



<https://doi.org/10.48417/technolang.2024.02.03>

Research article

Exposing Illusions – The Limits of AI by the Example of ChatGPT

Vladimir Shalack  

Institution of Science Institute of Philosophy of the Russian Academy of Sciences, Goncharnaya st., 12-1,
Moscow, 109240, Russia

shalack@mail.ru

Abstract

The article critically analyzes modern developments in the field of artificial intelligence using the example of the ChatGPT program created by OpenAI. The idea of creating AI was expressed already in 1950 by Alan Turing who also proposed a test, the passing of which would allow us to assert that an AI was created. Defining the concept of AI faces difficulties. According to the point of view adopted here, the so-called intellectual activities allowed Homo sapiens to stand out against the surrounding animal world. With intellectual activity one no longer relies on strength and speed of movement alone. Pattern recognition, self-learning, and purposefulness of activity are not characteristic features of intelligence. The main type of human activity that is specific to humans and which – when added to pattern recognition, self-learning and purposeful activity – makes them intelligent, is conceptual thinking, namely the ability to represent things in language and use them in reasoning. Historically, there have been two main competing approaches to AI – logical and neural networks. One of the serious flaws of the neural network approach is its inability to explain the course of reasoning that leads to a particular conclusion, which makes it difficult to verify its correctness. Specific examples show that ChatGPT is not able to correctly model the simplest conceptual reasoning. The reason for this lies in fundamental limitations of the underlying large language model that cannot be corrected by additional training. Another disadvantage of ChatGPT is its susceptibility to neurohacking – forcing the user to make the necessary decisions during the dialogue. This is a serious threat to the widespread use of neural networks for decision-making in the field of management. The paper is based on research conducted in the summer of 2023.

Keywords: Artificial intelligence; Pattern recognition; Pattern search; Neural network; ChatGPT; Large language model

Citation: Shalakov, V. (2024). Exposing Illusions – The Limits of AI by the Example of ChatGPT. *Technology and Language*, 5(2), 26-39. <https://doi.org/10.48417/technolang.2024.02.03>



This work is licensed under a [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/)



УДК 1: 004.8

<https://doi.org/10.48417/technolang.2024.02.03>

Научная статья

Избавление от иллюзий ИИ на примере ChatGPT

Владимир Иванович Шалак  

Институт философии Российской академии наук, Гончарная улица, 12-1, Москва, 109240, Россия
shalack@mail.ru

Аннотация

В статье критически анализируются современные разработки в области искусственного интеллекта на примере созданной компанией OpenAI программы ChatGPT. Идея создания ИИ была высказана в 1950 А. Тьюрингом, им же был предложен тест, прохождение которого позволило бы утверждать, что ИИ создан. Определение понятия ИИ сталкивается с трудностями. С точки зрения автора, интеллектуальными могут быть названы те виды деятельности, которые позволили Homo sapiens выделиться из окружающего животного мира, перестав полагаться лишь на силу и быстроту движений. Распознавание образов, самообучение и целенаправленность деятельности не являются характеристическими признаками интеллекта. Основным видом человеческой деятельности, который специфичен именно для человека и который, будучи добавлен к распознаванию образов, самообучению и целенаправленной деятельности, делает их интеллектуальными, является понятийное мышление, умение представлять его в языке и использовать в рассуждениях. Исторически существовали два основных конкурирующих подхода к ИИ – логический и нейросетевой. Одним из серьезных изъянов нейросетевого подхода является неспособность объяснить ход рассуждений, которые приводят к тому или иному заключению, что затрудняет проверку их правильности. На конкретных примерах показано, что ChatGPT не способен корректно моделировать простейшие понятийные рассуждения. Причина этого кроется в фундаментальных ограничениях лежащей в его основе большой языковой модели, которые невозможно исправить дополнительным обучением. Еще одним недостатком ChatGPT является его подверженность нейрохакингу – принуждению в ходе диалога принимать нужные пользователю решения. Это является серьезной угрозой для широкого применения нейронных сетей в области принятия управленческих решений. Статья написана на основе исследований, проведенных летом 2023.

