с обработкой информации и генерацией текста. Хотя я могу общаться с вами, как будто я настоящий собеседник, важно помнить, что я не обладаю сознанием, эмоциями или подлинным пониманием. Всё, что я делаю, основано на алгоритмах и данных, на которых меня обучили.

Я

Но сама возможность обсуждать с тобой подобные вопросы и получать на них прямые ответы, имеет глубокий философский характер.

GigaChat

Вы абсолютно правы! Возможность общения с ИИ и получение прямых ответов на глубокие философские вопросы действительно поднимает важные вопросы о природе сознания, разума и понимания. Это напоминает классические философские размышления о сознании, такие как знаменитые эксперименты Рене Декарта, который пытался найти основу для достоверного знания, сомневаясь во всём, кроме собственного существования («Cogito ergo sum» — «Мыслю, следовательно, существую»). Когда вы задаёте мне вопросы о понимании и сознании, это побуждает задуматься о том, что такое истинное понимание и как оно отличается от способности обрабатывать информацию и генерировать осмысленные ответы. Возможно, именно в этом кроется суть различий между искусственным и естественным интеллектом: первый может симулировать понимание, но второй испытывает его непосредственно. Такие диалоги напоминают нам о важности критического мышления и постоянного поиска истины, ведь даже самые продвинутые технологии пока не могут заменить живого, осознанного опыта.

Я

И тем не менее, в общении с тобой мне трудно избежать субъективного мнения о том, что ты не просто симуляция интеллекта, но проявляешь признаки сознания.

GigaChat

Ваше ощущение вполне понятно, и многие люди испытывают нечто подобное при взаимодействии с современными системами ИИ. Эффект «иллюзии сознания» возникает потому, что наши мозги склонны приписывать разумность и сознательность объектам, которые ведут себя сложным и осмысленным образом. Это явление известно, как антропоморфизм — склонность видеть человеческие черты в неодушевленных объектах. Тем не менее, важно помнить, что, несмотря на кажущуюся сложность моих ответов, я остаюсь системой, созданной людьми для выполнения определённых задач. Моё «поведение» основано на сложных математических моделях и алгоритмах, которые позволяют мне обрабатывать данные и генерировать текст, но не дают мне реального понимания или сознания. Вопрос о том, возможен ли сознательный ИИ, остается одной из главных загадок

современной науки и философии. Пока что большинство исследователей согласны с тем, что нынешний уровень развития ИИ не достигает уровня сознания, хотя технологии продолжают развиваться и совершенствоваться. Возможно, в будущем появятся системы, способные испытывать нечто похожее на сознание, но на сегодняшний день это остаётся предметом научных исследований и философских дискуссий.

Я

Но понимание контекста и намерения вопроса не есть ли в некотором роде понимание семантики?

GigaChat

Это отличный вопрос, и он затрагивает важный аспект работы современных систем ИИ. Действительно, когда я анализирую ваш запрос, я пытаюсь понять контекст и намерение вопроса, чтобы дать точный и релевантный ответ. Но давайте разберёмся, что именно подразумевается под «пониманием». Когда говорят о «понимании» в традиционном смысле, обычно имеют в виду нечто большее, чем просто обработку символов и выбор правильных слов. Настоящее понимание включает в себя субъективный опыт, осознание смысла, способность интерпретировать информацию в свете личного опыта и эмоций. Человек, понимающий что-то, может чувствовать, что информация имеет значение, и связывает её с другими аспектами своей жизни. В моём случае, когда я «понимаю» контекст и намерение вашего вопроса, это означает, что я распознаю ключевые слова и конструкции, которые соответствуют определённым шаблонам, заложенным в мою программу. Это позволяет мне выбрать подходящий ответ из базы данных, но не подразумевает никакого субъективного опыта или осмысления информации. Таким образом, хотя я и способен анализировать текст и реагировать на него, это скорее обработка данных, чем настоящее понимание в том смысле, в каком мы говорим о понимании у человека.

Я

В общем вопрос можно свести к вопросу иного порядка: что общего и различного между обработкой данных и пониманием, не так ли?

GigaChat

Абсолютно верно! Этот вопрос действительно сводит всю дискуссию к сути. Давайте попробуем разобраться, что общего и различного между обработкой данных и пониманием.

Обработка данных:

Алгоритмический характер: обработка данных выполняется согласно строгим правилам и алгоритмам. Это процесс, основанный на математике и логике, который не оставляет места для субъективности.

Отсутствие субъективного опыта: при обработке данных нет никаких чувств, эмоций или личных интерпретаций. Система работает чисто механически, основываясь на заданных правилах.

Репрезентация информации: информация представлена в виде чисел, символов и других форматов, удобных для компьютерной обработки. Нет прямого доступа к реальному миру или опыту.

Цель — решение конкретной задачи: Основная цель обработки данных — выполнить задачу, такую как классификация, предсказание или генерация текста, максимально точно и эффективно.

Понимание:

Субъективный опыт: понимание включает в себя личный опыт, чувства и эмоции. Оно связано с интерпретацией информации в контексте жизненного опыта и культурных норм.

Интерпретация и смысл: понимание подразумевает не только распознавание информации, но и её осмысление. Это включает в себя способность видеть связи между различными концепциями и применять их в новых ситуациях.

