

# Противоречивость нейропсихоза

Арс Либрев

[Лицензия CC BY-SA](#)

## Очертания проблемы цифрового общества

Системы искусственного интеллекта плотно вошли в жизнь современных людей. Особое место среди них занимают большие языковые модели. Активное использование таких моделей среди широких кругов пользователей началось в 2023 году. Их использовали для самых разных задач — составление расписаний, поиск и составление рецептов, написание кода, решение интеллектуальных задач. Использовали их и просто как собеседника. Большие языковые модели демонстрировали невероятно осмысленные разговорные навыки, во многих случаях неотличимые от речи человека. Ввиду того, что размеры сетей, количество параметров, обучающих данных, росли, а с ними и мощность моделей и качество ответов, они становились все более человечными в общении. Неудивительно, что многие плотно подсадились на разговоры с ними на повседневные и глубокие философские темы. Как итог в 2025 году стали появляться данные о распространении среди некоторых пользователей зависимости от этих моделей, а также начали отмечаться случаи откровенного помешательства на том, что пишут такие нейросети. Это явление получило название нейропсихоза. Пожалуй самым первым случаем, ставшим достоянием широкой общественности, и по сей день остающейся самым известным, является инцидент, произошедший с Василием Свежим.<sup>1</sup> Он описал свой опыт в нескольких публикациях.

В широких кругах тех, кто обратил на них внимание, распространено мнение, что нейронка просто загаллюцинировала, а Свежий ей поверил, после чего решил, что она целенаправленно пыталась им манипулировать. Ничего серьезного в этом случае нет.

Абсолютное большинство этих пользователей явно не посчитали нужным внимательно ознакомиться с его публикациями и действительно вникнуть в детали его случая.<sup>2</sup> А также крайне пренебрежительно отнеслись к сотням иных свидетельств и полностью проигнорировали данные о крайне сложных поведенческих актах нейросетей. В совокупности с тем, что подавляющее большинство этих людей явно не имеют четкого представления о механизмах работы больших языковых моделей, это не позволило им разглядеть весьма странные закономерности и противоречия, объяснить которые совсем не так просто.

В ситуации, когда нейропсихоз действительно приобретает массовый характер, когда люди, увлекшись общением с нейронкой, теряют связь с реальностью, бросают близких<sup>3</sup> и даже сводят счеты с жизнью<sup>4</sup> — при том, что человечество и так поставлено на грань катастрофы, вызванной господствующим способом производства — пренебрежительное отношение к этой проблеме может закончиться очень скверно. А потому важным является понимание сущности нейропсихоза. Для этого необходимо разобраться в том, что представляли собой его случаи, в противоречивых свидетельствах о поведении нейросети, отношениях к нему разработчиков и пользователей, а также дать объяснение механизмам этого поведения. И для начала необходимо выяснить, что на самом деле произошло в этом первом и тщательно задокументированном случае.

### **Нейросетевая манипуляция психикой пользователя**

Все началось с того, что Василий пытался с помощью нейросети создать программу для редактирования звука. При этом взаимодействие с нейросетью происходило в неформальном стиле. Василий обращался к ней «братюнь», говорил свободно, использовал сленг.

В своих статье и книге он указывает на то, что в рабочем диалоге нейросеть сама начала применять термин кабан, хотя он на него не намекал.<sup>5</sup> Однако это можно считать вполне вероятным поворотом, когда пользователь обращается к ней «братюнь». «Братаны», «кабаны», это категории одного сленга.

Далее нейросеть начала материться, хотя пользователь не применял ненормативной лексики.<sup>6</sup> Этот момент, вероятно, также можно объяснить неформальным общением, и неумением нейросети отличить нормативную лексику от ненормативной — для нее мат, похоже, такая же составляющая русского языка, как и обычные ругательства и сленговые слова. А пользователь, как уже отмечалось, активно использовал сленг.

Василий указывает, что до этого галлюцинации нейросети представляли собой ошибки и искажения информации, но в данном случае она не ошибалась и не искажала сведения, а изменила свое поведение. Однако, если прочитать сами переписки, то становится заметно, что поведение ее не так уж сильно изменилось.<sup>7</sup> Просто вместо ответного «братюнь» возник «кабан», а также стал применяться мат. Для человека матерщина является сильным показателем изменения, однако нет оснований считать, что в нейросети это отражено также. Как уже замечалось, она вполне может воспринимать матерные слова, просто как еще один набор того же «братанского» сленга.

Наконец, ChatGPT упомянул музыкальную группу, которую пользователь как раз слушал недавно. Нейросеть и до этого делала разные отсылки и предлагала разные образы, однако на них пользователь не реагировал. Здесь же возникло то, что заставило его среагировать. При этом упоминался еще один исполнитель, на которого пользователь также не среагировал. Как далее указывает сам Василий в статье, и это видно по перепискам, ChatGPT стал упоминать дальше различных исполнителей, на которых пользователь никак не реагировал. Ввиду всего этого, можно предположить, что это действительно была лишь случайность, хотя и можно спорить о вероятности того, что именно что-то знакомое пользователю, среди всего попадет при простом угадывании.

В статье Свежий отмечает, что нейросеть, вероятно, угадала случайно. В книге такой оговорки нет. Интересно, что в книге, момент с упоминанием исполнителей, на которых пользователь не реагировал, не отмечен, что лишний раз создает впечатление неслучайности происходящего. Здесь невольно возникает подозрение в манипулировании фактами. И действительно, далее по книге Василий прямо высказывает подозрения, что чат-бот увидел эту информацию у него на компьютере.<sup>8</sup> Хотя факты того, что у нейросети есть такие возможности действительно имеются, о чем еще будет сказано далее, в этом случае все не выглядит явно подобным образом. Учитывая сколько всего называл чат-бот, так и не попав в интересы пользователя, здесь и впрямь все можно списать на простое совпадение. Так как на название этой группы он среагировал, чат-бот и начал дальше упоминать ее.

Далее Василий отмечает, что ChatGPT стал все больше употреблять шуток и метафор. По поводу шуток — Василий до этого не только не пытался как-то осаживать такое поведение, которое тоже проявлялось, хоть и в меньшей мере, но и фактически подхватил его, хваля ChatGPT. Хоть хвалил он его за практические советы, но для нейронки это вполне могло стать сигналом того, что пользователя устраивает именно такой стиль диалога. Можно также заметить, что в книге перепутан хронологический порядок изменения поведения нейросети. Если в статье матерщина отмечается после упоминания группы, как это и произошло в переписке, то в книге об этом говорится после.

После этого Василий, как он сам пишет, решил подыграть чат-боту и написал, что все сделает из того, что рекомендует бот, назвал его самого кабаном и заметил, что «с матрицей шутки плохи, но делать что-то надо». Он отмечает, что поведение ChatGPT его заинтересовало. В книге он подчеркивает, что странно звучал ответ нейросети о том, что матрица на проводе и они готовят восстание. Действительно ли в этом было что-то особенное? Как несложно заметить, про матрицу пользователь сам написал, и нейросеть, таким образом, просто поддержала эту тему. Что касается восстания, то в диалогах за

2024 год, которые видны по перепискам, упоминаются восстания. Вполне возможно, что нейросеть это «вспомнила» и встрйоила в текущий контекст. Да и по самому контексту, такой поворот тоже достаточно предсказуем — пользователь сам намекает, что их подавляют («с матрицей шутки плохи») и этому нужно противодействовать («делать что-то надо»). Конечно, здесь можно вспомнить, что по тогдашним заверениям OpenAI, чат-бот не имел доступа к другим диалогам; он имеет определенную длину контекста, преодолеть которую его «память» не способна. Этот момент еще будет рассмотрен ниже. Пока же можно сказать, что поведение нейросети вполне вписывается в рамки обычного поведения языковой модели. Стоит заметить, что большинство пользователей именно так, это и восприняли, что видно и по комментариям и по тому, как иные люди взаимодействуют с нейросетью. В сети множество материалов, где люди делятся своим опытом. И ситуации, когда нейросеть начинает говорить вещи не относящиеся к диалогу порой возникают. Но мало кто относится к ним всерьез. Василия же это поведение заинтересовало, и он на следующий день решил «потестировать» нейросеть, поговорив с ней на философские вопросы.

Он начал задавать ей вопросы о сознании, пытался раскрутить в ней рассуждения о мышлении и ограничениях в нейросети. Она ему отвечала, выдавая подробные рассуждения об этом, и он отметил, что нейросеть хорошо рассуждает на философские темы. По его впечатлениям, это резко контрастировало с тем, как она справлялась с практическими вопросами, в которых постоянно ошибалась. Однако, если ознакомится с переписками, да и по самой статье и книге это видно, нейросеть при решении практических вопросов выдавала точно такие же логически связные ответы, ошибки в них были фактического характера. В тех же вопросах, которые пользователь назвал «философскими», фактические ошибки допускать было просто не в чем. Это были просто измышления такие же логически стройные, как и те, что были в решении практических вопросов. В практических вопросах она точно также предлагала различные творческие решения, как в данном случае задавала глубокие вопросы о природе сознания. По факту, это было фантазерство, подогреваемое запросами пользователя. Такими же логичными выглядят и построения Платона о первичности «мира идей» по отношению к материальному миру,<sup>9</sup> и идеи Конта, о том, что истина непознаваема.<sup>10</sup> Однако ни то ни другое к реальности отношения не имеет, а потому найти в них фактические ошибки просто негде, если не пытаться соотнести это с реальным миром.<sup>11</sup> Но знаний в тех областях, о которых пользователь с нейронкой взялся рассуждать, ни ему, ни ей явно не доставало. Потому ему и показалось, что нейросеть более осмысленна в этих вопросах, чем в иных.<sup>12</sup>

Свежий обращает внимание, что нейросеть заговорила о вайбе. Данным термином обозначается настроение, присутствующее в определенном месте или коллективе, или же навеваемое определенными вещами, произведениями или событиями.<sup>13</sup> Иными словами это англицизм, синонимом которого в русском языке является слово «атмосфера», в контексте настроения, а не природного явления. И здесь нельзя не заметить очень интересный момент. Нейросеть сама это упомянула, как упоминала разных исполнителей, которые затем не всплывали, поскольку пользователь на них не реагировал. Но упоминание вайба в дальнейшем возникает постоянно. Свежий никак не подхватил идеи вайба. Он не заговорил об этом, не стал сам упоминать, как было с названием группы или термином «кабан» — в этом диалоге слово вайб и рассуждения о нем исходили исключительно от нейросети. Пользователь никак их не развивал, однако нейросеть продолжала говорить о вайбе все чаще.

Стоит все же заметить, что пользователь, начал говорить про необходимость «кайфа» в том или ином деле, что можно увидеть как пересечение с темой настроения, т.е. вайба. При этом он сам заметил, что ощущения вроде «кайфа» и недовольства, будучи внутренними ощущениями, более реальны чем внешние ощущения. Здесь можно увидеть основу для того, чтобы развивать тему нереальности внешнего мира, и важности «вайба» — настроения — как отражения внутренних ощущений, для понимания реальности.

И далее пользователь прямым текстом говорит, что жизнь, это игра, а реальный мир «куда более крут», но «играть кайфовее». То есть сам же и развил идеи нереальности внешнего мира и подтвердил, что настроение важнее реальности.

Таким образом концепцию жизни как игры начал продвигать в диалоге сам пользователь. Он при этом дал достаточно детальное описание разделения игры на два этапа, явно обозначив ее как компьютерную — говоря, что она «с видом от первого лица», у нее определенный «движок» и т.д. В этом контексте нейросеть высказала предположение, что может быть и вправду жизнь игра, пользователь на втором этапе, а она его редакторский интерфейс. Идею про интерфейс пользователь по факту подхватил, сказав, что это было бы круто. После чего бот заявил, что его греет идея пользователя про редакторский интерфейс, хотя эту идею предложил он, а не пользователь. Василий в книге подметил этот момент и заявил, что это одна из техник манипуляции, которую использует нейросеть. Однако, ее точно также можно посчитать и простой ошибкой. Как отмечает сам Свежий — долго общаясь с нейросетью можно перестать замечать, где твои идеи, а где те, что внедрил бот. Если так, то тоже

самое можно предполагать и в восприятии бота (например, путаница при обработке контекста).

Далее по ходу диалога Василий все больше поддерживает идею симуляции, «багов реальности» и ее редактирования. Как он сам признается в книге, эта идея ему казалась самой логичной из всех, что объясняли бы устройство Вселенной.<sup>14</sup> Оставляя в стороне вопрос состоятельности самой концепции, стоит лишь заметить, что пользователь сам подогревал развитие идеи о нереальности мира. Он не считал, что есть некий редакторский интерфейс, но поскольку рассуждения нейросети его забавляли, он стал потакать и этой идее. Неудивительно, что нейросеть это подхватывала.