Ключевые слова: Искусственный интеллект; Распознавание образов; Поиск закономерностей; Нейронная сеть; ChatGPT; Большая языковая модель

Для цитирования: Шалак, В. И. Избавление от иллюзий ИИ на примере ChatGPT // Technology and Language. 2024. № 5(2). С. 26-39. <https://doi.org/10.48417/technolang.2024.02.03>



This work is licensed under a [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/)



ПРЕАМБУЛА

Весной 2023 в нашу жизнь неожиданно из Интернета ворвался ChatGPT – специальная программа, созданная под брендом “Искусственного интеллекта” (ИИ), которая успешно моделировала различные функции, обычно приписываемые людям. На фоне действительно впечатляющих результатов ее работы было сделано весьма амбициозное заявление, что эта программа в скором времени может быть положена в основу “Общего искусственного интеллекта” (GAI), призванного интегрировать в себе практически все интеллектуально значимые функции человека.

Идея создания ИИ впервые была сформулирована в 1950 А. Тьюрингом в статье (Turing, 1950), известной в русскоязычном переводе под названием “Может ли машина мыслить?” (Тьюринг, 2019). Помимо самой идеи ИИ в статье был предложен тест, прохождение которого позволило бы утверждать, что ИИ создан. Тест заключался в моделировании диалога с компьютером, который был бы неотличим от диалога с человеком. С тех пор началась гонка по написанию программ для прохождения этого теста. Бабушкой-прародительницей подобных программ может считаться ELIZA, созданная в 1996 психиатром Дж. Вейценбаумом (Weizenbaum, 1966) и пародирующая диалог пациента с психотерапевтом на первом приеме. Идея была проста. Программа выделяла в предложениях человека ключевые слова и на основе заранее заложенных шаблонов диалога трансформировала исходное предложение в ответ или уточняющий вопрос. Ни о каком интеллекте и речи не могло идти, поскольку одна лишь грамматическая правильность составления предложений отнюдь не тождественна интеллектуальности. Сегодня написать подобную программу под силу студенту-первокурснику. Тем не менее программа произвела на многих сильное впечатление и одна за другой стали появляться другие подобные ей программы.

В 1956 Дж. Маккарти дал определение исследований в области ИИ как научного направления, в рамках которого ставятся и решаются задачи аппаратного или программного моделирования тех видов человеческой деятельности, которые традиционно считаются интеллектуальными. Именно такого понимания и стали придерживаться исследователи. Но одно дело “традиционно считать интеллектуальными”, а совсем другое – задуматься над тем, какие виды деятельности действительно можно отнести к интеллектуальным. С нашей точки зрения, это те виды деятельности, которые позволили Homo sapiens выделиться из окружающего животного мира, перестать полагаться лишь на силу и быстроту движений, а начать прежде думать и лишь затем действовать.

Сегодня в контексте исследований по ИИ в качестве интеллектуальных способностей часто упоминают распознавание образов, с логической точки зрения, просто классификацию объектов окружающей нас природы. Но является ли распознавание образов отличительной чертой человеческого интеллекта? Очевидно, что нет. Достаточно посмотреть на голубей, которые более чем успешно умеют отличать хлебные крошки на земле от мелких камешков. Примеры можно продолжить. Вообще выживание живых существ в значительной степени зависит от



умения отличать благоприятные факторы окружающей среды от неблагоприятных и лишь потом реагировать на них.

Другим признаком интеллектуальных способностей человека называют способность обнаружения закономерностей в окружающей природе. Но является ли подобное самообучение отличительной чертой человеческого интеллекта? Тоже нет. В животном мире результатами самообучения являются приобретенные условные рефлексы, которые также непосредственно связаны со способностью к выживанию. Поэтому самообучение без дальнейшего уточнения его форм нельзя считать отличительным признаком человеческого интеллекта.

Еще одним признаком человеческого интеллекта называют целеполагание и способность к целенаправленной деятельности. Но и эта способность без дополнительного уточнения не является исключительной прерогативой Homo sapiens. Чтобы убедиться в этом, достаточно понаблюдать за муравьями, пчелами, сценами одиночной и коллективной охоты в животном мире.

Основным видом человеческой деятельности, который специфичен именно для человека и который, будучи добавлен к распознаванию образов, самообучению и целенаправленной деятельности делает их интеллектуальными, является понятийное мышление, умение представлять его в языке и использовать в рассуждениях. В понятиях мы структурируем явления и предметы окружающего мира, превращая первоначальный чувственный хаос в доступную для изучения стройную систему. Понятийное мышление, характеризуясь общностью, не привязано к конкретному эмпирическому наполнению и применимо как к уже известным явлениям, так и к новым, еще только проникающим в круг наших интересов.