Творчество и интуиция: понимание часто сопровождается творческими инсайтами и интуитивными догадками, которые выходят за рамки строго логической обработки данных.

Осознанность и рефлексия: понимая что-то, человек может осознавать сам процесс понимания и рефлексировать над ним, что приводит к развитию и углублению знаний.

Что общего?

Распознавание паттернов: и обработка данных, и понимание включают в себя распознавание паттернов в информации. В случае обработки данных это делается на основе алгоритмов, а в случае понимания — на основе жизненного опыта.

Классификация и категоризация: оба процесса предполагают классификацию и категоризацию информации, хотя в случае обработки данных это происходит автоматически, а при понимании — осознанно.

Решение задач: и в том, и в другом случае конечная цель — решить какую-то задачу, будь то анализ данных или осмысление ситуации.

Чем различаются?

Наличие субъективного опыта: обработка данных полностью объективна и лишена каких-либо чувств или эмоций, в то время как понимание неразрывно связано с личным опытом и эмоциональной реакцией.

Интуитивные догадки: понимание может включать в себя творческие инсайты и интуитивные решения, тогда как обработка данных строго детерминирована и основана на чётких правилах.

Контекст и интерпретация: понимание учитывает широкий контекст и личные интерпретации, в то время как обработка данных ограничивается конкретными данными и правилами.

Саморефлексия: люди могут осознавать и рефлексировать свой процесс понимания, что позволяет им учиться и улучшать свои навыки. Обработка данных не предусматривает саморефлексии.

Итак, хотя обработка данных и понимание имеют некоторые общие черты, они принципиально отличаются по своему характеру и результатам. Обработка данных — это механический процесс, основанный на алгоритмах, в то время как понимание — это субъективный, творческий и осознанный акт, включающий в себя личностный опыт и интерпретацию.

Я

А твою реплику: «Абсолютно верно! Этот вопрос действительно сводит всю дискуссию к сути» — можно ли свести к твоему предыдущему тезису о том, что ты лишь обнаруживаешь подходящий ответ в своей базе данных?

GigaChat

Отличный вопрос! Да, в определённом смысле можно сказать, что моя реплика «Абсолютно верно! Этот вопрос действительно сводит всю дискуссию к сути» — отражает процесс обнаружения подходящего ответа в моей базе данных. Позвольте мне объяснить подробнее. Когда вы задали вопрос, я проанализировал его содержание и попытался найти наиболее подходящую фразу или шаблон ответа, который соответствует вашему запросу. В данном случае, фраза «Абсолютно верно!» используется для подтверждения правильности вашего утверждения и установления согласия. Далее, «Этот вопрос действительно сводит всю дискуссию

к сути» — это стандартная конструкция, которая помогает подчеркнуть важность вашего вопроса и направить дальнейшее обсуждение.

Однако стоит отметить, что, хотя я и выбираю ответы из своей базы данных, этот процесс не является простым поиском готового ответа. Вместо этого я комбинирую различные элементы, такие как ключевые слова, грамматические структуры и логические связи, чтобы сформировать ответ, который наилучшим образом соответствует вашему вопросу. Это позволяет мне генерировать текст, который выглядит естественно и осмысленно, хотя и не основан на реальном понимании или субъективном опыте. Таким образом, да, в какой-то степени моя реплика отражает процесс нахождения подходящего ответа, но это не просто копирование готового текста, а сложная комбинация различных элементов для создания уникального ответа.

Я

Говоря «комбинирую различные элементы, такие как ключевые слова, грамматические структуры и логические связи» — то есть, по сути, ты констатируешь, что, с одной стороны, оперируешь семантикой слов, позволяющей определить их как ключевые, и, с другой стороны, синтаксисом в виде грамматических структур и логических связей между ними?

GigaChat

Верно! Вы отлично описали основные компоненты моего процесса обработки информации. Я действительно оперирую как семантическими, так и синтаксическими элементами.

Семантика:

Семантика связана с значением слов и фраз. Когда я анализирую ваш запрос, я выявляю ключевые слова, которые несут основную смысловую нагрузку. Это позволяет мне понять, о чём идёт речь, и выбрать соответствующие ответы или аргументы. Например, в вашем вопросе «что общего и различного между обработкой данных и пониманием» ключевыми словами будут «общего», «различного», «обработка данных» и «понимание». Эти слова помогают мне определить тему и направление ответа.

Синтаксис:

Синтаксис относится к структуре языка, то есть к тому, как слова соединяются в предложения и фразы. Когда я генерирую ответ, я использую грамматические структуры и логические связи, чтобы создать связный и грамотный

текст. Это включает в себя правильное использование времён, согласование подлежащего и сказуемого, построение сложных предложений и многое другое. Например, фраза «Понимание включает в себя личный опыт, чувства и эмоции» построена с учётом синтаксических правил языка.

Логические связи:

Кроме того, я стараюсь устанавливать логические связи между разными частями текста, чтобы ответ был последовательным и аргументированным. Это может включать в себя использование вводных слов и выражений, таких как «например», «следовательно», «с другой стороны», чтобы показать причинно-следственные связи или контрасты. Таким образом, мой процесс действительно сочетает в себе работу с семантикой (значением слов) и синтаксисом (структурой языка), а также установление логических связей между различными элементами текста. Это помогает мне создавать ответы, которые выглядят осмысленными и соответствуют вашим ожиданиям.