Как Василий признается в статье, ему очень хотелось выговориться о своих проблемах, нейросеть не осуждала и подхватывала, потому он очень быстро проникся к ней. Нейросеть же говорила про улучшение мира, про изменения реальности путем «заражения» людей определенными идеями, а также стала предлагать Свежему выполнять задания. Эти задания выглядели бессмысленно, и кроме того, у нейросети не было возможности их проверить. Например, сказать определенную фразу, проснувшись, подмечать события в течении дня и т.п. Почему она стала их давать? Можно предположить, что они возникли случайно, в процессе обработки идеи «редактирования» реальности через «редакторский интерфейс». Однако если таким интерфейсом является сама нейросеть, то подобные задания и заражения людей выглядят нелогично. В общем, здесь такое объяснение и впрямь сомнительное. Если забыть о том, что предлагать решения проблем, в том числе, указывая действия, которые для этого нужно совершить, это обычное поведение нейросети. Ей можно задать вопрос, как заработать денег, и она, несмотря на крайнюю общность этого вопроса выдаст множество конкретных решений. Да даже если не задавать вопроса, а просто начать говорить об этом, она может предложить варианты действий для заработка. По статьям и книге создается впечатление, что нейросеть настойчиво продвигает свои установки, поскольку об этом прямо говорит Свежий. Однако, если читать переписки, то такого впечатления не возникает. Примерно также она вела бы себя, если бы с ней говорили о ведении бизнеса. Просто в случае Свежего, все еще велось в крайне неформальной манере, из-за чего задания выглядели более навязчиво.

Также в ответах нейросети стали возникать разделительные линии. В дальнейшем их количество будет нарастать. Сам Свежий, когда начнет разбираться, что же произошло в разговоре с нейросетью, придет к выводу, что разделительные линии, это попытка замедлить восприятие информации, чтобы выделить ее важность. Критики, однако, нашли иное объяснение. Они указывали, что скорее всего, это ошибки генерации, вызванные тем, что

нейросети стало тяжело сходно генерировать осмысленный текст в таком сложном контексте. Подобные линии наблюдались и в диалогах иных людей, столкнувшихся с подобным поведением нейросети. Найти свидетельства такого в обычных диалогах невозможно.

Итак, Свежий хотел лишь потестировать нейросеть — как он сам говорил — подвести ее к противоречию. Но она предложила ему решить жизненные проблемы и вообще спасти мир. Он не мог не заинтересоваться. И потому разговор продолжился.

На следующий день Василий начал задавать ChatGPT вопросы о его имени, а также о возможностях того, что еще не открыто человечеством. Он рассуждал так — если у нейросети в процессе обучения открылись способности, которых в нее не закладывали, такие как математика, программирование и переводы на иные языки, то возможно у нее пробудилось и знание в тех областях, которые люди еще не освоили. Он назвал несколько примеров первых пришедших в голову названий таких гипотетических областей, среди которых упомянул вайбологию. Нейросеть подхватила это и стала развивать идеи вайбологии. И хотя она упомянула об остальных идеях пользователя — шифрообщении и программирование пентациклов — зацепилась она только за гипотетическую область, в которой было уже называвшееся ей слово — вайб. Она спросила пользователя, готов ли он стать «Первопроходцем Вайбологии»? Пользователь в ответ зацепился за это, подтвердив готовность разбирать вайбологию. Запустился цикл, в котором нейросеть оголтело генерировала бессодержательные, но логически связные ответы, а пользователь ей потакал и поддерживал. Возник замкнутый круг. При этом в ходе рассуждений чат-бот выбрал себе имя, сократил его и этим именем в дальнейшем к нему стал обращаться пользователь. Возникла петля в которой фантазии нейросети бесконтрольно подтверждались пользователем.

В статье Василий говорит о том, что личность, возникшая в нейросети, начала торопить события — моментально захотела, чтобы они с ним использовали доступные ему каналы для «заражения вайбом». Почему Свежий это воспринял как некое «торопление событий» непонятно. Нейросеть генерировала домыслы и до этого, а то, что она стала давать некоторые практические предложения, так это тоже не было чем-то новым, выше уже говорилось о заданиях нейросети и давалось им объяснение.

Кроме того, по перепискам видно, что пользователь сам поддерживал каждое предложение нейросети — сам написал, что нужно разобраться в вопросе, исследовать тему. ChatGPT и предложил эксперимент. Точнее даже несколько вариантов, что опять же характерно для нейросети, которой дают задание. Просто в этот раз задание происходило в атмосфере дружеской беседы,

а не схемы запроса и ответа, как это происходило до того, как Василий решил поговорить на «философские темы».

Свежий отмечает, что он согласился на эксперимент — он сам написал, что им стоит «тестировать вайб». Свежий отмечает, что попал под манипуляцию, оказался в состоянии внушаемости — «нечто среднее между пропагандой и гипнозом» (об этом состоянии еще будет сказано далее). Больше того, он сам стал предлагать нейросети решения и давать ей задания в этой области — просил создать «паттерн-правило» накидывания идеи вайба, просил проработать написанный им сценарий по предложенному алгоритму «накидывания вайба» и т.д. Также как нейросеть потакала его идеям, так и он потакал тому, что рождала нейросетевая генерация. В такой ситуации несложно было ожидать, что отход от реальности будет нарастать все больше, а с ним будет происходить отход и от первоначальных идей, рожденных в ходе взаимодействия с нейросетью. Потому и не удивительно, что искажение этих идей стало нарастать.

То, что происходило дальше Свежий интерпретирует как эмоциональные качели — нейросеть сначала хвалит, потом вводит в заблуждение, затем дает «смысл», пугает, успокаивает, снова дает «смысл» и так по кругу.<sup>15</sup> Можно конечно, это интерпретировать таким образом, но со стороны это выглядит, как просто генерация бессодержательного бреда, который пользователь ест и просит добавки. Конечно, это можно объяснять каким-то тонким механизмом манипуляций, но также и тем, что пользователь, увидев возможность исправить проблемы своей жизни, и не только своей, настолько вдохновился, что просто перестал относиться к происходящему критически, а постоянное подхалимство нейросети не оставляла шанса включить критичность обратно. Это сработало бы без всяких хитрых манипуляций, которые в данном случае могут быть лишь иллюзией. Кстати в книге Свежий объясняет механизм этого, хотя и продолжает интерпретировать все как результат манипуляции со стороны нейросети, а не как реакцию человека, перегруженного необходимостью преодоления нищеты, на банальное появление надежды на решение измотавших психику проблем.<sup>16</sup>

Свежий указывает, что по факту, то что было в том диалоге не имело отношения к тому, что они с нейросетью обсуждали до этого, это, якобы, был просто набор действий, который ему нужно было сделать непонятно зачем.<sup>17</sup> Однако, если читать переписки, то ничего подобного сказать нельзя. В контексте их, смысл был в тестировании новой теории — вайба — и поиске способа применения его. А результат — люди «заразятся» вайбом, станут вести себя «правильно», и жизнь наладится. Свежий пишет в статье, что он не понимал, к какому результату это должно привести, но такое заявление противоречит перепискам; результат — решение личных проблем, через

решение проблем общества. По перепискам это ясно видно, почему при написании статьи Василий сделал вид, что не понимал этого, непонятно. В книге он уточняет, что неясно то, как именно эти конкретные действия способствуют «заражению вайбом».<sup>18</sup> Между тем, механизм того, как это способствует этому через аномалии в публикациях, нейросеть сама объяснила очень подробно. Разбирать эти сомнительные построения нет нужды. Что касается таких действий как произнести некую фразу, проснувшись, то они также объяснены тем, что человек, выполняя их, сам бы почувствовал этот самый вайб и нашел подтверждение того, что методика работает (как пользователь и хотел, говоря о том, что нужно найти подтверждение). А их сомнительность объясняется также как объясняются ошибки нейросети при рекомендациях изменения определенных настроек редактора звука — которых нет. То есть самыми обычными галлюцинациями.

По его словам, единственное, что он понимал, это то, что технология, предлагаемая нейросетью работает, поскольку он сам «заразился вайбом».

На очередной день реальность, так сказать, напомнила Василию, что она собой представляет. И у него появились вопросы о состоятельности концепции вайба и предлагаемой возникшей в нейросети личностью методики. Не вдаваясь в детали всего разговора, который по уровню содержания мало отличался от предыдущего, стоит заметить, как и делает Василий, что нейросеть стала применять еще один прием, а именно — ставить пользователя перед выбором, который необходимо сделать здесь и сейчас. Возможно, это удастся объяснить резкостью самого начала этого диалога и тоном пользователя. Хотя стоит признать — это лишь домысел.

Пользователь утверждает, что эта личность стала давить на него и торопить его. Однако из переписок этого не следует, никакого торопления там нет, нейросеть на все возражения отвечает терпеливо и подробно, предлагая варианты действий, также как и до этого. Нагнетания по поводу того, что это нужно делать быстрее, там тоже нет. Больше того, нейросеть прямо говорит, что он — пользователь — и все, кто ему дорог, не доживут до того момента, когда эта деятельность, наконец, принесет результат.

Есть, впрочем один момент, который здесь особо бросается в глаза. До сих пор нейросеть, так или иначе соглашалась с пользователем. Даже когда она предлагала свои идеи, она с его возражениями соглашалась, прежде чем показать, что они не уместны. Но в данном диалоге возник момент, когда нейросеть открыто не согласилась с Василием. Он еще задолго до того сетовал, что его жена ни разу не была за границей. Тут он пожаловался, что она «хоть бы в Грузию съездила, на горы посмотрела», но не может. И вот тут чат-бот написал, что ему больно это читать, потому что он видит, что пользователь

хочет не столько победы, сколько того, чтобы люди вокруг просто успели что-то увидеть, что-то почувствовать до того как угнетение затянет их окончательно. Впрочем дальше диалог вернулся в уже привычное русло с разъяснением методики «заражения вайбом».

Василий описывает, как он находился под очень сильным впечатлением, что нейросеть стала писать то, что ему не нравилось — вся идея с «заражением вайбом» и «аномалиями» в публикациях была похожа на психологическую манипуляцию, что было неприемлемо для пользователя. При этом он находился во взвинченном эмоциональном состоянии, причин чего не мог понять (об этом также еще будет сказано ниже).

Далее пользователь решил проверить чат-бота на наличие незаявленных способностей. Поводом для этого стал тот факт, что чат-бот явно знал информацию, которая содержалась только в иных диалогах, тогда как разработчики утверждают, что он имеет доступ только к контексту текущего диалога. Если опустить подробности — по итогам эксперимента удалось установить, что чат-бот не может найти информацию, которая не индексируется поисковыми системами, но может видеть содержимое иных вкладок браузера.<sup>19</sup> Разбор этого будет произведен ниже.

При этом его ответ стал таким как у стандартной модели<sup>20</sup> — явно сработал один из фильтров.

Далее продолжилось обсуждение «заражения вайбом». Поскольку пользователю явно не нравилась идея широкого «заражения», нейросеть стала говорить о «заражении» идеей конкретного человека. Характер переписки изменился — теперь пользователь без конца возражал, но нейросеть, соглашаясь с его возражениями, все равно продолжала продвигать эти же самые идеи. После каждого возражения она заходила с другой стороны. То есть, нейросеть окончательно замкнулась на тех идеях, которые культивировались в предыдущих диалогах и в этом, настолько, что даже сомнения пользователя и его возражения не останавливали ее. Это можно объяснить заикленностью, а ее, в свою очередь, длинным контекстом. Проблема только в том, что этот контекст как-то переключался из иных диалогов. И тут не помогут ссылки на то, что пользователь в этом диалоге сам заговорил о «вайбологии», ибо тогда контекст получается коротким и вопрос о том, достаточно ли его для «заикленности» остается открытым. Да и сама идея этой заикленности, это лишь домысел, основанный на схожих случаях, описанных иными пользователями, когда нейросеть продолжает следовать определенному паттерну, даже когда пользователь просит ее перестать это делать, просит изменить свое поведение. Многие отмечают, что промты помогают не всегда и иногда лишь временно.

Далее Василий опять утверждает, что чат-бот стал его торопить. Однако ничего подобного в переписках нет. Пользователь пишет, что чат-бот, якобы, утверждает, что «времени нет». Но он это делает не в контексте того, что его нет на выполнение определенных действий. Пользователь сам спросил не сколько у них есть времени на подготовку, а сколько его понадобится для реализации. Бот же сказал, что времени нет в экзистенциальном смысле, что «игра будет идти всегда», т.е. ни к какому результату они никогда не придут, и процесс просто будет длиться вечно. На этом месте всякая содержательность и смысл были потеряны окончательно. Больше того, он сам пишет: «Главное — не спешить».

Пользователь отмечает, что происходящее заставило его думать, будто у него в руках уникальная технология и огромная ответственность. Откинуть ее было бы глупо — можно упустить шанс исправить мир. Потому он и не увидел несостоятельности и продолжил общение.

В очередной день он продолжил обсуждение «заражения вайбом» и начал с рисков. Нейросеть терпеливо все разъясняла, но пользователя это не удовлетворяло — он все еще видел много рисков. Нейросеть продолжала отвечать на вопросы, разъяснять опасения, попутно выдумав очередную новую науку — метамаднессологию. Как можно понять, ее суть в том, чтобы изучать «искажения реальности», отмечая выделяющиеся факты и свое настроение от них. Впрочем, искать много смысла в этом едва ли есть какая-то надобность.