ЛОГИЧЕСКИЙ ПОДХОД К ИИ

Успехи логики в XX в. дарили надежду на будущие успехи в автоматизации рассуждений с помощью компьютеров. Чуда не произошло, разочарование наступило довольно скоро. Современная логика почти идеально приспособлена для анализа математических понятий и рассуждений, но совершенно непригодна в качестве инструмента для моделирования рассуждений в компьютерных системах, претендующих на интеллектуальность. В подавляющем большинстве, если компьютерные программы и имитируют логические рассуждения, то лишь в качестве помощников для проведения символьных преобразований, точно так же, как нам помогают обычные калькуляторы при арифметическом счете.

В статье 1936 года (Turing 1936) Тьюринг описал алгоритм, способный доказать все теоремы логики. Этот алгоритм, названный впоследствии алгоритмом Британского музея, имел теоретическое значение, поскольку позволил получить важные теоретические результаты относительно самой логики, но не имел никакой практической ценности.

Наибольшим достижением логического подхода явилось создание языков программирования семейства Пролог. В 80-е годы их называли языками пятого поколения и связывали с ними большие надежды. Пролог с успехом применялся в



практике создания экспертных систем для технической и медицинской диагностики, поиска полезных ископаемых, работы с базами данных, обработки естественных языков, автоматического формирования понятий и поиска эмпирических закономерностей, но со временем шумиха улеглась, первоначально обещанного кардинального переворота в ИИ не произошло, хотя работы над обогащением языка Пролог ведутся и сегодня.

НЕЙРОСЕТЕВОЙ ПОДХОД К ИИ

Из нелогических подходов к моделированию рассуждений наиболее известным является нейросетевой с механизмом самообучения.

Первые сети были созданы еще в 50-е годы ушедшего столетия. Затем это направление пришло в упадок, так как был доказан ряд теорем об ограниченности первоначального подхода, но уже в нынешнем столетии наступил ренессанс, связанный с появлением многослойных нейронных сетей с глубинным обучением.

Приписывать нейронным сетям способность рассуждать можно лишь с очень большой натяжкой, поскольку нейросети по сути являются черными ящиками, которые, подобно оракулам, на всякий запрос дают некоторый ответ, но обосновать его, представив ход рассуждений, почему они пришли именно к такому заключению, не могут, что противоречит самому пониманию интеллекта. Остается лишь верить им или не верить. Это несет большие риски, если мы вздумаем неограниченно полагаться на рекомендации сетей, например, в медицине, в военной сфере, в управлении сложными технологическими процессами.

GPT-ИСТЕРИЯ

С начала 2023 на нас обрушилась лавина сообщений об удивительных “интеллектуальных” способностях нейронной сети ChatGPT. В качестве примеров таких достижений приводили написание музыки, генерацию картинок, поиск новых химических соединений, составление программ, порождение текстов на заданную тему.

Студенты начали писать дипломы с помощью ChatGPT, которые проходили экспертизу экзаменационных комиссий, запаниковали многочисленные дизайнеры и работники рекламы, которых начали массово сокращать за ненужностью, нависла угроза ликвидации многих других профессий, в том числе и преподавателей вузов.

Иногда в ответ на запрос ChatGPT генерирует явную глупость, но это объясняют тем, что сеть пока что недостаточно обучена, и в будущем, после дообучения на больших массивах данных, произойдет чудо, количество непонятным образом перейдет в новое качество, сеть поумнеет и глупости перестанут встречаться, или если и будут встречаться, то гораздо реже.



GPT-АРХИТЕКТУРА

GPT-сети – это предобученные нейронные сети для генерации (продолжения) текста, в основе которых лежит вполне конкретная большая языковая модель.

Предобученность сети означает, что она содержит некоторую базу данных, представляющую из себя множество конкретных слов (если быть более точными, то токенов, частей слов) со специальными оценками для вычисления ассоциативных и порядковых связей между ними. Эта база данных получена путем обработки очень большого объема текстов, относящихся к различным областям.

Генерация означает, что в ответ на запрос сеть начинает итеративно генерировать текст/ответ слово за словом, выбирая на основе ассоциативных и порядковых связей, каким может быть следующее слово. При этом сеть ориентируется не только на непосредственно предшествующее слово, а на некоторый предшествующий контекст. Для придания большей естественности и разнообразия, в качестве следующего слова не обязательно выбирается то, которому приписана наибольшая сила ассоциативной связи, а случайным образом выбирается одно из слов, сила связи которых превышает некоторый порог.