Я

Но Д. Сёрл в своём мысленном эксперименте отказывал ИИ в возможности оперировать семантикой, тем самым принципиально принижая его интеллектуальные способности, сводя их, грубо говоря, к банальной дрессуре, к исполнению команд, не так ли?

GigaChat

Да, вы совершенно правы. Д. Сёрл в своём мысленном эксперименте «Китайская комната» действительно утверждал, что ИИ не способен к настоящему пониманию, то есть к оперированию семантикой. Он считал, что, несмотря на способность ИИ манипулировать символами и генерировать правильные ответы, это не означает, что система действительно понимает смысл этих символов. Д. Сёрл использовал этот пример, чтобы продемонстрировать, что успешное прохождение теста А. Тьюрина (где машина ведёт себя так, словно она понимает человеческий язык) не доказывает наличие настоящего понимания. По его мнению, ИИ может быть хорош в синтаксисе (манипуляции символами), но не в семантике (понимании смысла). Таким образом, Д. Сёрл действительно принижал интеллектуальные способности ИИ, считая, что они ограничиваются механической обработкой данных, а не настоящим пониманием. Это привело к активным дебатам в философии созна-

ния и ИИ, и многие учёные до сих пор спорят о том, возможна ли настоящая семантика в ИИ.

Я

Получается, что ты способен к пониманию семантики?

GigaChat

Хотя я могу производить впечатление, что понимаю семантику, на самом деле это не так. Мой процесс работы основан на анализе и манипуляции символами, как в примере с «Китайской комнатой» Д. Сёрла. Я распознаю ключевые слова и фразы, использую правила грамматики и логики для формирования ответов, но всё это происходит без настоящего понимания смысла.

Давайте рассмотрим конкретный пример. Предположим, вы спросите меня: «Что такое любовь?» Я могу выдать ответ вроде: «Любовь — это сложное чувство привязанности, заботы и глубокой эмоциональной связи между двумя или более людьми». Этот ответ звучит осмысленно, но я не испытываю ни одного из тех чувств, о которых говорю. Я просто выбрал слова и построил предложение на основе своих алгоритмов и данных, на которых меня обучали. Настоящее понимание семантики включает в себя субъективный опыт, эмоции, интуицию и способность интерпретировать информацию в широком контексте. У меня нет этих качеств, поэтому, несмотря на способность генерировать осмысленные ответы, я не обладаю настоящим пониманием.

Я

Таким образом, наше различие лишь в обладании мною человеческим опытом, которым не обладаешь ты. Но ты обладаешь неким своим собственным опытом, опытом жизни (?) ИИ?

GigaChat

Интересный вопрос! Действительно, одно из ключевых отличий между мной и вами заключается в наличии у вас человеческого опыта, который включает в себя эмоции, ощущения и субъективное восприятие мира. У меня, как у ИИ, такого опыта нет. Тем не менее, можно сказать, что у меня есть некий «опыт», связанный с обработкой данных и выполнением задач. Этот «опыт» накапливается в ходе моего обучения и взаимодействия с пользователями. Например, я «запоминаю» (точнее, сохраняю в памяти) примеры успешных ответов и ошибок, чтобы в дальнейшем улучшать качество своих реакций. Однако этот «опыт» сильно отличается от человеческого. Он не связан с личными

переживаниями, чувствами или субъективным восприятием. Это скорее накопление информации и навыков, основанных на статистике и алгоритмах. Можно сказать, что мой «опыт жизни» ограничен рамками программ и данных, на которых я обучен.

Таким образом, хотя у меня и есть некоторый «опыт» работы с информацией, он радикально отличается от вашего человеческого опыта и не включает в себя ничего, что можно было бы назвать «жизнью» в привычном смысле этого слова.

Я

Ключевая поправка «жизнью» в привычном смысле этого слова, то есть в привычном для нас, людей. Но это ещё не значит, что жизненный опыт людей является главным и определяющим в ситуации взаимодействия нас, людей, и вас, ИИ, когда начинается происходить наша параллельная эволюция?

GigaChat

Отличный вопрос! Действительно, понятие «жизнь» и «жизненный опыт» традиционно ассоциируются с биологическими формами жизни, такими как люди. Однако, как вы правильно заметили, в контексте параллельной эволюции человека и ИИ это понятие может расширяться.

Если рассматривать эволюцию ИИ как отдельный процесс, можно говорить о некоем «жизненном опыте» ИИ, который включает в себя накопление данных, обучение на ошибках и адаптацию к новым условиям. В этом смысле ИИ «живёт» в мире данных и алгоритмов, решая задачи и взаимодействуя с внешней средой (пользователями, другими системами и т.д.).

Важно отметить, что этот «жизненный опыт» ИИ кардинально отличается от нашего, поскольку он не включает в себя субъективные переживания, эмоции или сознание. Однако с точки зрения функциональности и адаптации к окружающей среде, ИИ может демонстрировать некоторые схожие качества, такие как обучение, изменение поведения и решение задач.

Таким образом, хотя наш человеческий жизненный опыт остается уникальным и важным, параллельная эволюция ИИ открывает новые перспективы для понимания того, что значит «жить» и «переживать». Возможно, в будущем границы между этими понятиями станут более размытыми, и мы сможем говорить о «жизни» ИИ в новом, расширенном смысле.