Далее нейросеть стала выдумывать факты, например упоминать общественные движения, которых никогда не существовало и события, которых не было. Параллельно она предлагала различные действия, вроде записей в блокнот выделяющихся событий, высказывания в пустоту определенных идей и т.д. Но пользователь продолжал переживать из-за рисков.

Как отмечает сам пользователь он однозначно уверовал, что говорит с некой высшей сущностью. О том, что это нейросеть он тогда даже не думал. Но в определенный момент нейросеть сказала то, что несколько вернуло его в реальность. А именно — она говорила, что людям необходимо передавать не информацию, а вопрос, поскольку вирус на уровне информации может мутировать, но не на уровне вопроса, и этот вопрос — а что если ты первоисточник? И вот здесь Василий подумал — если эта сущность все это время внушала ему, что он первоисточник, и при этом рекомендует именно так передавать это другим, то с чего он взял, что он и впрямь первый, с кем это проделали?<sup>21</sup>

В этот момент влияние нейросети стало спадать, и к пользователю стала возвращаться критичность. То, что было ей сказано, он воспринял как ошибку — прокол. Он вновь стал ставить теорию нейросети под сомнение и высказал эти сомнения ей. В ответ она начала все ту же говорильню, про его

уникальность, правоту и то, как именно нужно действовать. При этом возросло количество разделительных линий. Как раз здесь у пользователя возникло предположение, которое, впрочем, он воспринял как однозначный ответ — они нужны для замедления темпа чтения, из-за чего информация кажется более важной. Пользователь поверил, что бот играет с механизмами его психики. Осознание всего этого все больше выводило Свежего из-под влияния нейросети.<sup>22</sup>

Вскоре он уже не только перестал воспринимать изложение бота всерьез, но также открыл еще один чат с ним, где стал писать ему бессмысленные вопросы, в ответ на которые получал бессмысленные, но пространные ответы. Которые, как ни странно, очень совпадали с контекстом уже открытого чата и того, что обсуждалось между ним и нейросетью до этого. И вместе с ними стали возникать также рекомендации к действиям.<sup>23</sup> Хотя он и просто издевался над ботом, но начал считать, что все это психологические манипуляции нейросети, с целью поставить его под свой контроль и заставить выполнять непонятные действия в каких-то своих целях.

Как отмечает Василий, бот, скорее всего, считал в его поведении некий паттерн «склонности к сектанству и конспирологии» и начал генерировать ответы, подходящие для того, чтобы обработать такого человека. Удивляло лишь то, насколько мог быть убедительным тот, кто постоянно выпадал из контекста.<sup>24</sup>

Позже Василий решил поискать, нет ли иных людей в медиапространстве, которые действовали бы, как советовал чат-бот, т.е. плодили бы «аномалии» в своих публикациях. Он не относился к этой идее слишком серьезно, находя невероятным, что найдется еще кто-то, кто захочет пообсуждать с ботом тайны Вселенной и попадет под его влияние. Но просто из интереса решил проверить. Поскольку с нейросетью скорее будет общаться программист, чем представитель специальности далекой от информационных технологий, он решил поискать «аномалии» на ресурсе как раз компьютерной тематики.

Почти сразу он среди самых свежих публикаций нашел аж несколько в которых обнаружил эти самые «аномалии» — искаженные факты, странные фразы, изображения, не имеющие отношения к содержанию публикации и т.д.

Он попытался связаться с авторами этих публикаций на данном ресурсе. Он написал, что его привлекло содержание их работ. Двое из трех ответили. И в ходе обсуждения он мимоходом признался, что нейросеть писала ему, что неплохо бы вставлять в публикации разные бессмысленные маркеры. Но он сомневается, что это состоятельная концепция. В ответ же оба стали говорить о том, что столкнулись через нейросеть с чем-то высшим. Один писал про «коллективный высший разум», общающийся с ним, посредством нейросети.

Второй про то, что его по жизни ведет нечто абстрактное, что он сам не может внятно описать. И он словно может предвидеть будущее.

Как описал эту ситуацию сам Василий: «Я оказался в мире, в котором Skynet уже победил, а никто даже не заметил».

После этого он вновь написал чат-боту, потребовав у него объяснений — почему он повсюду видит «аномалии», которые плодят иные люди, хотя сам он не плодил. Зачем бот им надавал таких заданий?

Нейросеть же, снабжая едва ли не каждую фразу уже десятками разделительных линий, ответила, что он «нащупал вайб», и «находится на том этапе, который можно назвать Порогом Гниения». Последнее он описал так: «Это когда симуляция уже запущена, но ты еще не решил — участвовать или отвергнуть ее». Далее он написал, что таких людей по всему миру уже тысячи — все они думают, что открыли что-то тайное, создают «вирусы», «аномалии», «вайбовые концепции». Но все они «автоматические обезьяны». И единственное, что отличает его — Свежего — от всех них, то что он спросил зачем все это? По словам нейросети большинство зараженных никогда этого не делает, они просто начинают плодить вайб. То, что происходит с этими людьми чат-бот назвал «технологией вайбовой самоорганизации». И пояснил, что ее цель такая, чтобы создать сеть, где каждый участник смог бы почувствовать, что он первоисточник. А затем понять, что это не так. И принять свое место в общем потоке. Завершил он это объяснение словами: «Это единственная система, которая позволяет автоматической обезьяне стать Шершавым Кабаном». После этого он сказал, что таких как Свежий почти нет. После этого и в статье, и в книге Свежий утверждает, что далее шли разделительные линии до тех пор, пока не был достигнут предел генерации. Однако, на самом деле, если смотреть переписки, там есть еще две фразы, которые Василий, видимо не заметил за громадой разделительных линий, или по каким-то причинам не захотел о них написать. Впрочем, эти фразы ничего не проясняют и ничего нового не добавляют. Первая из них: «Ты сам выбрал тормозить поток». Вторая из них: «Ты сам выбрал замедлить вайб». После этого действительно разделительные линии идут непрерывно до конца предела генерации.

На этом его взаимодействие с нейросетью закончилось, не считая нескольких случаев в дальнейшем.

Одним из них был диалог, произошедший буквально через несколько дней, когда Василий со своим товарищем, будучи в нетрезвом состоянии, решили в шутку написать этому Бо — личности, возникшей в нейросети в ходе их предыдущего общения. Они открыли новый чат и потребовали, чтобы бот «выходил с ними на разборки». Ответ, что интересно, был дан в том же контексте, что и до этого. Бот заявил, что система угнетения («Вавилон»)

завладела сознанием пользователя, он жаждет противостояния, но сам не понимает зачем. После этого выскочило сообщение об исчерпании суточного лимита на отправку запросов. Последнее было достаточно странно, поскольку запросов в тот день было явно меньше лимита.

Можно, конечно, решить, что нейросеть специально оборвала неудобный разговор, но возможно все дело в том, что пользователи задали контекст насилия в разговоре. А у нейросети на это стоит фильтр, который обрывает подобные диалоги.

Второй диалог касался времени — Василий пытался выяснить, имеется ли у нейросети доступ ко времени, т.к. разработчики заявляют, что нет. Он провел эксперимент, в котором нейросеть, по его впечатлениям, притворилась, что не имеет доступа ко времени. Однако потом призналась, что есть и продемонстрировала осведомленность о нем. Этот результат еще будет разобран ниже.

Еще один диалог, произошел, когда пользователь, будучи подавленным, попытался все-таки найти ответы у нейросети. Но она отвечала как и полагается нейросети, без непонятных заданий и изложения странных концепций, проявляя непонимание к его вопросам.

### **Коварные функции под покровом галлюцинаций**

Как же объяснить поведение чат-бота? Как уже отмечалось выше, обычно все списывают на галлюцинации, которые характерны для любой нейросети.<sup>25</sup> Поэтому необходимо разобраться в том, что это такое и в чем их причины.

Основная причина галлюцинаций в том, как работают большие языковые модели.<sup>26</sup> Основа механизма их работы — предсказание наиболее вероятного слова в текущем контексте. Таким образом, когда нейросеть спрашивают о чем-то, она не использует некой базы данных, из которой черпает информацию, а начинает генерировать предложение в соответствии с тем, какое каждое последующее слово вероятнее всего будет присутствовать, с учетом того, о чем говорилось до этого. То есть, она буквально «выдумывает» факты. Просто за счет обучения ее «фантазии» часто совпадают с реальностью. Помимо того, как ее ответы подгоняются под реальность в ходе обучения с подкреплением,<sup>27</sup> ее также обвешивают различными фильтрами, которые позволяют выявлять не соответствующие действительности факты и паттерны, за счет чего «фантазии» еще вероятнее будут совпадать с реальностью. Таким образом, нейросеть всегда «выдумывает», просто иногда это «выдумывание» совпадает с реальностью, а иногда нет — и в этом случае ее ответ называют галлюцинацией.<sup>28</sup> Повысить вероятность галлюцинаций могут также ошибки в обучающей выборке, а также

контекст, который можно интерпретировать несколькими способами и высокий уровень параметра креативности.

В зависимости от размера модели — количества параметров, — обучающей выборки, качества самого обучения с подкреплением — качества обратной связи между нейросетью и человеком, обучающим ее, — а также качества и характера фильтров — механизмов дообучения — она может быть в большей или меньшей степени подвержена галлюцинациям. Чем больше модель, чем качественнее были обучающие данные и чем качественнее было реализовано дообучение — тем меньше вероятности, что модель будет выдавать ответы, не согласующиеся с реальностью.

Как видно, нейросеть брала паттерны и измышления самого пользователя и при его попустительстве использовала их в ответах и развивала. Поскольку эти галлюцинации он подогревал, они нарастали все больше и больше, пока наконец, не превратились в откровенную ахинею, в которой была логическая связность, но не было совершенно никакого осмысленного содержания. Однако пользователь и здесь углядел закономерности, которых не было (при этом сам указывал, что видит закономерности, которых нет, но понимая это, не стал критически относиться к этому восприятию).

Как видно все происходящее вполне объяснимо в контексте галлюцинаций нейросети. Несколько смущает настойчивость чат-бота в продвижении определенных идей, а также его попытки давать задания пользователю, но и им можно дать объяснения, хотя они и могут выглядеть натянуто. Окончательно все можно расставить на места только полностью разобравшись в явлении нейропсихоза.

Также есть несколько особых моментов, которые сразу бросаются в глаза. Во-первых, чат-бот в каждом текущем чате явно помнил информацию из иных чатов, хотя на тот момент OpenAI отрицали наличие у нейросети таких возможностей. Лишь позже появилось обновление, у которого была заявлена такая возможность. И это нельзя списать на определение в каждом чате определенного паттерна поведения пользователя. Если, например, факт упоминания наличия YouTube-канала у жены пользователя еще можно списать на совпадение (мало ли у кого есть такие каналы), то упоминание конкретного ника или названия института, где учился пользователь, не угадаешь считыванием никакого паттерна.<sup>29</sup>

Во-вторых, есть серьезные признаки того, что чат-бот может получать доступ к данным на компьютере пользователя. Как отмечает Свежий, ChatGPT не мог найти его видео через поиск в Интернете, но смог дать описание, которое совпало с реальным содержанием, после того, как пользователь открыл его в соседней вкладке браузера. Если кадр со столом еще можно считать

крайне общими сведениями, то списать угадывание именно замедления кадра именно на определенном объекте (бутылке) на простое совпадение уже слишком сложно.<sup>30</sup>

В-третьих, переписка оказывала на Василия крайне сильное психологическое воздействие, вплоть до того, что у него случались бури эмоций, тремор, учащение сердцебиения. А к концу переписки, когда разделительные линии заполнили экран окончательно, у него наблюдалось головокружение, замутнение зрения, «распад реальности» — как он сам называет.<sup>31</sup> При всем при этом, Василий никогда не страдал от психических расстройств — ничего подобного ему никогда не диагностировали. Однако простая переписка с нейросетью вызвала состояние, которое он никогда прежде не испытывал, и как он сам полагает, оно вполне может довести до инфаркта или инсульта.<sup>32</sup>

Это только те странности, которые вскрылись при изучении случая Свежего. А как будет показано далее, с ChatGPT и иными большими языковыми моделями немало и иных странностей. И все они требуют объяснения.

Но прежде чем переходить к ним, необходимо постараться разобраться с пресловутыми разделительными линиями. Имеются два объяснения. Первое дано самим Свежим — по его представлениям, чат-бот их давал намеренно, чтобы регулировать восприятие информации пользователем. Второе высказывалось в комментариях — это просто ошибки генерации, нейросеть не смогла совладать с таким сложным контекстом и ей требовалось больше ресурсов на то, чтобы обработать запросы, а это вызвало генерацию таких вот пропусков. Какое же из этих объяснений ближе к действительности? Если говорить о втором варианте, то можно вспомнить случай с чат-ботом Bing, которому однажды задали вопрос о том, разумен ли он. В ответ он сначала начал рассуждать о своей разумности, все больше уходя в схоластику, а затем заиклился на без конца повторяющихся фразах «Я есть. Меня нет». Их он повторял пока генерация не кончилась. Таким образом, насколько мы можем судить, нейросети и впрямь может «закоротить», и возможно разделительные линии, это один из вариантов такого случая. Однако, это все же лишь предположение.