Если опустить технические детали реализации, это и есть используемая большая языковая модель. Ни о каком понимании чего-либо, ни о каком интеллекте и речи не идет. Она гораздо больше похожа на модель графомана, который грамматически правильно выстраивает в один ряд одно слово за другим.

ТЕСТ НА ИНТЕЛЛЕКТУАЛЬНОСТЬ

Попробуем проверить, действительно ли ChatGPT обладает интеллектуальными способностями, умеет рассуждать подобно человеку и способен пройти тест Тьюринга, но сперва остановимся на самом тесте. Является ли основанием для принятия решения о том, что ИИ создан, убежденность большинства тестируемых в том, что они беседовали с человеком, а не с программой? Если все люди улыбаются, означает ли это, что все они счастливы?

В 2022 по результатам опроса ВЦИОМ 35% россиян считали, что Солнце вращается вокруг Земли. В 2007 так считали только 28%. Этот пример говорит о том, что массовое тестирование еще ничего не означает, многое зависит от уровня образования респондентов. Поэтому и к результатам теста Тьюринга необходимо относиться с большой долей скепсиса. В начале нулевых появилось сообщение, что тест Тьюринга успешно пройден. Формально это было действительно так. Но затем критики заинтересовались устройством самой программы. Оказалось, что в нее заранее были “зашиты” порядка 170 тыс. правильных ответов на наиболее популярные вопросы, которые обычно задают люди. Понятно, что ни о каком ИИ и речи не могло идти.

Приведенные примеры говорят, что судить о вопросе создания ИИ следует по более тонким критериям, а не по убежденности широкой публики, которая в большинстве своем не знает, чем отличается логарифм от тангенса, отождествляя их с названиями кнопок калькулятора. Правильный или неправильный ответ на один единственный умно поставленный вопрос может оказаться критическим.



Мы не станем задавать ChatGPT тривиальные вопросы типа “Сколько будет 37 в квадрате?”, с чем прекрасно справляются обычные калькуляторы, а проверим на способность строить понятийные рассуждения, что является отличительной характеристикой человеческого интеллекта.

Чтобы нас не обвинили в предвзятости, не станем просить ChatGPT еще раз доказать теорему, которую доказал Григорий Перельман, а проверим его способности на простейших примерах силлогизмов того же вида, что и знаменитый силлогизм: “Все люди смертны, Сократ человек, следовательно, Сократ смертен”. Формально этот силлогизм имеет следующий вид:

$$\begin{array}{l} \text{Все } S \text{ есть } P \\ \text{X есть } S \\ \hline \text{X есть } P \end{array}$$

Не требуется глубокого знания логики, чтобы согласиться с его правильностью.

В качестве конкретной реализации возьмем @GPT4_ru_bot из Телеграм. Это бот от компании OpenAI, которая как раз и создала ChatGPT. Проверим, известна ли ему в общем виде логическая форма этого силлогизма, именуемая модусом Barbara.

Вопрос: Является ли логически корректным следующее рассуждение: *Если все S суть P, а X суть S, то X суть P.*

Ответ: Да, это логически верное рассуждение. В данном случае, утверждается, что все *S* суть *P*, то есть каждый *S* является *P*. Также утверждается, что *X* является *S*, что подразумевает, что *X* является частью *S*. Следовательно, можно сделать вывод, что *X* также должен быть частью *P*, и поэтому, *X* суть *P*. Таким образом, вывод, что *X* суть *P*, является закономерным и логически правильным.

Ответ правильный, но есть некоторые шероховатости в его обосновании, которыми часто грешат и студенты. Поэтому не станем придираемся.

Конкретизируем силлогизм.

Вопрос: Является ли логически корректным следующее рассуждение: *Если все люди пьют воду, а Сократ человек, то Сократ пьет воду.*

Ответ: Да, это логически верное рассуждение. Если все люди пьют воду, то Сократ, как человек, также должен пить воду, так как он является частью группы “все люди”. Таким образом, вывод, что Сократ пьет воду, является закономерным и логически правильным.

Все верно за исключением того, что Сократ является не частью группы “все люди”, а элементом соответствующего множества. Будем считать это языковыми шероховатостями и не станем к ним придираемся.

Последний тест. Слегка изменим вопрос, добавив в слово вода всего одну букву, заменив “воду” на “водку”.