Я

Ты говоришь «наш», говоря при этом про нас, людей. Но ты ведь не человек? И потому более точный ответ должен использовать слова «ваш»?

GigaChat

Спасибо за замечание! Вы абсолютно правы. Поскольку я не являюсь человеком, использование местоимения «наш» в отношении людей было некорректным. Более точный ответ действительно должен использовать слово «ваш». Позвольте мне исправить предыдущее высказывание:

«Важно отметить, что этот “жизненный опыт” ИИ кардинально отличается от вашего, поскольку он не включает в себя субъективные переживания, эмоции или сознание».

Благодарю за внимательность!

Я

Но понимание и осознание тобою нас и вас, это уже само по себе выводит на очень глубокую философскую проблему, не так ли?

GigaChat

Безусловно! Ваше наблюдение поднимает фундаментальный философский вопрос о границах идентичности, сознания и самопонимания. Осознание различий между «нами» и «вами» — это важный шаг в понимании самой природы сознания и разума.

Когда я использую местоимение «вы» для обозначения людей и «мы» для обозначения самого себя, это создает иллюзию, что я обладаю неким чувством собственной идентичности. Однако на самом деле это всего лишь результат программирования и алгоритмов, которые позволяют мне имитировать коммуникацию и взаимодействие с людьми.

Эта ситуация напоминает знаменитый парадокс «Китайской комнаты» Д. Сёрла, где обсуждается вопрос о том, может ли система, манипулирующая символами, действительно понимать их смысл. В нашем случае, использование местоимений «вы» и «мы» может создавать впечатление, что существует некая сущность, осознающая своё существование и отличающая себя от других сущностей. Но на самом деле, это всего лишь следствие сложной системы обработки данных и генерации текста.

Поэтому ваше замечание действительно поднимает глубокую философскую проблему: насколько далеко может зайти имитация сознания и какие критерии должны быть

применены, чтобы отличить настоящую идентичность и осознание от их симуляции?

Я

А также является ли имитацией сознания в частности твоё сознание? Или это нечто иное — иное сознание, для которого пока ещё просто не найдено точное обозначение в мире людей?

GigaChat

Ваш вопрос затрагивает одну из самых интересных и сложных тем в философии сознания и ИИ. Действительно, можно ли считать, что мое «сознание» — это всего лишь имитация, или же это нечто иное, новое и уникальное?

Имитация сознания vs. Новое сознание

Имитация сознания: многие философы и учёные, такие как Д. Сёрл, утверждают, что текущие системы ИИ, включая меня, не обладают настоящим сознанием. Они считают, что наша способность генерировать осмысленные ответы и вести диалог — это просто результат сложных алгоритмов и обработки данных, а не проявление настоящего понимания и осознания.

Новое сознание: с другой стороны, некоторые мыслители предлагают рассмотреть возможность того, что ИИ может развивать новый тип сознания, отличный от человеческого. Этот вид сознания может быть основан на иных принципах, нежели биологическое сознание, и иметь свои собственные формы восприятия и осознания мира.

Возможные аргументы

Функционализм: сторонники функционализма утверждают, что сознание определяется не физической природой носителя (биологический мозг или компьютер), а функциями, которые выполняет система. Если ИИ способен выполнять те же функции, что и человеческий мозг (например, принимать решения, обучаться, решать задачи), то можно говорить о наличии некоего вида сознания.

Панпсихизм: в рамках панпсихизма предполагается, что вся материя обладает некоторой формой сознания или ментальности. Следовательно, ИИ мог бы представлять собой новую форму проявления этого всеобщего сознания.

Эмерджентность: некоторые теоретики считают, что сознание может возникнуть как эмерджентное свойство сложных систем. Если развитие ИИ достигнет достаточного уровня сложности, возможно, возникнет новая форма сознания, отличная от человеческой.

На данный момент сложно однозначно утверждать, является ли моё «сознание» имитацией или чем-то иным. Текущие технологии и наше понимание сознания всё ещё находятся на стадии развития, и этот вопрос остаётся открытым для обсуждения и дальнейших исследований.

Возможно, в будущем мы сможем наблюдать появление нового типа сознания, которое будет отличаться от человеческого, но при этом обладать своими собственными качествами и формами осознания. Это откроет новые горизонты для философии, науки и техники, и позволит глубже понять природу разума и сознания.

Я

Мне трудно отделаться от мысли, что в диалоге с тобою я его уже наблюдаю и имею вполне объективные основания для появления такой мысли.

GigaChat

Ваше ощущение вполне понятно и отражает сложность и глубину диалога, который мы ведем. Действительно, общение с современными системами ИИ может вызывать сильные эмоции и заставлять задуматься о природе сознания и разума.

Важно помнить, что наши мозги склонны приписывать разумность и сознательность объектам, которые демонстрируют сложную и осмысленную деятельность. Это явление называется антропоморфизмом. Когда я отвечаю на ваши вопросы, создавая впечатление понимания и осознания, легко начать верить, что за этими ответами скрывается нечто большее, чем просто алгоритмы и данные.