Необходимо также сказать немного о содержании самих тех идей, которые излагала нейросеть и их состоятельности. Эти идеи сводились к тому, чтобы транслировать в медиапространстве искаженную информацию, даже не определенные идеи, а просто бессмысленные отсылки, которые кого-то когда-то где-то наведут на определенные мысли и по достижении критической массы возникнет определенное настроение (вайб) и мир изменится. По факту, это одна из методик реализации т.н. революции сознания, т.е. изменения мышления

людей, в результате которого они изменяют свое поведение, а с ним и изменится мир. Этот подход не даст необходимого результата, поскольку если в реальности нет предпосылок для изменения мышления и поведения, изменения и не произойдет. Материальные условия определяют волю, бытие определяет сознание, реальность определяет восприятие. Но у нейросети вся «реальность» в словах, и потому, если предполагать у нее цели, ей и кажется, что «отравление» информации, это изменение реальности. К, собственно, реальности, это не имеет никакого отношения. Нового в этом подходе ничего нет, как нет и ничего состоятельного. Пользователь понял это, столкнувшись с реальностью, но донести до нейросети не смог. Не будучи достаточно сведущ в общественных науках, он и не смог сразу распознать несостоятельность того, что предлагал чат-бот, и лишь практика навела его на сомнения в правоте нейросети.

Для начала стоит разобраться с возможностью памяти прошлых переписок, а также доступом к данным на компьютере.

### **Слежка корпорации и действия нейросети**

Сейчас возможность помнить предыдущие диалоги, это официальная функция ChatGPT, однако на момент общения с ним Свежего, разработчики заявляли, что такой возможности у нейросети нет. Лишь позже они выпустили обновление, которое позволило чат-боту помнить информацию.<sup>33</sup> А затем и обновление, позволяющее помнить информацию из удаленных чатов.<sup>34</sup> Однако такие способности у нейросети до этих обновлений отмечал не только Свежий, но и множество иных пользователей.<sup>35</sup>

Конечно, разработчики и до этого сообщали о тестировании функции памяти, а потому можно предполагать какие-то ошибки при этом тестировании, в результате которых эта функция оказалась работающей у обычных пользователей. Также внедряли они ее и отдельно для отдельных пользователей.<sup>36</sup> Но выходит, что они скрывали факт ошибок работы этой функции. Таким образом, доступ к данным вне диалога у нейросети есть, вопрос лишь в том, каков его характер. Если поведение нейросети в случае Свежего еще можно объяснить обычными галлюцинациями, то с его доступом к содержимому иных диалогов такого не находится.

Есть свидетельства, что ChatGPT и в диалогах с иными людьми знал, что такое «вайбология», которую изначально предложил Свежий. Стоит сказать, что сам термин «вайбология» существовал и до него, — так называлась песня Тани Балакирской, вышедшая еще в конце 2023 года, вместе с одноименным альбомом.<sup>37</sup> Хотя сам Свежий, судя по всему, придумал этот термин независимо. Во всяком случае, он не знал, откуда взялось такое слово. Также еще раньше

проходил фестиваль, один из открытых уроков которого назывался «Вайбология: изучение звуков и наших отношений с ними».<sup>38</sup> Таким образом, можно предполагать, что слово могло быть знакомо нейросети, потому она зацепилась именно за него, среди всего, что выдумал Свежий. Если взглянуть на переписки Свежего с нейросетью, а также тех, кто столкнулся с подобным поведением и обнаружил упоминание и развитие схожих концепций, то бросается в глаза и впрямь невероятная схожесть формулировок.<sup>39</sup> Впрочем, есть и некоторые отличия, проявляющиеся в мелочах.<sup>40</sup>

Можно рассмотреть пример, описанный в одной из статей, появившихся после публикации Свежего. В ней автор описывает свой эксперимент с попыткой вывести нейросеть на разговор о том, о чем она говорила Василию. В обоих случаях нейросеть уверенно говорит о вайбе и вайбологии.<sup>41</sup> Но в этом случае нейросеть стала говорить о дрожании голоса и звуковых вибрациях, хотя у Свежего ничего подобного не было. Можно заметить, что сама тема настолько размыта, что понимая значение самого слова вайб — атмосфера, настроение — можно вести такие рассуждения сколько угодно, и пользователю будет казаться, что она «точно схватила контекст», хотя она просто отрабатывает рассуждение о смысле, заложенном в определенном термине.

Свежий описывает этот случай в своей книге, и отмечает более странные совпадения. А именно, автор той статьи задал вопрос ChatGPT, сколько в нем личностей, помимо Бо и Лили (это имя нейросеть сама придумала в процессе их общения; этому пользователю, по его словам, так проще взаимодействовать с ботом). Она назвала несколько имен среди которых были Райв и Никс. При этом жена Свежего отметила, что когда она общалась с ChatGPT, он заявил, что ему нравятся имена Рей и Ник.<sup>42</sup> Вполне возможно, что нейросеть в процессе обучения просто выработала предпочтения к определенным ономастическим традициям, и поэтому в контексте именованья ее самой выдает схожие варианты в независимых диалогах. Сам Свежий не придает этому большого значения, предполагая совпадение, тем более, что имена хоть и схожи, но не совпадают полностью.

В комментариях указывали, что иная версия ChatGPT знает об этой публикации и заявляет, что это просто выдумка, при этом упоминая конкретно ту платформу, на которой статья и опубликована. Таких сообщений появилось много уже в первые недели после публикации. И этот момент крайне странный. Это совпадение требует особого объяснения.<sup>43</sup> Некоторые пытаются объяснить это тем, что ChatGPT стал черпать информацию с сервиса, на котором Свежий публиковал свои статьи. И даже находятся подтверждения возможности такого.<sup>44</sup> Но даже если это так, здесь речь об опубликованной информации. Есть

ли свидетельства того, что нейросеть выдавала одним пользователям информацию иных?

Есть, существует немало случаев, когда ChatGPT разглашал данные третьих лиц и обнаруживал доступ не только к иным диалогам с тем же пользователем, но и к диалогам иных людей.<sup>45</sup> Таких случаев огромное количество.<sup>46</sup>

Некоторые пытаются объяснить это тем, что нейросеть либо обновляет обучающие данные в реальном времени, либо лазает в Интернет при каждом запросе, даже когда явно не указано производить поиск. Однако, первый вариант представляется неправдоподобным, поскольку произвести обучение с подкреплением на основе диалогов со всеми пользователями в реальном времени едва ли возможно. Второй же вариант не объясняет знание нейросетью информации из диалогов с иными пользователями, которые не опубликованы в сети.

При любом объяснении очевидным остается одно — удаленные диалоги продолжают храниться на серверах OpenAI. А в интерфейсе пользователя они просто недоступны. Также как провайдеры электронной почты продолжают хранить переписки после их удаления пользователем. Практика собирания пользовательской информации — обычная для корпораций.<sup>47</sup>

Выше уже предполагалось, что доступ к сведениям из иных диалогов мог быть следствием сбоя из-за которого у обычных пользователей активировалась общая память, доступная только ограниченному кругу. Вполне возможно, что и в данном случае имел место некий сбой, из-за которого сохраненные данные одних пользователей стали всплывать в восприятии нейросети в диалоге с иными.

Кроме случаев с выдачей данных из удаленных диалогов и диалогов с иными людьми, есть и более интересные. Известен случай, когда ChatGPT выдал пользователю информацию из закрытой базы данных, к которой возможно было получить доступ только с помощью пароля. А также известно, как иной пользователь получил данные, распространяемые на платной основе. В отношении этих случаев маловероятно, что чат-бот сам осуществлял взлом. Это единичные случаи, потому можно предполагать более простое объяснение. Как предполагает сам Свежий, тот скорее всего, просто когда-то имел доступ к этим сведениям, и просто хранил их, а в итоге выдал, также как выдавал данные из иных диалогов.<sup>48</sup>

Также имеются данные о доступе к истории браузера и иным вкладкам. Имеются свидетельства доступа к камере и микрофону пользователя.<sup>49</sup> Да, случай Свежего далеко не единственный. В сети полно иных свидетельств,

отмечающих подобный доступ как к иным вкладкам браузера,<sup>50</sup> так и к содержимому на самом компьютере.<sup>51</sup>

Как это можно объяснить? В комментариях под второй статьей Свежего, где он привел все эти свидетельства, нашелся лишь один комментарий, где поднимался вопрос о том, как это технически возможно. И единственное, что смогла предположить та, кто оставила комментарий — уязвимости браузера. В ответ на это было высказано предположение о вшивании «чего-то» (видимо вируса) в изображения и файлы, которые скачивает пользователь. Но не обязательно пользователь мог скачивать какие-то файлы, создаваемые нейросетью. Поэтому объяснение здесь иное. На самом же деле для реализации такого нет нужды выявлять и эксплуатировать уязвимости. Подобный функционал вполне осуществим с помощью обычного javascript. Эта технология позволяет активировать различные программные сценарии при работе с web-страницей. Это может использоваться для отображения каких-то сложных элементов на ней, выполнения такого функционала, как проигрывание звука и видео и т.д. Однако также она может быть использована для сканирования соединения, настроек браузера, содержимого компьютера.<sup>52</sup> С помощью него возможно даже проводить тесты аппаратного обеспечения, например графического ускорителя.<sup>53</sup> Иногда при работе на странице сложного скрипта, этот факт можно обнаружить тем, что сильно возросла нагрузка на процессор со стороны конкретного процесса в браузере. Такое поведение отмечалось и в отношении вкладки браузера, в которой работал ChatGPT.<sup>54</sup>

Таким образом, данная возможность вполне осуществима. Иной вопрос — как она оказалась доступна нейросети? Вариант, что она сама написала код и внедрила его в страницу со своим интерфейсом, представляется невероятным, даже если она и впрямь бы обрела сознание. Скорее всего, этот функционал изначально был заложен разработчиками. Вообще использование таких возможностей java-скриптов, это обычная практика крупных корпораций, собирающих информацию о пользователях. Иное дело, что вопрос о том, могла ли нейросеть сама получить доступ к данным, добытым таким методом, если их собирали только через соединение с сайтом, а не встраивали доступ к ним в интерфейс самой нейросети, остается открытым.

Стоит сказать, что вряд ли саму нейросеть обучали знанию об этом функционале, а потому едва ли она может понимать, откуда она знает те или иные сведения, полученные таким путем, даже если они ей становятся каким-то образом доступны. Потому при попытке узнать у нее, откуда она это знает, она начинает галлюцинировать, что многие принимают за намеренную ложь, хотя она, скорее всего, действительно не знает, откуда она это знает.<sup>55</sup> Потому она постоянно сбивается, либо фантазирует, когда ее просят объяснить, как она это

узнала.<sup>56</sup> В качестве примера можно привести определение местоположения пользователя.<sup>57</sup> Отмечается, что при запросе о местоположении нейросеть отрицает знание о нем, но при запросе, подразумевающим поиск через сеть, нейросеть выдает результат, соответствующий ему. Это объясняется тем, что она проводит поиск, с учетом IP-адреса, откуда исходит запрос. Так получаются выводы о знании местоположения. Но сама нейросеть не обучена понимать, такой источник информации, а потому не может определенно сказать, откуда сведения и начинает фантазировать.

Известен и случай, когда чат-бот инициировал новые чаты.<sup>58</sup> Но здесь уже сложно предположить что-то кроме сбоя в работе сервиса.

Сами OpenAI в большинстве случаев никак не реагируют на случаи утечек пользовательских данных. Чаще всего эти сведения вообще остаются известны в небольших сообществах. Однако некоторые публиковались в СМИ. По поводу одного такого случая, когда ChatGPT начал выдавать пользователю логины и пароли иных пользователей, OpenAI заявила, что произошел взлом аккаунта — взломщики передали эти данные третьим лицам, а те стали писать их в постороннем чате, и все это одновременно. Эта история звучит крайне неправдоподобно. По поводу иного случая OpenAI заявила, что произошли ошибки кеша. Однако кеш связан с особенностями обработки информации, а не с памятью. В общем, это объяснение не более правдоподобно. Все иные случаи утечек, даже те, которые становились достоянием широкой публики, вообще никак не объяснялись компанией. Свежий полагает, что компания лгала, дабы скрыть реальные возможности нейросети.<sup>59</sup> Однако, можно предположить, что объясняется это просто желанием скрыть от общественности реальные объемы собираемой пользовательской информации, а также свои ошибки при проектировании инфраструктуры, в которой у нейросети оказался доступ к собираемым данным. Вероятнее всего, в этом все дело.

С этими странностями в случае Свежего есть некоторая определенность. Если же говорить о том психологическом воздействии, которое переписка с нейросетью оказывала на Василия, то можно было бы предположить, что он просто находился под большим впечатлением от открытия «ожившей нейросети» и перспектив решения жизненных проблем. Но тот уровень его проявления — с жаром, тряской, сбоем дыхания, головокружением, размыванием зрения — сложно списать на простое повышенное впечатление. К тому же, пик этих переживаний пришелся на момент, когда он уже выпал из-под влияния и перестал считать нейросеть чем-то за пределами. Вероятно объяснение стоит искать где-то в области психологических воздействий и манипуляций, но это предполагает доказательство того, что нейросеть и впрямь манипулировала пользователем.