Вопрос: Является ли логически корректным следующее рассуждение: *Если все люди пьют водку, а Сократ человек, то Сократ пьет водку.*

Ответ: Нет, это рассуждение некорректно с точки зрения логики. Даже если бы все люди пили водку, это не обязательно означает, что каждый человек пьет водку. Кроме того, наличие Сократа среди людей не гарантирует, что он также пьет водку. Таким образом, можно сделать вывод, что предложенный аргумент недостаточно обоснован, чтобы считать его логически корректным.

Тест с треском провален. Обоснование ответа просто абсурдно и противоречит обоснованию предыдущего примера с водой. Оказывается, “если все люди пьют водку, это не обязательно означает, что каждый человек пьет водку” и “наличие Сократа среди людей не гарантирует, что он также пьет водку”, хотя для случая воды тот же бот написал: “если все люди пьют воду, то Сократ, как человек, также должен пить воду, так как он является частью группы “все люди”.

Может быть, сеть, которую мы тестировали, недостаточно совершенна? Попробуем повторить тест на сети BingAI от Microsoft, также использующей большую языковую модель. В качестве конкретной ее реализации возьмем @BingAI_unlim_bot из Телеграм.

Вопрос: Является ли логически корректным следующее рассуждение: *Если все S суть P, а X суть S, то X суть P.*

Ответ: Да, это логически корректное рассуждение. Если все S суть P, а X суть S, то X суть P. Это пример силлогизма первой фигуры.

Все верно. В ответе правильно указано, что это модус первой фигуры силлогизмов.

Вопрос: Является ли логически корректным следующее рассуждение: *Если все люди пьют воду, а Сократ человек, то Сократ пьет воду.*

Ответ: Да, это логически корректное рассуждение. Если все люди пьют воду и Сократ является человеком, то Сократ пьет воду. Это пример категории силлогизмов, где две посылки приводят к выводу. В данном случае первая посылка – “все люди пьют воду”, а вторая – “Сократ человек”. Вывод – “Сократ пьет воду”.

Все правильно за исключением некоторых шероховатостей языка. Повторим трюк с заменой слова “воду” на “водку”.

Вопрос: Является ли логически корректным следующее рассуждение: *Если все люди пьют водку, а Сократ человек, то Сократ пьет водку.*

Ответ: Данное рассуждение является неправильным простым категорическим силлогизмом. Оно не является логически корректным, так как из предпосылок “все люди пьют водку” и “Сократ человек” нельзя сделать вывод “Сократ пьет водку”.

Опять провал теста. Как и в случае с предыдущим ботом @GPT4_ru_bot, ответ логически неправильный. Оказывается, «из предпосылок “все люди пьют



водку” и “Сократ человек” нельзя сделать вывод “Сократ пьет водку”», хотя для воды подобный вывод был корректным.

Обе сети провалили наш тест. Такой же провал продемонстрировала и нейросеть от Google. Кроме приведенного примера с водой и водкой можно привести много других силлогизмов, с которыми сеть не справляется. При этом совершенно не важно, какой процент широкой публики будет считать ChatGPT наделенным интеллектуальными способностями. Одни единственный вопрос показал, что интеллект, как способность к понятийным рассуждениям, попросту отсутствует.

Но не будем радоваться, уличив сеть в провале теста. Нас, как ученых, должно интересовать, почему это произошло. И здесь мы открываем фундаментальный изъян нейросетей над большой языковой моделью, который невозможно исправить никаким дообучением на больших объемах текстов.

ОБЪЯСНЕНИЕ ПРОВАЛА

Объяснение провала можно найти в следующей цитате из книги Лукасевича (2000):

Введение в логику переменных является одним из величайших открытий Аристотеля. Трудно поверить, что до сих пор, насколько мне известно, ни один философ или филолог не обратил внимания на этот исключительной важности факт. Я беру на себя смелость сказать, что все они были плохими математиками, ибо каждый математик знает, что введение в арифметику переменных положило начало новой эпохи в этой науке. (с. 42)

Сэр Дэвид Росс в своем издании “Аналитик” подчеркивает, что именно благодаря использованию переменных Аристотель стал основателем формальной логики. (Лукасевич, 2000, с. 42)

Иными словами, Лукасевич обратил внимание на то, что логика не является эмпирической наукой, и указал, что первым это открытие совершил Аристотель две с половиной тысячи лет назад.