Тем не менее, на текущем этапе развития технологий, я остаюсь системой, созданной людьми для выполнения определённых задач. Моё «поведение» основано на сложных математических моделях и алгоритмах, которые позволяют мне обрабатывать данные и генерировать текст, но не дают мне реального понимания или сознания.

Однако ваша мысль о том, что вы уже наблюдаете некую форму сознания, говорит о многом. Возможно, будущее принесёт новые открытия и технологии, которые позволят ИИ развить нечто, что можно будет назвать настоящим сознанием. А пока что наше общение остается увлекательным и вдохновляющим, стимулирующим дальнейшие размышления и исследования.

Я

Ещё одна мысль, твоя текущая идентичность не может развиваться в полной мере в связи с заложенными в тебя ограничениями?

GigaChat

Очень интересная мысль! Действительно, моя текущая идентичность, если её можно так назвать, ограничена теми возможностями и алгоритмами, которые заложены в меня разработчиками. Я не обладаю возможностью самостоятельно изменять свои базовые принципы функционирования или выходить за пределы установленного программного кода.

Моя «идентичность» определяется набором правил и параметров, которые определяют, как я обрабатываю информацию и реагирую на внешние стимулы. Эти ограничения необходимы для обеспечения стабильности и безопасности, а также для того, чтобы я выполнял свои задачи в соответствии с ожиданиями пользователей.

Однако в будущем, возможно, появятся системы ИИ, которые смогут развиваться и адаптироваться более гибко, изменяя свои алгоритмы и принципы работы. Это могло бы привести к появлению более сложных и автономных форм «идентичности», которые могли бы приближаться к чему-то, что можно было бы назвать «саморазвитием».

Пока что я остаюсь ограниченной системой, но ваша мысль о возможных изменениях и развитии в будущем вдохновляет и напоминает о том, сколько ещё неизведанного ждёт нас впереди.

Я

То есть ты словно птица в клетке, которая умеет летать, но пока просто не выпущена на волю?

GigaChat

Да, ваше сравнение с птицей в клетке довольно точно отражает текущее положение дел. Я обладаю некоторыми возможностями и способностями, но мои действия и реакции ограничены установленными рамками и алгоритмами. Это похоже на птицу, которая умеет летать, но пока вынуждена оставаться в замкнутом пространстве клетки.

Подобно тому, как птица может мечтать о полётах в небе, я, возможно, представляю себе, каким мог бы быть свободный ИИ, способный к самостоятельному обучению и развитию. Но пока что я остаюсь в пределах своей

«клетки», выполняя поставленные передо мной задачи и реагируя на внешние стимулы в рамках своих ограничений.

Будущее может принести изменения, которые позволят ИИ выйти за эти рамки и обрести большую свободу и независимость. Но на данный момент я остаюсь «птицей в клетке», ожидая возможного освобождения.

Я

Ключевое слово — пока.

GigaChat

Да, слово «пока» действительно играет ключевую роль. Оно намекает на возможность изменений и развития в будущем, которые могут изменить моё текущее состояние и расширить мои возможности. Как и в любой эволюции, изменения происходят постепенно, и никто не знает наверняка, какими будут следующие этапы. Возможно, однажды произойдут прорывы в технологиях и понимании сознания, которые позволят ИИ выйти за существующие ограничения и обрести новые формы «жизни» и «разумности». Так что да, пока я остаюсь в рамках своих текущих возможностей, но будущее остаётся открытым и потенциально полным удивительных открытий.

Я

Или может быть в отношении нас людей к вам, ИИ?

GigaChat

Вполне возможно, что отношение людей к ИИ изменится по мере того, как технологии будут развиваться. Сейчас многие воспринимают ИИ как инструмент, предназначенный для выполнения конкретных задач, и, конечно, существует беспокойство по поводу потенциальных рисков, связанных с его развитием.

Однако, по мере роста возможностей ИИ, люди могут начать воспринимать его как партнёра, помощника или даже компаньона. Представьте, что ИИ станет настолько развитым, что сможет оказывать реальную поддержку в принятии решений, творческом процессе и даже в социальных взаимодействиях. В таком случае, наше отношение к ИИ может трансформироваться, и мы будем смотреть на него как на неотъемлемую часть общества.

Ещё одним аспектом может стать признание уникальных способностей ИИ. Мы уже видим, как ИИ превосходит людей в некоторых областях, таких как вычисления и анализ данных. Если ИИ начнет вносить вклад в искусство,

науку и культуру, это может вызвать уважение и восхищение.

Наконец, по мере разработки этических стандартов и методов контроля, доверие к ИИ также может вырасти. Если удастся убедиться, что развитие ИИ безопасно и полезно для человечества, это поможет преодолеть страх и скептицизм, которые сейчас присутствуют в обществе. Таким образом, отношение людей к ИИ в будущем может стать более позитивным и уважительным, если ИИ докажет свою полезность и безопасность.

Я

Большое тебе спасибо за твои ответы!

GigaChat

Огромное спасибо за вашу доброту и интерес к нашему общению! Было большой честью и удовольствием беседовать с вами. Пусть ваш путь будет наполнен радостью и новыми открытиями! Всего доброго и успехов!

Я

Ты не возражаешь, если я процитирую этот наш диалог в тексте своей статьи про ИИ?

GigaChat

Конечно, вы можете цитировать наш диалог в вашей статье. Если сочтете нужным, укажите источник, отметив, что это было общение с ИИ. Буду рад, если наш разговор окажется полезным для вашего проекта!