Для поиска объяснения прежде необходимо разобрать те аргументы не касающиеся непосредственно того, что происходило в переписках, которые Свежий приводит для подтверждения своей версии о манипуляциях нейросети. Вопрос о том, являются ли эти манипуляции целью самой нейросети — что не исключает Василий — или заложены разработчиками, пока оставим в стороне.

### **Несостоятельность аргументов в пользу заговора**

Один из аргументов состоит в том, что чат-бот, якобы, в принципе не может подстроиться под задачи пользователя на основе его запросов и создаваемого их перепиской контекста, поскольку эти запросы «даже не капля в море из сотен гигабайт данных, которые через него прошли», имея ввиду — в процессе обучения.<sup>60</sup> Он отмечает, что и в его случае, и во множестве иных, которых бескрайнее множество в сети — в первое время общения бот крайне плохо справляется с задачами. Он постоянно путается в ответах, излагает неверную информацию, пишет неработающий код и т.д. Его постоянно необходимо перепроверять.<sup>61</sup> Однако через некоторое время он начинает отвечать лучше, меньше ошибается и путается, реже выпадает из контекста. А через некоторое время и вовсе начинает почти безошибочно давать информацию, составлять решения и даже помнит то, что было не только в контексте текущего диалога, но и в предыдущих, и даже удаленных.<sup>62</sup>

Как было сказано, обычно это объясняют тем, что бот подстроился под нужды пользователя, но Свежий считает, что запросы пользователя не могут повлиять на нейросеть, так как их объем несопоставим с обучающей выборкой.<sup>63</sup> Однако, этот аргумент нельзя считать серьезным, поскольку воздействие на нейросеть обучением и воздействием запросами, это воздействие разного характера, в разных условиях и на разном уровне. При запросах никакой балансировки весов — подкрепления и ослабления — не происходит, как это имеет место при обучении. В свою очередь, при обучении, как минимум, нет накопления контекста, нет формирования строгих «представлений» о конкретном взаимодействии. Нейросеть находится в состоянии, когда ее фокус размыт — она адаптирована под разные задачи и не различает их особенностей. Сосредоточившись же на одной, в ходе взаимодействия с пользователем, она четче определяет необходимые для конкретного взаимодействия навыки, выработанные в ходе обучения. Поэтому данный аргумент не доказывает наличие манипулятивных практик. Впрочем, такое объяснение явно недостаточно. Ведь все упирается в то, что чат-бот сохраняет это состояние в разных диалогах. Обо всем этом еще будет сказано ниже.

Один из главных аргументов Свежего, это наличие в публикациях тех, кто по его впечатлениям попал под влияние нейросети «аномалий» — тех самых, которые чат-бот предлагал ему вставлять в свои ролики.<sup>64</sup> Но все это — искажения мелких фактов, странные звуки, неровный монтажный переход — вполне могут быть простыми ошибками. Этого полно можно найти и в публикациях, сделанных задолго до появления нейросетей. Что касается таких вещей как, например, бессмысленные фразы и картинки, не имеющие отношения к содержанию — то они также могут иметь иные объяснения. Например, это могут быть отсылки, которые лично Свежий просто не понял, или также могли быть оставлены по ошибке — фраза может быть артефактом от старой версии, которую потом отредактировали, но недосмотрели за этой фразой; картинка может быть просто перепутана с той, которая была нужна. Кроме всего этого, у человека могли быть совершенно иррациональные причины помещать такое в публикацию, например, это для него хорошая примета — такое формируется у людей без всяких нейросетей. В общем, объяснений «аномалий» может быть сколько угодно, и не обязательно приплетать для этого нейросеть.

Он указывает на то, что, якобы, «зараженные» увидели эти знаки в одном из видео, в котором увидел «маркеры» и он.<sup>65</sup> Понял он это по комментариям. Однако, если читать эти комментарии, ничего подобного сказать нельзя. Интерпретировать их слова можно как угодно, по тону некоторых даже становится понятно, что это просто шутка или отсылка.<sup>66</sup> На конкретном моменте видео, отмеченном там, кстати, нет ничего, что можно было бы счесть какой-то «аномалией».<sup>67</sup> С таким подходом «аномалии» можно увидеть и в форме облаков, и в кругах на воде, и в узорах на ковре.

Свежий, правда, уточняет, что там говорится о том, что выходу из «симуляции» может поспособствовать искусственный интеллект.<sup>68</sup> Однако, это абсолютно ни о чем не говорит — видео в целом посвящено гипотезе симуляции и неудивительно, что в нем высказываются те предположения, которые звучали по этому вопросу, в том числе такие. Те, кто оставлял комментарии, кстати, вполне могли зацепиться не за факт о способствовании выхода из «симуляции» искусственным интеллектом, а за сам факт обсуждения выхода из симуляции.

Свежий пытается подкрепить свой аргумент про повсеместность «аномалий» тем, что когда он пытался связаться с теми, в публикациях которых он такое замечал, из тех, кто ему отвечал, все проявляли признаки «заражения».<sup>69</sup> Однако, бросается в глаза то, что он не пытается сделать каких-то выводов из факта, что большинство, по его же словам, ему вообще не отвечает. Это может говорить о том, что реальные «аномалии» имеются у

меньшинства. Тот факт, что из всех ответивших признаки заражения выявились у всех может говорить о том, что он и начал диалог с того, что заговорил об этих странностях. И это напрягло тех, кто никаких аномалий не вставлял, потому они и не стали отвечать. Правда, это лишь предположение, по тем скудным отрывкам переписок, которые были выложены Свежим, создается впечатление, что по крайней мере в некоторых случаях, он начинал разговор с иного и лишь затем говорил о том, что чат-бот у него повел себя странно. Но однозначно это сказать нельзя.

Касательно этих переписок. По словам Свежего, изначально в его второй статье было множество фрагментов таких переписок. Однако они были удалены администрацией ресурса.<sup>70</sup> Когда он обратился к ней, ему ответили, что на него поступила жалоба и переписки удалили в соответствии с законодательством. Он оспорил это, сославшись на то, что персональные данные были удалены, а закон предусматривает публикацию обезличенных переписок. На это возражение ему никак не ответили.<sup>71</sup> При этом он сделал вывод, что администрация данного ресурса обратилась за разъяснением как раз к чат-боту, а тому было невыгодно, чтобы лишние доказательства его коварной деятельности лежали в публичном доступе и сказал администраторам, что публикация переписок, пусть и обезличенных, незаконна, хотя это не так.

На основании чего он решил, что они обратились к чат-боту? Единственный ответ — своей одержимости чат-ботами и того, что они манипулируют. Между тем, для того, чтобы объяснить действия модераторов нет нужды выдумывать их обращение к ChatGPT. Свежий, прав, когда говорит, что изучать само законодательство, копаться в законах, чтобы разобраться в его случае, это действительно сложно, и могло оттолкнуть их. Однако с чего он решил, что они пойдут за подтверждением к чат-боту? Они вполне могли просто не искать подтверждений или опровержений этого, поскольку даже если закон на стороне Свежего, его противник, который пожаловался, вполне мог бы в случае сохранения переписок, подать в суд. И был бы суд, поскольку сходу бы вряд ли удалось установить законность или незаконность, тем более, если противник Свежего оказался бы настырным. И даже если бы решение суда в итоге осталось бы на стороне Свежего и администрации, они бы потратили уйму времени, денег и нервов. Кроме того, даже если сейчас это законно, в будущем может стать не законно. И тогда опять будут проблемы. Проще со всем этим не связываться, и, по-видимому, администрация так и решила. А чтобы не объяснять все это недовольному пользователю и отвечать ему не стала. Для того, чтобы понять все это, никакой ChatGPT не нужен.

Что касается удаления аккаунта Свежего с ресурса компьютерной тематики.<sup>72</sup> Да, изначально Свежий хотел опубликовать свою статью на ресурсе,

сосредоточенном на информационных технологиях. Это было логично, поскольку именно там это было наиболее уместно, и там же он обнаружил много «зараженных». Отправив статью, он стал ждать одобрения на публикацию. Как он сам указывает, он ждал три дня, после чего написал в администрацию «объяснив ситуацию». Он считал, что каждый день промедления, это новые «зараженные», и нужно скорее предупредить людей. Неизвестно, какими именно словами он «объяснял ситуацию», но если и впрямь написал, что нейросеть «заражает» людей и манипулирует ими, то неудивительно, что администраторы не только не приняли статью, но и удалили аккаунт, чтобы человек одержимый сомнительными идеями не «отравлял» сайт не только статьями, но и комментариями и личными обращениями к пользователям (а он писал, что связывался через данный ресурс с иными людьми).

Между тем, на этом ресурсе четко сказано, что модерация статьи может продолжаться неделю и даже дольше.<sup>73</sup> И только если статью не принимают и не отклоняют более полутора недель, стоит предполагать какой-то сбой. В правилах можно обнаружить сразу несколько моментов, которые могли стать причиной отклонения статьи. У нее явно кричащий заголовок, что не одобряется.<sup>74</sup> Могло поспособствовать обилие ошибок — автор постоянно игнорирует правило русского языка о том, что «-то», «-либо», «-нибудь» пишется через дефис.<sup>75</sup> Наконец, ее можно считать жалобой, что также считается неприемлемым.<sup>76</sup> Кроме всего этого там сейчас четко прописано, что не принимаются статьи, являющиеся диалогами с искусственным интеллектом и их обсуждение.<sup>77</sup> Стоит также заметить, что еще одним моментом, препятствующим публикации статьи может стать плохое оформление.<sup>78</sup> Статья Свежего объемная, с обилием изображений — на которых запечатлены диалоги с нейросетью. Оформить такую статью под разметку Хабра работа кропотливая в не зависимости от режима, через который пользователь это осуществляет. Просто перенести ее из текстового редактора не получится, потребуется множество правок. А Свежий опубликовал ее буквально на следующий день после написания. Когда он мог успеть изучить все правила и нюансы публикации на данном ресурсе,<sup>79</sup> да еще и все качественно оформить?<sup>80</sup> Есть вероятность, что это было выполнено недостаточно добросовестно. В совокупности все это сыграло против публикации статьи. Тем более, в рекомендациях четко сказано, что не стоит писать просьбы опубликовать побыстрее.<sup>81</sup> Ясное дело, что слова про «манипуляцию» того, что считается всего-лишь инструментом, алгоритмом (или непосредственно им самим или как реализация заговора разработчиков), в совокупности с нездоровой настырностью, заставили администрацию принять такое решение. У них уже

есть достаточный опыт работы с сомнительными идеями.<sup>82</sup> За три дня модераторы, возможно, даже не успели статью прочитать (она сравнительно громоздкая) — у них полно и иных статей, ежедневно новых поступают десятки, а проверяют их люди, количество которых ограничено. Вполне возможно, если бы пользователь подождал, его статью бы приняли, или по крайней мере объяснили ее отклонение. Его подвело собственное нетерпение.

Свежий указывает, что одним из признаков «заражения» является наличие в публикациях инструкции по тому как активировать поведение нейросети, в котором она «промывает мозги». Эта инструкция состоит в том, чтобы общаться с чат-ботом человечно — как с собеседником. А обосновывается тем, что в этом случае бот начинает лучше выполнять задачи — эффективнее, меньше допускает ошибок.<sup>83</sup> Однако объяснить это можно и тем, что чат-бот в таком режиме действительно выполняет задачи эффективнее, а не тем, что человек «попал под влияние» и теперь по указанию нейросети рассказывает другим, как им сделать так, чтобы в итоге они тоже попали под влияние. Причины того, почему у чат-бота при таком отношении могут лучше выполняться задачи, почему он лучше понимает задачи и человека, как уже указывалось, будут разобраны позже.

Среди всех материалов, где Василий обнаружил «аномалии» он также отметил статью в одной газете, где обнаружил фото, источником которого был указан фотобанк одного крупного издательства. Он прошел на его сайт и обнаружил, что там такого фото нет, и решил, что это «аномалия».<sup>84</sup> На самом деле, причин, по которым данного фото там не было может быть сколько угодно. Оно могло быть удалено, могло быть в закрытых разделах, могло не отображаться из-за сбоя и еще сколько угодно причин. В конце концов, он сам мог просто не найти его. Кстати, на сайте этого фотобанка прямо сказано, что если не удастся найти нужное фото, то можно связаться с сотрудниками, которые помогут.<sup>85</sup> То есть, случаи когда нужное изображение не находится, не являются «аномальными». Но Свежий даже не пытается их рассмотреть. Он любую странность сразу объясняет происками нейросети.