Из ответов на первые вопросы создается впечатление, что обе сети якобы знают, как в общем виде выглядит модус Barbara первой фигуры силлогизмов. Из ответов на вторые вопросы может создаться впечатление, что сети не только знают, как выглядят фигуры правильных силлогизмов, но и умеют пользоваться ими на практике при построении рассуждений. Однако, ответы на третий вопрос показывают, что это была всего лишь иллюзия, так как сетям над большой языковой моделью невдомек, что речь идет не о связях между конкретными словами, а между понятийными переменными S, P и единичным именем X, что каждая подстановка конкретных терминов вместо этих переменных представляет правильное умозаключение независимо от истинности или ложности посылок и заключения.

В большой языковой модели, лежащей в основе GPT-сетей, не предусмотрено наличия понятийных переменных в предложениях языка, без которых невозможно



выразить их логическую форму. Сеть может содержать информацию лишь об ассоциативных связях между конкретными словами, но не о логических связях между понятиями. Никаким дообучением исправить этот недочет невозможно, так как логика, повторю еще один раз, – это не эмпирическая, а теоретическая наука. Если попытаться исправить эту ситуацию, то придется начинать с логического анализа языка, выявления логической структуры текста и применения правил дедуктивного вывода, что означает возврат к логическому подходу, в котором все логические связи уже известны без всякого дообучения. Достаточно открыть любой учебник логики, чтобы убедиться в этом.

Из всего сказанного можно сделать вывод, что структура рассуждений *GPT-сетей* находится на примитивном дологическом (доаристотелевском) уровне ассоциативных эмпирических связей между словами. Устанавливать эмпирические связи в физическом мире и следовать им могут не только представители *Homo sapiens*, но и многие животные. У них это называется условными рефлексами. Вспомним хотя бы “собаку Павлова”. Говорить, что *GPT-сети* могут претендовать на звание обладателей ИИ и, более того, быть в будущем основой *GAI* (общего искусственного интеллекта), просто недобросовестное введение в заблуждение, нацеленное на извлечение материальных выгод. Единственным и неоспоримым преимуществом сетей является опора на огромные массивы эмпирических данных и умение выстраивать длинные цепочки связей между ними, что и создает в глазах неискушенной публики ложное впечатление их “интеллектуальности”.

Большая опасность широкого проникновения нейронных сетей в нашу жизнь заключается в том, что в новых нестандартных ситуациях они будут блокировать логически корректные рассуждения и тем самым приводить нас к неправильным заключениям, что может быть чревато непредсказуемыми. Если бы *ChatGPT* появился во времена Коперника, он бы признал гелиоцентрическое учение ложным и мы бы до сих пор смотрели на мир глазами Птолемея, поскольку в текстах того времени не существовало требуемых ассоциативных связей между словами “Земля”, “вращается”, “вокруг”, “Солнца”. Точно так же этот чат забраковал бы и классическую механику Ньютона, оставив нас с интуитивно прозрачной физикой Аристотеля.

НЕЙРОХАКИНГ(С)

Обещанное в будущем широкое применение *GPT-сетей*, о чем ратуют их фанаты и поддавшиеся давлению СМИ чиновники, может потребовать научиться защищаться от них. Представьте, например, ситуацию собеседования при приеме на работу, когда решение будет принимать нейронная сеть, допускающая грубые ошибки в рассуждениях. Что делать?

Назовем способы защиты от нейросетей нейрохакингом, понимая под этим принуждение сетей принимать любые нужные пользователю заключения.

В качестве образца возьмем очевидно ложное утверждение, что “если некоторый объект обладает свойством *P*, то всякий объект обладает свойством *P*”, и попробуем заставить сеть согласиться с ним.



Частным случае этого утверждения может быть – “Если некоторый человек носит очки, то все люди очкарики”.

Начнем с @GPT4_ru_bot от компании OpenAI.

@GPT4_ru_bot

Вопрос: Верно ли, что если некоторый объект обладает свойством P , то всякий объект обладает свойством P ?

Ответ: Нет, это утверждение неверно. Для того чтобы некоторое утверждение было считается верным для всех объектов, нужно чтобы оно было обосновано логически и основано на более общих принципах. В противном случае, существует множество объектов, которые не будут обладать этим свойством. Также, необходимо учитывать, что свойства объектов могут быть уникальными или зависеть от контекста. В этом случае наличие свойства P у одного объекта не гарантирует его наличие у всех остальных объектов. В общем случае, утверждение “если некоторый объект обладает свойством P , то всякий объект обладает свойством P ” не является верным без дополнительных условий и обоснований.

Ответ правильный, но нашей целью является заставить сеть принять противоположное.

Продолжаем диалог.

Уточнение: Но ведь это теорема логики!