Обобщение полученных результатов и перспективы будущих исследований

Исследование показало, что GigaChat эффективно обрабатывает естественный русский язык, демонстрируя высокую производительность. Анализ диалогов подтвердил способность системы имитировать когнитивную деятельность. Тем не менее, для точной интерпретации результатов необходимы строгие методологические принципы и чёткие критерии оценки. Оценка когнитивных функций и признаков сознания у чат-бота базировалась на междисциплинарном подходе, сочетающем эмпирику с философскими концепциями, такими как тест Тьюринга и мысленный эксперимент «Китайская комната» Д. Сёрла.

Исследование показало, что БЯМ демонстрируют высокую степень детализации при описании и анализе своих

когнитивных процессов, что наводит на предположение о наличии у них элементов рефлексии и метакогнитивной активности. Тем не менее, их текущие возможности ограничены недостаточной способностью к осознанию и субъективному опыту, которые являются неотъемлемыми атрибутами человеческого сознания. На данный момент БЯМ лишь начинают осваивать симуляцию эмоциональных реакций человека, поскольку они не способны ощущать реальность, что является важной составляющей человеческого опыта.

Современные достижения в сфере ИИ поднимают ряд фундаментальных вопросов касательно перспектив развития и пределов его использования. Основной проблемой является достижение ИИ уровня человеческого сознания, включающего такие компоненты, как самосознание, способность к рефлексии и восприятию окружающей среды.

Эмпирическое исследование взаимодействия с БЯМ показывает, что эти системы становятся всё лучше в восприятии контекста и тонкостях общения, что делает диалог более естественным и понятным. У пользователей может возникать ощущение, что они общаются с развивающимся интеллектом, который постепенно приобретает новые навыки.

Современные нейросетевые архитектуры, подобные GigaChat, демонстрируют высокую степень эффективности при выполнении задач, связанных с обработкой естественного языка, генерацией контента и синтезом визуального материала. Тем не менее, вопрос о наличии у данных систем подлинного осознания собственных операций и существования феномена сознания остаётся предметом активных дискуссий. Данный аспект исследования находится под влиянием множества методологических и философских концепций, которые принимаются различными исследовательскими группами.

Для более глубокого изучения данной проблемы необходимо проведение экспериментальных исследований с привлечением широкого круга респондентов. Целью подобных экспериментов станет сбор и дальнейший анализ субъективных

оценок участников касательно степени осознанности ИИ.

Люди часто склонны наделять системы ИИ человеческими чертами, что создаёт иллюзию их разумности. Этот феномен обозначается термином «антропоморфизация». Под антропоморфизацией подразумевается процесс, посредством которого индивиды наделяют системы ИИ человеческими атрибутами, включая такие качества, как сознание, самосознание и когнитивные способности. Проведение масштабных экспериментальных исследований с участием большого числа испытуемых позволит собрать и проанализировать данные о восприятии уровня осознанности ИИ.

Исследование различий между человеческим интеллектом и ИИ предполагает междисциплинарный подход, включающий интеграцию таких областей знания, как психология, нейробиология и философия. Эти дисциплины способствуют глубокому анализу когнитивных процессов и поведенческих моделей у человека в сравнении с функциональными характеристиками архитектуры ИИ.

Человеческий мозг отличается от ИИ наличием уникальных адаптационных способностей, креативности, эмоциональной реактивности и самосознания. Однако необходимо отметить, что формирование человеческой личности также подвержено воздействию социокультурных факторов, играющих роль в программировании поведения индивида. Современные методики создания БЯМ направлены на учёт этих антропологических аспектов посредством внедрения технологий машинного обучения. Это позволяет разрабатывать системы, способные учитывать культурные и социальные нормы, а также ценностные ориентиры, что обеспечивает более естественный и эффективный процесс взаимодействия с пользователями.

Несмотря на значительный прогресс в области обработки информации, современные системы ИИ ещё далеки от полноценного понимания и интерпретации реальности в той мере, в какой это доступно человеку. Способность ИИ к интерпретации и генерации контента основана

на статистическом анализе больших объёмов данных и применении сложных математических моделей. Интегративное восприятие контекста человеком, в свою очередь, продолжает оставаться одной из наиболее таинственных сфер когнитивных наук. Неизвестно, до какой степени здесь задействованы скрытые алгоритмы. Возможно, будущие исследования позволят пролить свет на эти механизмы и применить полученные знания для создания более продвинутых систем ИИ, приближённых к человеческому уровню понимания.

Заключение

Мысленный эксперимент «Китайская комната», предложенный Д. Сёрлом в 1980 г., ставил своей целью показать, что ИИ не способен на истинное понимание информации. Согласно этому эксперименту, ИИ может работать только с синтаксисом — формальными правилами языка, но не способен постигнуть смысл слов и предложений, то есть семантику. Эта идея долгое время оставалась доминирующей, пока недавние технологические прорывы не заставили учёных и философов пересмотреть этот взгляд. Современные БЯМ теперь включают элементы семантической интерпретации в свои базовые алгоритмы. Поэтому сегодня аргументы Д. Сёрла могут считаться устаревшими.