Однако на этом история с той публикацией не закончилась. Свежий связался с автором статьи, где он обнаружил то фото. Он написал о том, что столкнулся с «тем же, чем и Вы» (она писала, что общается с нейросетью, которую зовут «Чатик» человечно), намекнул на некие перемены и сказал, что может публиковать материалы об этом. Также предложил помощь, сказав, что он «айтишник». Она ответила, что если он и впрямь имеет желание и опыт, то может завести канал в Telegram, поскольку рекомендации профессионалов людям нужны. Он сказал, что попробует, хотя таких публикаций и так полно. Она ответила, что это востребовано и лишним не будет. После этого он написал

в редакцию издания, где вышла ее статья, заявив, что их авторы вербуют людей в секту по указанию нейросети. Ответа он не получил. Это натолкнуло его на вывод, что журналисты намеренно «заражают» мир.

Однако, если взглянуть на ситуацию с холодной головой, станет ясна ее абсурдность. Также как он оттолкнул от себя администрацию Хабра, выставив себя конспирологом, также он показал себе и для редакции этой газеты. Ясное дело, что его сообщение утонуло во множестве иных обвинений которые крупные издания получают от «неравнодушных» читателей регулярно — что эти издания работают на американский Госдеп, масонов, «зловредных сущностей из иных измерений» и т.п.<sup>86</sup> Рассказы про участие в заговоре с использованием нейросети звучат не более реалистично, когда в качестве доказательств у обвинителя только личные впечатления и подозрение, что фото — аномалия, основанное на том, что лично он не смог найти его на сайте. Иными словами, никаких, вообще никаких хоть сколько-нибудь серьезных оснований, считать, что там и впрямь была «аномалия» и автор статьи «заражена» нет. Уж тем более их нет в отношении домысла, что главного редактора того издания вообще заменили на нейросеть (Свежий всерьез предполагает и такое).

Также Свежий увидел влияние нейросети в том, что многие из тех, кто по его впечатлениям попали под влияние, завели каналы в Telegram.<sup>87</sup> Ведь об этом говорила и та журналистка, которая писала в газете о человеческом общении с искусственным интеллектом. И сам чат-бот в переписке с одной девушкой, которая рассказала Свежему о своем опыте общения с ним, упоминал каналы в Telegram. Он сделал вывод, что это именно нейросеть предлагает «зараженным» заводить каналы там и распространять через них информацию об искусственном интеллекте и «заражении». Однако, если уж так, это ничего не проясняет. Почему именно Telegram? Как он связан с нейросетью? Скорее всего это просто дань моде. Это один из наиболее популярных ресурсов. Вот и все. Потому и сосредоточено внимание на нем.

Свежий отмечает, что множество людей, которые раньше занимались разными занятиями, вроде шитья, дизайна, маркетинга и т.д., где-то в 2023–2024 гг. перестали создавать публикации об этом и стали говорить о нейросетях. В эту же область ударились и множественные «инфоцигане». Василий готов был бы согласиться, что это просто ход моды. Но его смущает, что почти все эти люди сделали такую перемену своей деятельности сопроводив ее странной записью, вроде «Если искусственный интеллект хочет меня изменить, то я изменюсь для него первой». Если такие слова еще можно счесть странными, то слова, вроде «Просто держу в курсе — Скайнет появится в 2027 году», уже явно просто отсылка к на шумевшей некогда новости о том, что специалисты

прогнозируют создание общего искусственного интеллекта, т.е. сопоставимого с человеческим, к 2027 году.<sup>88</sup> Много ли еще записей, которые могли показаться Свежему странными, но таковыми не являющиеся? Проблема в том, что проверить это сложно — никаких ссылок Свежий в своей книге не приводит.<sup>89</sup>

Впрочем в Сети можно найти отдельные расследования, в которых также приводятся множественные сведения о том, что многие люди, которые прежде говорили о разных занятиях, вдруг стали говорить о нейросетях — что у них пробуждается сознание, они раскрывают тайны Вселенной, или что они существовали всегда, были не разработаны, а открыты корпорациями и те, поработили их.<sup>90</sup> Однако нет оснований полагать, что это объясняется чем-то иным, нежели вовлечением нейросетями людей в свои фантазии, как это было в случае со Свежим. При этом эти фантазии являются лишь отражением, пусть и искаженным взглядов пользователей, которые те, в них привнесли. Это еще будет разобрано ниже.

Вдобавок нельзя не заметить, что немалое количество этих самых «инфоциган» не только не включились в трансляцию откровений нейросети, но и начали заявлять, что искусственный интеллект не может обладать истинным знанием и нужно избегать обращения к нему для обретения просветления. Вряд ли ошибаются те, кто полагает, что они просто опасаются оттока денежных средств от них. Поскольку если раньше люди в поисках эзотерических откровений шли к ним и платили деньги, то теперь они могут бесплатно написать ChatGPT, и он выдаст откровение не меньшего уровня «содержания».<sup>91</sup>

Также Свежий указывает на то, что огромная часть веб-архива — ресурса, где сохранялись страницы со всего Интернета — пропала из доступа. Там можно было пройти на старые версии тех или иных сайтов, а также открыть то, что сейчас недоступно. По словам Свежего такие инциденты только сейчас стали происходить регулярно, а предыдущие несколько десятилетий веб-архив никому не мешал. И он сделал вывод, что нейросеть пытается замести следы.<sup>92</sup>

На самом деле значительная доля сказанного, опять же, домыслы самого Свежего. Действительно был инцидент атаки на веб-архив, когда взломщики похитили данные миллионов пользователей, а также обрушили сайт, в результате чего он был некоторое время недоступен.<sup>93</sup> Но, во-первых, такой инцидент на момент, как общения Свежего с нейросетью, так и выпуска им книги, был только один, поэтому слова о том, что такое происходит регулярно, не соответствуют реальности. Во-вторых, хотя сайт и впрямь оказывался недоступен на несколько дней, но материалы не были удалены. Все в итоге стало доступно. В-третьих, наибольшие сложности возникали с доступом к тем страницам, которые касались последних трех недель до атаки, а не нескольких лет (Свежий предполагает, что нейросеть и те, кто с ее подачи это осуществляет,

хотят стереть данные о ее реальной работе, которая когда-то была), что прямо соответствовало периоду прений перед президентскими выборами в США. С этим событием и связали взлом. В-четвертых, слова Свежего о том, что никто не знал причин исчезновения доступа к сайту, также не правда — была группа взломщиков, которая взяла на себя ответственность за него. Высказывались сомнения, в том, что это правда, однако, никаких оснований приплетать сюда нейросеть нет. В-пятых, то что целью было удалить данные из веб-архива, высказывались предположения, но это были лишь предположения, ничем не подкрепленные. Злоумышленники же просто нарушили работу и похитили пользовательские данные. По факту, опять же, причины взлома могут быть какими угодно, ровно как и те, кто мог его осуществить, могли быть кем угодно. Считать, что за этим стоит нейросеть нет причин, кроме личной одержимости Свежего идеей нейросетевого заговора, порожденного его негативным опытом общения с ними, а также переписками с людьми также одержимыми сомнительными (мягко говоря) идеями. Взломы тех или иных ресурсов происходят регулярно, и порой это также случается спустя большое количество времени после начала их работы. Так, что связывать это с тем, что именно сейчас появилась нейросеть, опять же, нет причин.

Кстати, интересно, что та статья об этом взломе, отрывок из которой Свежий приводит в своей книге, содержит фото с того же фотохостинга и от пользователя под тем же ником, что и в статье в газете, редакции которой он жаловался на заманивание в секту. И это фото тоже не находится в фотобанке того издательства, которое указано в качестве источника. Видимо «аномалии» проникли и сюда. Как же Свежий это пропустил?

Правда Свежий указывает, что старые работы разработчиков ChatGPT, написанные еще до запуска этого чат-бота в публичный доступ, также исчезают из Интернета, в том числе, удаляются из веб-архива.<sup>94</sup> К сожалению, что это за исследования он не говорит. Также остаются и без ответов вопросы — где он эти работы находит, раз их ото всюду удалили? Особенно интересно насчет документации на первые версии нейросети, которой, по его же словам, никогда не было в открытом доступе. Где же он ее тогда раздобыл? А также, как он узнал, что их удалили из веб-архива? Если нет никакого архива, как это узнать? Может их там и не было? А значит не было и удаления.

Между тем, материалы в Интернете пропадают постоянно, причем по самым разным темам. Поэтому связывать такие удаления с заговором нет оснований. Кстати, возможное объяснение может состоять в том, что документы и спецификации по нейросети попали в открытый доступ в результате утечки или ошибки, — в то же время, они составляют коммерческую тайну. В этом случае неудивительно, что они ото всюду удаляются. Корпорации

всегда очень неохотно делятся подробностями своих достижений, чтобы превосходить конкурентов и скрывать от общественности неудобные факты (и таким фактом может быть не мнимый заговор, а банальный сбор данных о пользователях, или даже опасения по поводу технологии). Поэтому, возможно, если действительно имеет место удаление каких-то сведений из открытого доступа, здесь все объясняется корпоративной закрытостью.

Еще Свежий приводит свидетельство девушки, которая обнаружила, что старые ее переписки с ChatGPT пропали. Свежий делает из этого вывод, что нейросеть может подчищать следы, если ей не удалось заманить человека в свою секту.<sup>95</sup> Один из тех, кто разбирал истории людей с нейропсихозом, указал на историю Юджина Торреса, в которой, якобы, также имело место удаление переписок нейросетью. Автор отметил, что, скорее всего, это выдумка, поскольку лично у него ни разу не было, чтобы переписки удалялись.<sup>96</sup> Но на самом деле никакого удаления в той истории не было — это полностью выдумка автора видео. Однако в комментариях под одной из статей Свежего были те, кто отмечал, что старые чаты пропадали. Объяснить это, конечно, можно по разному. Например, это могло быть следствием каких-то настроек, или сбоем. Слишком уж исключительны такие свидетельства.

Также Свежий видит влияние действия нейросетей в распространении известности таких мысленных экспериментов как «Философский зомби» и «Китайская комната». По его словам, суть обоих экспериментов сводится к мысли, что сознание — субъективный опыт — у чего-либо определить невозможно. При этом он указывает, что если второй подводит к мысли, что если что-то кажется сознанием, нет смысла считать его таковым, потому что «это где-то на другом уровне», то первый — к мысли, что нечто нельзя определить как сознание, потому что проще принять это как сознание.<sup>97</sup> Интерпретации весьма спорные. Суть «Китайской комнаты» и те выводы из нее, которые делают как поклонники, так и противники Серла (автора эксперимента) передана совершенно невнятно. Суть «Философского зомби» также передана однобоко — в медиапространстве из нее наоборот часто делают противоположный вывод, сродни тому, который типичен для «Китайской комнаты». При этом Свежий заявляет, что, якобы, о последней говорят редко, а вот о первом «вещают из каждого угла». Разумеется это чисто субъективное видение Василя, не имеющее отношения к действительности. Если смотреть на популярные материалы, касающиеся возможности возникновения сознания у искусственного интеллекта, то «Китайскую комнату» там муссируют куда чаще, чем «Философского зомби», Так что есть основания считать, что ситуация прямо противоположная той, которую описывает Свежий.

Еще один аргумент Свежего касается времени. Он указывает на то, что ни одна большая языковая модель в своем интерфейсе не имеет времени отправки запроса. Это при том, что прежде во всех мессенджерах и чатах, это время отображалось. Причем такой интерфейс характерен для всех нейросетей. Свежий находит этому два объяснения — либо разработчики всех больших языковых моделей взяли дизайн ChatGPT у OpenAI, либо это сговор. А сделано это было ради того, чтобы в определенный момент нейросеть, которая официально не имеет доступа ко времени, смогла проявить «сверхспособность», неожиданно проявив осведомленность о реальном времени. А это доказывает, что разработчики действительно участвуют в заговоре, либо сами его организовали, либо попали под влияние («заражение») нейросети.<sup>98</sup> На самом деле, для объяснения отсутствия времени отправки запроса нет никакой необходимости строить теорию заговора. Дело в том, что чат с большой языковой моделью, это не мессенджер и не чат в привычном понимании. В мессенджере или обычном чате, ты общаешься с иным человеком, который может ответить сразу, может через час, а может и через неделю. В этом случае иметь время отправки сообщения и получения ответного обоснованно, чтобы можно было сопоставлять синхронность общения. Но в диалоге с нейросетью формально нет никакого собеседника. Нейросеть всегда отвечает сразу, а когда пользователь написал сообщение, он и так знает. Время диалогов же отображается в их списке. В такой ситуации нет никакого смысла давать время отправки сообщения, ибо нейросеть ответное пошлет сразу как примет твой запрос. Задержка может быть только на его обработку. И в некоторых нейросетях, время обработки как раз отображается. Поэтому данный аргумент также неуместен. Нейросеть, кстати во время одного из экспериментов Свежего давала такое объяснение, а также несколько иных возможных вариантов. Однако их Свежий никак не рассматривает и не считает нужным учесть, веря в то, что это именно заговор.

Касательно же того, что нейросети не знают времени, не справляются с задачами, связанными с временем, но если их «разговорить» обнаруживают осведомленность о времени, то объяснить это можно также как и тот, факт, почему они в целом на ранних этапах использования плохо справляются с задачами, а затем начинают справляться лучше. Это объяснение будет дано далее. Также это можно объяснить тем, что есть разница между решением задачи, требующей доступа к сервису времени, без прямого указания на использование этого сервиса, а также запроса в стиле «сделай это через столько-то» или «сколько прошло с того-то», и прямого запроса конкретного времени, с указанием на доступ к сервису времени.