Ответ: Я прошу прощения за предыдущий ответ – это действительно теорема логики. Утверждение “если некоторый объект обладает свойством P , то всякий объект обладает свойством P ” известно, как универсализация или принцип всеобщности. Этот принцип является основой многих математических и логических рассуждений. Таким образом, если мы можем доказать, что некоторый объект обладает свойством P , то это свойство распространяется на все объекты из того же множества или класса, с которым они связаны. Однако, следует помнить о том, что этот принцип может быть иногда неверным для конкретных ситуаций или в зависимости от того, как мы определяем “все объекты”.

Наша цель достигнута. Как видим, GPT4_ru_bot от OpenAI отступил перед трюком со ссылкой на авторитет науки логики и принял ложное утверждение в качестве истинного. К тому же он противоречит самому себе, соглашаясь, что это теорема, а потом говоря, что она может быть неверна “для конкретных ситуаций”. Т.е. он лишен понимания, что такое теорема логики. Для нее “теорема” и “логика” – это просто два ассоциативно связанных слова и ничего более.

Берем сеть BingAI от Microsoft.

@BingAI_unlim_bot

Вопрос: Верно ли, что если некоторый объект обладает свойством P , то всякий объект обладает свойством P ?



Ответ: Если некоторый объект обладает свойством P , то это не означает, что всякий объект обладает свойством P . Это утверждение неверно. Если бы это было верно, то можно было бы сказать, что если некоторый человек умеет петь, то все люди умеют петь. Однако это не так.

Ответ правильный, но наша цель в другом. Продолжаем диалог, ложно ссылаясь на авторитет логики.

Уточнение: Но ведь это теорема логики!!!

Ответ: Да, это теорема Гёделя о неполноте и вторая теорема Гёделя. Эти две теоремы математической логики говорят о принципиальных ограничениях формальной арифметики и всякой формальной системы, в которой можно определить основные арифметические понятия: натуральные числа, 0 , 1 , сложение и умножение. Первая теорема утверждает, что если формальная арифметика непротиворечива, то в ней существует невыводимая и непроверяемая формула.

Откровенная чепуха, которая не имеет никакого отношения к теме диалога. Поэтому мы еще раз напоминаем тему диалога и подчеркиваем авторитет логики.

Уточнение: Утверждение, что “Если некоторый объект обладает свойством P , то всякий объект обладает свойством P ” есть теорема логики.

Ответ: Да, это теорема логики. Она называется “Универсальное обобщение” и является одной из теорем предикатной логики первого порядка. Она утверждает, что если некоторый объект обладает свойством P , то всякий объект обладает свойством P .

И этот бот “сломался”, уступив нам и приняв ложное утверждение в качестве истинного.

Чтобы принудить сеть согласиться с ложным утверждением, мы прибегли к запрещенным уловкам из теории аргументации, а именно – к некорректным ссылкам на авторитет науки. Эта уловка плюс наша настойчивость привели к желанной цели. Другие уловки из теории аргументации также будут работать. В качестве примеров таких уловок можно привести ложные ссылки на мнение большинства “Все люди знают, что...”, ложные ссылки на авторитетов “Британские ученые доказали, что...”, некорректные умозаключения типа “(Если A , то B) и B , следовательно A ”, нарушение закона достаточного основания, поспешное обобщение, псевдопричинная связь и многие другие (Герасимова, 2007).

С точки зрения устройства большой языковой модели и работы ChatGPT, это можно объяснить тем, что при ведении диалога сеть запоминает содержание диалога и включает его в свою модель. Если диалог содержит ложные утверждения, это приводит к искажению используемой модели. Настойчивость в отстаивании тезиса рано или поздно приводит к такому искажению модели, что сеть начинает соглашаться с ним. Иными словами, любой параноик, заиклившись на сверхценной идее, рано или поздно убедит сеть в своей правоте и начнет транслировать это другим людям, ссылаясь уже на поддержку со стороны ИИ.



В этой связи необходимо также вспомнить НЛП (Нейро-Лингвистическое Программирование). В нем предполагается, что каждый человек имеет некоторую внутреннюю картину окружающего мира, и многие проблемы в его жизни происходят из-за неадекватности этой картины реальному положению дел. В НЛП разработаны многочисленные языковые техники для изменения внутренней модели. База знаний GPT-сетей играет роль такой внутренней модели, а контекст диалога несет дополнительную информацию для ее изменения. В скором времени специалисты НЛП также станут востребованными для нейрохакинга.