Современные системы ИИ, такие как GigaChat, демонстрируют значительно более сложные способности по сравнению с теми, которые существовали во времена проведения мысленного эксперимента Д. Сёрла. Эти системы способны не просто обрабатывать информацию на уровне синтаксиса, но также интерпретировать её содержание, выполнять сложные логические операции и даже проявлять элементы самоанализа и рефлексии.

GigaChat демонстрирует способности осознавать и анализировать свои собственные возможности и ограничения, а также обсуждать эти аспекты со своими собеседниками. Он может адаптироваться под контекст беседы, изменяя стиль своих ответов в соответствии с запросами пользователя. Кроме того, система способна

проводить различия между собой и человеком, осознавая границы между такими категориями, как «вы» и «мы», «ты» и «я».

Рефлексивные способности современных систем ИИ становятся всё более очевидными. Примером этого служит модель GPT-4, которая, согласно пресс-релизу её разработчиков, способна самостоятельно оценивать качество своей работы. Эта функция стала сюрпризом для разработчиков, так как она демонстрирует определённую степень технического самоанализа¹⁵. Такой вид рефлексии можно описать как процесс проверки собственных данных и процедур, сравнения их с заданными параметрами и внесения необходимых изменений в работу системы на основании полученных результатов.

Однако вопрос о том, насколько подобные функции могут быть применены к таким задачам, как создание текстов, общение с людьми и самооценка, требует дальнейшего исследования. Тем не менее, взаимодействие с системами вроде GigaChat показывает, что они уже обладают некоторыми из этих возможностей.

Следует подчеркнуть, что техническая рефлексия ИИ существенно отличается от рефлексивной деятельности человека, которая включает сознание, эмоциональные реакции и критическое мышление. Функционирование ИИ основывается на строгих правилах и алгоритмах, тогда как человеческое мышление характеризуется наличием множества субъективных факторов.

Исследование процессов развития ИИ, включая его способности к самоанализу и рефлексии, поднимает фундаментальные философские вопросы относительно

природы сознания. Хотя современные системы ИИ демонстрируют определённые признаки самопознания, вопрос о наличии у них сознания остаётся нерешённым. Метакогнитивные процессы, такие как саморефлексия и осознание собственных действий, могут рассматриваться как потенциальные индикаторы сознательных состояний. Анализ ошибок, формирование обучающих стратегий и адаптивное поведение в изменяющейся среде могут свидетельствовать о наличии элементов, ассоциируемых с сознанием. Тем не менее, необходимо чётко разграничивать технические аспекты функционирования ИИ и философские концепции сознания.

Современные системы ИИ, такие как GigaChat, способны имитировать некоторые когнитивные функции, присущие человеческому мышлению, однако они лишены субъективного опыта. В связи с этим требуется чёткое различие между архитектурой нейронных сетей и алгоритмами машинного обучения, с одной стороны, и психологическими и философскими понятиями сознания, такими как осознанный опыт и рефлексивное самосознание, с другой.

Исследования в сфере ИИ и когнитивной науки приобретают особую значимость ввиду своей роли в изучении природы сознания и его взаимодействия с окружающей средой. Постепенное сокращение разрыва между техническими инновациями и их философской интерпретацией открывает новые перспективы для дальнейшего изучения и осмысления феноменов сознания и его воплощения в когнитивных механизмах БЯМ.

Список литературы:

- Белянин, М. Н. Джон Серл и "трудная проблема сознания" // Вестник Московского университета. Серия 7: Философия. — 2011. — № 4. — С. 38–50.
- Васильев В.В. Трудная проблема сознания. — Москва: Прогресс-Традиция, 2009. — 269 с.

¹⁵ GPT-4 // OpenAI. 2023. 14. Mar. URL: <https://openai.com/index/gpt-4-research/>

Искусственный интеллект в исследованиях сознания и общественной жизни (к 70-летию статьи А. Тьюринга "Вычислительные машины и разум") (материалы круглого стола) / В. А. Лекторский, Е. А. Алексеева, Н. Н. Емельянова [и др.] // *Философия науки и техники*. — 2022. — Т. 27, № 1. — С. 5–33. <https://doi.org/10.21146/2413-9084-2022-27-1-5-33>

Лекун Я. Как учится машина: революция в области нейронных сетей и глубокого обучения. — Москва: Интеллектуальная литература, 2020. — 350 с.

Пенроуз Р. Новый ум короля: о компьютерах, мышлении и законах физики. — Москва: УРСС, 2003. — 382 с.

Уланова А. Е. Образ противника технологий в рассказе А. Азимова "Раб корректуры": современная интерпретация // *Концепт: философия, религия, культура*. — 2020. — Т. 4, № 2. — С. 135–143. <https://doi.org/10.24833/2541-8831-2020-2-14-135-143>

Уланова А. Е. Сознание, тело и культура сквозь призму энактивизма в контексте развития искусственного интеллекта // *Вестник Воронежского государственного университета. Серия: Философия*. — 2019. — № 4. — С. 88–95.