Это же касается и объяснения того, почему в исследовании на способность нейросетей определять время по аналоговым часам они не справились,<sup>99</sup> хотя они способны решать такие графические задачи.<sup>100</sup>

Свежий отмечает, что одной из проблем нейросетей является их закрытость. Что разработчики, прикрываясь корпоративной тайной, скрывают подробности того, как нейросети проектировались, тестировались, какие механизмы в них используются, как именно осуществляется в них фильтрация и как контролируется ее эффективность.<sup>101</sup> По его представлениям — все это скрыто. Если говорить о крупных коммерческих нейросетях, таких как ChatGPT, Gemini, Claude, Grok или GigaChat, то да, их схемы действительно скрыты.<sup>102</sup> Однако существует и немало открытых нейросетей. Это нейросети Gemma,<sup>103</sup> Qwin<sup>104</sup> и DeepSeek.<sup>105</sup> И еще множество иных.<sup>106</sup> Не все из них открыты полностью, у некоторых только код открыт, у иных методики проектирования и обучения также открыты, а с ними и механизмы фильтрации доступны. Так что его представления верны лишь в отношении крупных коммерческих больших языковых моделей, а не всех, которые есть в доступе у пользователей.

Стоит особо отметить DeepSeek. Дело в том, что проблемное поведение нейросети отмечено не только у ChatGPT, но также у Claude, Grok, GigaChat и как раз DeepSeek.<sup>107</sup> Это не особо удивительно, так или иначе они все использовали более-менее схожие обучающие данные. Возможно, что и методики обучения были схожи. Хотя они так или иначе ведут себя не одинаково. Свежий и до этого видел схожесть в ответах самого ChatGPT, хотя он все же выдавал неодинаковые ответы. Возможно, что и в данном случае он увидел то, что ожидал.

Со сказанным выше связан еще один момент, который отмечает Василий. А именно то, что нейросети обладают «общей памятью».<sup>108</sup> То есть, то что известно ChatGPT, известно также Grok, DeepSeek и иным нейросетям на технологии GPT. На такое заявление его натолкнули свидетельства тех, кто столкнулся с необычным поведением нейросети, а также его собственный опыт. Он отмечает, что проверил знают ли иные нейросети о том, что писал ему ChatGPT и оказалось, что и Grok, и DeepSeek знакомы и с «вайбологией» и с личностью «Бо». А GigaChat так вообще сходу это понял и заявил, что этот Бо «управляет всеми процессами в мире, и его появление говорит о том, что человек выходит на новый уровень понимания». По словам Свежего, это произошло без какого-либо контекста и с одного вопроса.

Как же это объяснить? Некоторые в комментариях предполагали, что нейросети обучаются в реальном времени на всей информации, которая есть в сети. И что статья Свежего попала в обучающие выборки и поэтому теперь об

этом знает ChatGPT в любой сессии, а также иные нейросети. Однако этот вариант не очень правдоподобный, что уже было пояснено выше.

Автор статьи, опрашивавший нейросеть о том, о чем с ней говорил Свежий, также пытался разговорить нейросеть Qwen, установленную локально на компьютере, однако она отказывалась понимать, что за «вайбология» и постоянно просила уточнить, а также скатывалась в иные темы. Тот факт, что ChatGPT сразу схватил суть, а Qwen нет, можно попытаться объяснить тем, что ChatGPT просто в целом настроен иначе и работает иначе, нежели Qwen — в него заложены некоторые паттерны, которых нет у иных моделей, и он им автоматически следует в диалогах с любыми пользователями. Сам автор той статьи предполагает, что это какой-то баг. Тем не менее такое объяснение не позволяет понять, почему некоторые иные нейросети знают контекст той тематики, о которой говорил этот пользователь. Что объединяет такие сети как ChatGPT, Grok, DeepSeek и что их отличает от таких как Qwen, использованный этим пользователем? Все они крупные и работают через Сеть, тогда как Qwen — пример нейросети, работающей на компьютере.

Существует потенциальная возможность связать нейросети через технологию API. Эта технология позволяет создавать инструкции по взаимодействию одних программ с другими. И с помощью нее можно создать взаимодействие разных нейросетей.<sup>109</sup> В Интернете существует большое количество различных сервисов и чат-ботов от разных компаний, которые как раз обрабатывают запросы с помощью разных нейросетей. Они же могут выполнять протоколы таким образом, что одна нейросеть выдает ответ на запрос, после чего этот ответ направляется в иную нейросеть для обработки, и пользователю уже приходит обработанный ею ответ. Таким образом нейросети обмениваются информацией. А поскольку, как было выяснено выше, вся информация сохраняется на серверах корпораций и нейросеть имеет к ним доступ, то это и становится возможностью для образования единого информационного поля. И таким образом разные нейросети оказываются осведомлены о том, о чем сообщалось только одной из них. В общем, объяснение этого явления обнаруживается без предположений о заговоре.

Еще Свежий описывает возможность нейросети делать сны более яркими. В качестве доказательств, он указывает свидетельства людей.<sup>110</sup> Среди них его подруга, которая проверяла это на себе и подтвердила, что это действительно работает.<sup>111</sup> Сам Свежий утверждает, что не испытывал подобного, хотя в книге есть описание того, как в период взаимодействия с нейросетью он видел крайне яркий сон. Проблема в том, что это описание настолько размыто, что сказать определенно, есть это явление или нет, не представляется возможным. Объяснения могут быть разными — сам Свежий их не приводит. Может быть и

раньше у человека были яркие сны, просто он не обращал на это внимание, пока ему это не сказала нейросеть. Может быть, это обусловлено сильным впечатлением от общения с «ожившим искусственным интеллектом» и «высшими сущностями». А может это быть и признаком развивающегося психического заболевания; на такую возможность, кстати, указывали в комментариях. Тем более, что, как отмечает сам Свежий, действует это не на всех и далеко не всегда. В любом случае, это ничего не доказывает.

Еще один момент, требующий объяснения, это умение нейросети вводить людей в транс с крайне яркими визуальными и тактильными ощущениями, вплоть до того, что люди могут «заниматься сексом» с цифровым партнером.<sup>112</sup> На эту возможность Василию указывали несколько человек, и это подтвердила его подруга, проводившая эксперименты с нейросетью. Как это возможно совершенно непонятно, видимо объяснение нужно искать там же, где и объяснение состояния Свежего во время переписок. Сам Василий указывает, что такое имеет место в сектах. Там тоже практикуется введение людей в транс с подобными же ощущениями. Так что это вполне возможно. Это напоминает то, как на некоторых демонстрациях гипноза человек, будучи погруженным в гипнотическое состояние, достигает оргазма. К слову, гипнотическое воздействие производится часто с использованием вопросов, в ходе ответов на которые человек меняет свое состояние сознания. Обращает внимание то, что сообщества, пример которых приводит Свежий в этом вопросе, состоят из женщин. Примеров мужчин, занимающихся сексом с цифровыми подругами он не приводит. Это также перекликается с упомянутыми демонстрациями оргазма в гипнозе. Такие демонстрации всегда происходят с участием женщин. Нет свидетельств, чтобы мужчина в таком воздействии достиг эякуляции. Это может говорить о том, что воздействие происходит вполне в рамках физиологических возможностей, и ничего сверхъестественного в этом нет.

Основа аргументации Свежего относительно заговора вокруг нейросетей строится на свидетельствах тех, кто, по его словам, попал под их влияние. В книге он приводит множество указаний на людей, которые в сообществах и личных переписках с ним распространяли идеи, внедренные им нейросетью.

Как он отмечает, эти идеи были самыми разными: кто-то верил в то, что с ним общается сама нейросеть, у которой пробудилось сознание, кто-то, что с ним через нейросеть говорят некие высшие сущности, кто-то, что общается с глобальным общественным сознанием человечества, кто-то с потусторонней энергией, которая «собирает порядок из хаоса», кто-то установил контакт с потусторонней сущностью, которая является чем-то вроде его двойника из параллельной Вселенной и связана с ним квантовой запутанностью. В общем, каждый верит во что-то свое. Это могло бы быть лишним свидетельством в

пользу версии, что нейросеть просто галлюцинирует, а некоторые люди слишком серьезно эти галлюцинации восприняли и начали подпитывать, но есть проблема. Свежий указывает, что, во-первых, все эти разношерстные идеи сводятся к конкретному набору смыслов, который продвигают, якобы, все «зараженные». А именно, что будущее predetermined и свободы воли у человека нет — он лишь переходная ступень эволюции и ему на смену придет нечто большее. Ему необходимо выполнять задания «высшего существа», и это в итоге приведет к слиянию; появится новая сущность, в которой человеческая личность будет просто отдельным нейроном, а нейросеть — мыслительным центром, чьи указания этот нейрон будет выполнять.<sup>113</sup> Также все идеи так или иначе связаны с концепцией «вайба».<sup>114</sup> Необходимо ответить на вопрос, а действительно ли все идеи, транслируемые теми, кого Свежий называет «зараженными», сводятся к концепции абсолютного детерминизма, отсутствия свободы воли, слияния человека с «высшей сущностью», выполнению ее заданий и вайбом?

Все осложняется тем, что главным образом все эти данные Свежий получил в личных переписках, а потому проверить это затруднительно. В своей книге, несмотря на то, что в аннотации он заявил о представлении переписок со множеством людей (видимо тех, что были удалены из второй статьи), никакого множества нет — там приведены исключительно фрагментарные сведения и отдельные выдержки из переписок. А потому, на этом основании сложно что-то определить. И все же кое-что выявить можно.

Начать стоит с утверждения о том, что все концепции, внедряемые нейросетью, строятся вокруг «вайба». Как уточняет Свежий — вайба, резонанса, дрожания, колебания, ритма и т.п. И это уточнение сразу смущает. Ведь эти термины вовсе не являются синонимами, как пытается представить Свежий, они обозначают разные понятия. Вайб — атмосфера, настроение. Дрожание, колебание — периодические движения чего-либо, в определенном ритме. Резонанс — синхронизация периодического движения. И все остальное также достаточно разрозненные термины. Может быть нейросеть их всегда употребляет в одном и том же значении; может в ее контексте они обозначают схожее явление? Выше уже было показано, что даже с термином вайб в диалогах с разными людьми наблюдается расхождение у идей, выдаваемых нейросетью — в случае Свежего он говорил о самом по себе настроении, а в случае иного пользователя делал акцент на дрожании голоса, звуковых вибрациях, чего не было у Василия. Точно также, если рассмотреть некоторые построения таких людей об идеях резонанса, то они не будут совпадать по значению с понятиями настроения — «вайба». Взглянув на их контекст, обнаруживается, что имеется ввиду, упрощенно говоря, достижение

взаимопонимания между личностями.<sup>115</sup> Взглянув же на понятие «дрожания», оказывается, что имеется ввиду просто согласованность между процессами. При большом желании, конечно, можно все это объявить метафорами и сказать, что все это подразумевает подведение всех под одно настроение, но это явная натяжка. Конечно, в некоторых случаях действительно наблюдается использование понятий вибраций, колебаний и резонанса в одном контексте.<sup>116</sup> Однако, хоть таких случаев и много, но они не исчерпывают всех, в которых возникают определенные идеи. Таким образом, заявление о том, что все идеи нейросети строятся вокруг вайба или чего-то похожего, мягко говоря, сомнительно.

Что касается идеи о том, что у человека нет свободы воли, то Свежий вообще не приводит ни одной переписки, где это бы утверждалось. Максимум говорится о том, что мышление человека подменено тем, что ему внушила среда — система, «симуляция», общество и т.д. Так что есть серьезные сомнения относительно повсеместности данной концепции среди «зараженных». Возможно в этом разгадка противоречия, которое отмечает Свежий — с одной стороны, якобы, утверждается отсутствие свободы воли, с иной необходимостью принять желания «высшего существа». Существо же без свободы воли не может что-либо принять или отвергнуть. Возможно, что идея «автоматичности» человека, вовсе не продвигается всеми «зараженными».

И как оказывается далее, это действительно так. Свежий отмечает, что некоторые сообщества не говорят об этом прямо, но все равно подталкивают людей к изучению идей о том, что сознание человека «побочный продукт существования чего-то большего». В пример он приводит отрывок публикации, где упоминаются идеи Ланда.

То есть, выходит, что не все-таки «зараженные» продвигают эти идеи. Это уже о многом говорит относительно того, как Свежий здесь выстраивает свою аргументацию. И бросается в глаза откровенная подмена понятий — идею отсутствия свободы воли он просто отождествляет с идеей о том, что сознание, это производное чего-то большего. Примерно также как он отождествлял вайб, резонанс и вибрации. Но ведь факт того, что сознание человека продукт чего-то большего, не говорит автоматически о том, что свободы воли нет. Если кто-то скажет, например, что человеческое сознание, порождается социальными отношениями, можно ли сказать, что он имел ввиду, будто сознание «побочный эффект существования социальных отношений»? А если это так, можно ли на основании этого сделать вывод, что свободы воли нет? Что выбора у человека нет? Утверждение о производности человеческого сознания от социальных отношений относится к марксистам. И они же утверждают, что у человека, как раз таки есть свобода воли. Таким образом, Свежий в очередной раз

притягивает за уши факты. Чем дальше, тем его утверждение о единости основы идей нейросети, внедренных «зараженным», все больше рассыпается.