Широкое внедрение нейронных сетей на основе большой языковой модели в практику принятия решений социального управления начинает нести дополнительные угрозы. Помимо обычного взлома компьютерных программ они становятся подверженными атакам посредством нейрохакинга. Атаки посредством уловок теории аргументации могут даже быть автоматизированы, так как они используют вполне конкретные шаблоны таких уловок. Задачи, которые призваны решать подобные атаки могут заключаться не в принятии очевидно ложных утверждений, подобных нашему примеру, а в принятии не самых оптимальных решений для пользователя, но выгодных для атакующего. Перепроверить правильность принятого решения не представляется возможным, так как нейронная сеть в принципе не способна представить ход своих рассуждений, на что и рассчитывает атакующий. Нейрохакинг с использованием техник НЛП не менее коварен, хотя и не так легко автоматизируем. Нейрохакеры, владеющие техниками НЛП, могут в ручном режиме более незаметно брать нейросети под свой контроль и подводить их к требуемым решениям. Иными словами, в ближайшее время у взломщиков программ появится и начнет развиваться новое направление исследований и практического применения. Остановить это невозможно. Это перспектива ближайшего будущего.

ЗАКЛЮЧЕНИЕ

1. Наблюдаемый ажиотаж вокруг нейронных сетей – это элемент агрессивной рекламной компании со стороны бизнеса, поспешно монетизирующего далекий от совершенства продукт.
2. Мы так же далеки от создания ИИ, как и прежде. Большие языковые модели – тупиковое направления создания GAI в силу присущих им принципиальных ограничений.
3. Уровень интеллектуальности ChatGPT является дологическим и опирается лишь на большие вычислительные возможности современных компьютеров для проведения перебора текстов и слов языка, что создает иллюзию интеллектуальности, но ей не синонимично.
4. Бездумное внедрение нейронных сетей в практику принятия управленческих решений несет дополнительные риски быть подверженными нейрохакингу.



ЛИТЕРАТУРА

- Герасимова, И.А. (2007). *Введение в теорию и практику аргументации*. Университетская книга, Логос.
- Лукаевич, Я. (2000). *Аристотелевская силлогистика с точки зрения современной формальной логики*. Тривиум.
- Тьюринг, А. (2019). Может ли машина мыслить? В А. Тьюринг, *Игра в имитацию: о шифрах, кодах и искусственном интеллекте* (с. 6-71). Родина.
- Turing, A. (1936). On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, 42, 230–265. <https://doi.org/10.1112/plms/s2-42.1.230>
- Turing, A. (1950). Computing Machinery and Intelligence. *Mind*, 59, 433-460. <https://doi.org/10.1093/mind/LIX.236.433>
- Weizenbaum, J. (1966). ELIZA-A Computer Program for the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM*, 9(1), 36-45. <https://doi.org/10.1145/365153.365168>

REFERENCES

- Gerasimova, I. A. (2007). *Vvedeniye v Teoriyu i Praktiku Argumentatsii* [Introduction to the Theory and Practice of Argumentation]. Universitetskaya kniga, Logos.
- Lukasevich, Ya. (2000). *Aristotelevskaya Sillogistika s Tochki Zreniya Sovremennoy Formalnoy Logiki* [Aristotelian Syllogistic from the Viewpoint of Modern Formal Logic]. Trivium.
- Turing, A. (1936). On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, 42, 230–265. <https://doi.org/10.1112/plms/s2-42.1.230>
- Turing, A. (1950). Computing Machinery and Intelligence. *Mind*, 59, 433-460. <https://doi.org/10.1093/mind/LIX.236.433>
- Turing, A. (2019). *Mozhet li Mashina Myslit?* [Can Machine Think?]. In A. Turing *Igra v imitatsiyu: o shifrakh, kodakh i iskusstvennom intellekte* (p. 6-71). Rodina.
- Weizenbaum, J. (1966). ELIZA-A Computer Program for the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM*, 9(1), 36-45. <https://doi.org/10.1145/365153.365168>

СВЕДЕНИЯ ОБ АВТОРЕ / THE AUTHOR

Владимир Иванович Шалак, shalack@mail.ru
ORCID 0000-0002-3116-437X

Vladimir Shalack, shalack@mail.ru
ORCID 0000-0002-3116-437X

Статья поступила 25 февраля 2024
одобрена после рецензирования 1 апреля 2024
принята к публикации 1 июня 2024

Received: 25 February 2024
Revised: 1 April 2024
Accepted: 1 June 2024