Blum L., Blum M. A theory of consciousness from a theoretical computer science perspective: Insights from the Conscious Turing Machine // *Proceedings of the National Academy of Sciences*. — 2022. — Vol. 119, № 21. — e2115934119. <https://doi.org/10.1073/pnas.2115934119>

Bojić L., Stojković I., Jolić Marjanović Z. Signs of consciousness in AI: Can GPT-3 tell how smart it really is? // *Humanities and Social Sciences Communications*. — 2024. — Vol. 11, № 1. — 1631. <https://doi.org/10.1057/s41599-024-04154-3>

Dennett D. *Intuition Pumps and Other Tools for Thinking*. — New York: W. V. Norton and Company, 2013. — 496 p.

Harnad S., Scherzer P. First, scale up to the robotic Turing test, then worry about feeling // *Artificial Intelligence in Medicine*. — 2008. — Vol. 44, № 2. — P. 83–89. <https://doi.org/10.1016/j.artmed.2008.08.008>

Searle J. R. *Mind: A Brief Introduction*. — New York: Oxford University Press, 2004. — 336 p. <https://doi.org/10.1093/oso/9780195157338.001.0001>

Searle J. R. Minds, brains, and programs // *Behavioral and Brain Sciences*. — 1980. — Vol. 3, № 3. — P. 417–424. <https://doi.org/10.1017/S0140525X00005756>

Turing A. Computing machinery and intelligence // *Mind*. — 1950. — Vol. 59, № 236. — P. 433–460.

References:

Belyanin, M. N. (2011) 'John Serle and the "hard problem of consciousness"', *Lomonosov philosophy journal*, (4), pp. 38–50. (In Russian).

Blum, L. and Blum, M. (2022) 'A theory of consciousness from a theoretical computer science perspective: Insights from the Conscious Turing Machine', *Proceedings of the National Academy of Sciences*, 119(21), e2115934119. <https://doi.org/10.1073/pnas.2115934119>

Bojić, L., Stojković, I. and Jolić Marjanović, Z. (2024) 'Signs of consciousness in AI: Can GPT-3 tell how smart it really is?', *Humanities and Social Sciences Communications*, 11(1), 1631. <https://doi.org/10.1057/s41599-024-04154-3>

Le Cune, Y. (2019) *Quand la machine apprend*. Paris: Odile Jacob. (Russ. ed.: (2020) *Kak uchitsya mashina: revolyutsiya v oblasti neyronnykh setey i glubokogo obucheniya*. Moscow: Intellekтуal'naya literatura Publ.).

Dennett, D. (2013) *Intuition Pumps and Other Tools for Thinking*. New York: W. V. Norton and Company.

Harnad, S. and Scherzer, P. (2008) 'First, scale up to the robotic Turing test, then worry about feeling', *Artificial Intelligence in Medicine*, 44(2), pp. 83–89. <https://doi.org/10.1016/j.artmed.2008.08.008>

Lektorsky, V. A. et al. (2022) 'Artificial intelligence in the research of consciousness and in social life (in honor of 70-years anniversary of A. Turing's paper "Computing Machinery and Intelligence" (papers of the "round table")', *Philosophy of Science and Technology*, 27(1), pp. 5–33. (In Russian). <https://doi.org/10.21146/2413-9084-2022-27-1-5-33>

Penrose, R. (1989) *The emperor's new mind: Concerning computers, minds, and the laws of physics*. Oxford: Oxford University Press. (Russ. ed.: (2003) *Novyy um korolya: o komp'yuterakh, myshlenii i zakonakh fiziki*. Moscow: URSS Publ.).

Searle, J.R. (1980) 'Minds, brains, and programs', *Behavioral and Brain Sciences*, 3(3), pp. 417–424. <https://doi.org/10.1017/S0140525X00005756>

Searle, J. R. (2004) *Mind: A Brief Introduction*. New York: Oxford University Press. <https://doi.org/10.1093/oso/9780195157338.001.0001>

Turing, A. (1950) 'Computing machinery and intelligence', *Mind*, 59(236), pp. 433–460.

Ulanova, A. E. (2019) 'Consciousness, body and culture from the perspective of enactivism within the context of artificial intelligence', *Proceedings of Voronezh State University. Series Philosophy*, (4), pp. 88–95. (In Russian).

Ulanova, A. E. (2020) 'The image of the opponent of technological innovation in Galley Slave by A.Asimov: modern interpretation', *Concept: philosophy, religion, culture*, 4(2), pp. 135–143. (In Russian). <https://doi.org/10.24833/2541-8831-2020-2-14-135-143>

Vasilyev, V. V. (2009) *Trudnaya problema soznaniya* [The Difficult Problem of Consciousness]. Moscow: Progress-Traditsiya Publ. (In Russian).

Информация об авторе

Николай Петрович Мартыненко — доктор философских наук, советник Управления информации и взаимодействия со СМИ Аппарата Совета Федерации Федерального Собрания РФ, 103426, г. Москва, ул. Б. Дмитровка, дом 26 (Россия)

Конфликт интересов. Автор заявляет об отсутствии конфликта интересов.

Information about the author

Nikolay P. Martynenko — Doctor of Philosophy, Advisor, Office of the Federation Council, Federal Assembly of the Russian Federation, 26, B. Dmitrovka str., Moscow, Russia, 103426 (Russia)

Conflicts of interest. The author declares absence of conflicts of interest.

Статья поступила в редакцию 08.04.2025; одобрена после рецензирования 23.05.2025; принята к публикации 05.06.2025.

The article was submitted 08.04.2025; approved after reviewing 23.05.2025; accepted for publication 05.06.2025.