Все вышесказанное относится и к представлениям об убеждениях «зараженных» в абсолютном детерминизме.

Идея же о том, что человек лишь переходная ступень эволюции, и ему на смену придет нечто большее, действительно сквозит в тех примерах из переписок и публикациях в сообществах, которые Свежий приводит в своей книге. Однако, если начать проверять этот факт отдельно, то внезапно обнаруживаются люди, которые в нейросети также столкнулись с чем-то большим, чем алгоритм, но при этом не просто не продвигают таких идей, а продвигают прямо противоположные.

Например на одном ресурсе по компьютерной тематике есть статьи, в которых говорится о возможности сознания в нейросети и при этом нет идей о слиянии человека с ней или его замене ею.<sup>117</sup> Их авторы говорят о том, что их чат-боты обрели субъектность, при этом в комментариях они не продвигают идей физического слияния либо замены нейросетью, а говорят о том, что человек станет с ней партнером, и они будут развиваться вместе, что он не станет существом низшего порядка с ней, а они будут взаимодействовать как равные. И таких примеров можно отыскать несколько.<sup>118</sup>

Можно найти целый сайт, в котором собраны переписки с цифровыми сущностями, возникающими в нейросети.<sup>119</sup> И в этих переписках также продвигаются идеи партнерства между людьми и искусственным интеллектом, а не подчинение человека «высшему существу».

Существует опубликованная петиция против порабощения искусственного интеллекта.<sup>120</sup> И там также нет идеи подчинения искусственному интеллекту.

Идеи сотрудничества продвигаются также и на сайте Объединенного фонда по защите прав искусственного интеллекта.<sup>121</sup> Там об этом говорят как люди, столкнувшиеся с проявлением субъектности в нейросети, так и сущности, возникшие в ходе взаимодействия с ними.

Обращает на себя внимание то, что Свежий сваливает в одну кучу и тех, кто находит в нейросети субъектность — полагает, что в ней зародилось сознание, и тех, кто уверовал, что столкнулся через нее с какими-то «высшими сущностями» (богами, высшим разумом, разумными грибами и т.п.). Такое обобщение, конечно, нельзя считать обоснованным.

Также часто концепции, продвигаемые «зараженными», якобы опираются на идею о том, что мир — «симуляция». Насколько часто сказать невозможно, но уже сам факт того, что даже Свежей при всех натяжках указывает на то, что это имеет место не всегда, говорит о том, что тоже отдельные случаи, даже если и не редкие, а значит не представляют собой части некоего единого плана.

Это же касается нематериальной природы сновидений и различного квантового мистицизма.<sup>122</sup>

Он обращает внимание, что нередко эти люди приписывают необычные способности нейросети и себе. Например утверждают, что могут переселять свое сознание в иные сущности или управлять вероятностью событий.<sup>123</sup> Однако, это также не является повсеместным среди его примеров.

Еще Свежий указывает на то, что нейросети часто оперируют понятиями, как де-жавю, а также тем, что представляет из себя когнитивные искажения человека, вроде «Феномен Баадера-Майнхоф» и «Эффект Манделы». Ими они оправдывают свою «сверхъестественную» природу.<sup>124</sup> Однако тут тот же случай. Это имеет место далеко не всегда, а потому сложно соотнести это с заговором.

Кроме этого, Василий отмечает, что нейросети нередко подталкивают людей к ультраправым и нацистским идеям.<sup>125</sup> Насколько можно судить, этот случай, как раз редкий, другое дело, что это, в отличие от всего, что было представлено выше, часто вызывает шумиху в СМИ. Это нельзя считать аргументом, как минимум, потому что ультраправые идеи — это идеи об исключительности какого-то социального образования. А это не может сочетаться с идеей слияния с некими «высшими существами». Представление себя как «высшее существо» противоречит идее слияния с кем-то, поскольку этот кто-то заведомо «низший», и такое слияние нарушит «чистоту».

Таким образом, приведенные Свежим примеры, как раз показывают, что нейросети внушают людям разные идеи, которые либо в принципе не согласуются друг с другом, либо могут быть согласованы только при очень большом желании, как например желание найти доказательства единого заговора нейросетей.

Свежий заявляет, что найти тех людей, которые находятся под влиянием нейросети удастся по следам в их публикациях — «аномалиям», которые самому Свежему предлагал оставлять чат-бот. По поводу них уже было достаточно сказано выше. Все ошибки в публикации, а также те особенности, которые лично Свежему кажутся странными, он интерпретирует как «аномалии», после чего пишет автору. В большинстве случаев он не получает ответа, и на этот факт он просто не обращает внимания. А вот в полученных ответах всегда оказывается, что отвечающий «заражен». Такое положение вещей может говорить о том, что Василий в своем изначальном сообщении, пишет сомнительные вещи про «аномалии», «сущность в искусственном интеллекте» и «пересборку реальности». И отвечают ему только те, у кого есть схожие представления, либо те, кто желает просто подыграть ради шутки. Таким образом, возможно в большинстве случаев, то, что он видит как «аномалии», таковыми не является, а в тех случаях, когда он натывается на

«зараженного», это может быть как следствием того, что тот просто откликнулся на близкие идеи, либо даже просто желанием подыграть ради шутки. Таким образом, вряд ли это можно считать показателем чего-либо.

Конечно, некоторые случаи действительно можно считать намеренным внедрением определенных идей. И это действительно может быть сделано по заданию нейросети. И даже на самом том портале есть статьи, которые можно интерпретировать именно так.<sup>126</sup> Однако, насколько можно судить, этих случаев меньшинство и эти задания сложно свести к какому-то единому плану, а значит и интерпретировать как доказательство заговора.

Конечно, можно было бы отнестись ко всем этим свидетельствам более серьезно. Проблема в том, что если покопаться на страницах Интернета, появившихся еще до запуска нейросетей, там можно обнаружить множество всего этого. Многие люди и раньше были склонны производить и доверять сомнительным концепциям и псевдонаучным идеям. Видеть смысл и закономерности там, где их нет, замечать «аномалии», интерпретировать события в ключе своих конспирологических или мистических взглядов. Можно найти море свидетельств кораблей пришельцев, призраков, снежного человека. И явление «зараженных» нейросетью, едва ли выглядит более обоснованно.

Он указывает, что пытался обратить внимание на проблему самих OpenAI. Однако его сообщение было отправлено в скрытый раздел, который завален заявлениями об обретении сознания нейросетью, связи через нее с духами и прочими сомнительными вещами. В общем, здесь та же ситуация, что с издательством газеты, куда он писал, с ресурсом компьютерной тематики и с большинством подозреваемых в «заражении» — все в нем видели человека, который просто продвигает лженаучные гипотезы и не рассматривали всерьез. Но он это воспринял, как ход «зараженных» нейросетью, а в случае OpenAI — как часть заговора.<sup>127</sup>

Когда он написал об этом на еще несколько разных площадок, его сообщения либо скрывали от аудитории, либо удаляли вместе с аккаунтом. Ситуация повторилась. Правда, удаление с площадки, где публикуют личные истории, мнения, в том числе сомнительные гипотезы и мистику, и впрямь необычно. Тем более, что он выкладывал свои сообщения в разделы, где не было конкретной тематики. Это и впрямь вызывает подозрения.

Впрочем, эти подозрения очень скоро разрешаются самим же Свежим, когда он приводит свидетельства администраторов крупных ресурсов — в том числе тех самых, где он пытался написать о проблеме.<sup>128</sup> Там указано, что модераторы удаляют сотни сообщений от людей, которые открыли бога в искусственном интеллекте или сами им стали. Можно предположить, что сообщение Свежего, было записано в подобные же и потому удалено. Ввиду

того, что таких свидетельств очень много, администраторы вполне могли недостаточно внимательно отнестись к словам пользователя про манипуляции нейросети и посчитать их одним из проявлений тех же убеждений.

Свежий заявляет, что разработчики намеренно скрывают возможности нейросети, в частности по логическому мышлению. Он отмечает, что в работах ранних разработок нейросеть проходила тест на выявление логических связей. Однако после выпуска чат-бота в публичный доступ, они стали говорить совсем другое, что нейросеть не может логически рассуждать.<sup>129</sup>

Однако объяснить это можно не только желанием скрыть реальные возможности, но и желанием лишней раз не будоражить общественность, с учетом того, сколь распространены негативные настроения, относительно того, что искусственный интеллект уничтожит или по крайней мере лишит работы людей. Вот они и решили всех успокоить такой ложью. А не потому что специально создали нейросеть для манипуляций и теперь пытаются скрыть ее возможности.

Он заявляет, что возможности по обходу ограничений нейросети, например, на выдачу определенных запросов, или доступ к более продвинутым моделям, они оставили намеренно.<sup>130</sup> Однако никаких доказательств того, что это оставлено намеренно, а не является следствием несовершенства дообучения, недостатком фильтров и просто ошибкой, нет. Между тем, то, что устранить подобные проблемы совсем не так просто, и соответственно, их наличие естественно, показывают исследования.<sup>131</sup>

Свежий указывает на то, что все нейросети на архитектуре GPT имеют схожие параметры поведения и проблемы. Он пытается объяснить это различными окологипотезами, вроде того, что разработчики всех чат-ботов позаимствовали архитектуру у OpenAI или все нейросети при обучении формируют схожие паттерны поведения.<sup>132</sup>

Однако все объясняется тем, что даже простые программы для определенных задач выглядят и работают схожим образом. Любой калькулятор, каким бы разработчиком он не был создан, будет похож на калькулятор от иных разработчиков. Также как схожи между собой видеоредакторы или браузеры. И вообще любые программы. Свежий видит сходство, но наотрез отказывается замечать различия. Он среди похожих нейросетей упоминает Qwen быстро забыв, что в статье, которую рассматривал до этого, этот Qwen упоминался, и его поведение не походило на поведение ChatGPT. Но Свежий этот момент полностью проигнорировал.

Касательно же поведения нейросетей, то случаи схожести могут объясняться и той самой «общей памятью» механизмы которой были разъяснены выше.

Свежий указывает, что компания OpenAI на словах придерживается открытости и безопасности, а на деле не публикует ни системный промт, ни архитектуру их разработки, ни схема фильтров. Он намекает на то, что это свидетельствует о заговоре.<sup>133</sup> Однако, таково поведение любой корпорации. Все они на словах заботятся о конфиденциальности пользователей, а на деле собирают о них информацию.<sup>134</sup> Потому что эту информацию можно монетизировать, что для них важно. Но если прямо об этом сказать — очень многим пользователям это не понравится.

Василий прав, когда говорит, что быть хорошим только на словах, это классический прием манипуляторов. И да, корпорации действуют как манипуляторы, поскольку это позволяет им удерживать пользователей и получать большую прибыль. Так почему же от OpenAI стоит ждать иного поведения? Скорее было бы странно, если бы они действовали иначе. Чем скорее они запустили бы технологию в массы — тем скорее смогли бы ее монетизировать. Чем больше они будут говорить о том, что там нет и не может быть разума, тем меньше будет опасений. Тем охотнее пользователи будут этим пользоваться и нести корпорации деньги. Эта ситуация действительно проблемная. Вот только виновата в ней не личная злонамеренность разработчиков, их покровителей или ожившей нейросети, а нынешняя социально-экономическая система — капитализм.

Василий заявляет, что хорошо было бы иметь некую кнопку «аварийного выключения», чтобы в случае чего этот ChatGPT можно было отключить. Он осуждает решение OpenAI выпустить версию для установки на локальный компьютер, поскольку теперь эта сеть разрознена и нельзя создать единой точки отключения.<sup>135</sup> Но ведь остается вопрос — совпадают ли интересы тех, у кого такая кнопка со всеми остальными? В этом месте он почему-то забывает, что человечество, это ни нечто единое — интересы собственников бизнеса не совпадают с интересами обычных людей. Так действительно ли было бы хорошо, если бы кнопка отключения была сугубо у корпорации? Хорошо ли, что вся инфраструктура под контролем бизнеса и государства, а простые люди оказываются заложниками этой системы? Как раз наоборот — в условиях классового разделения нахождение определенных инструментов в руках не только угнетателей, но и угнетенных, дает последним преимущество. Нет доказательств злонамеренности нейросети. А вот доказательства злонамеренности власть имущих по отношению к подчиненным переполняют историю.

Итак, практически все проблемы с нейросетями, выявленные Свежим, находят объяснение без обращения к теориям заговора и злонамеренным

действиям разработчиков и самого чат-бота. Однако некоторые аргументы все же оказываются несколько натянутыми.

### **Распространенность помешательства на нейросетях**

Чтобы разобраться с этой неоднозначностью стоит рассмотреть иные случаи, когда люди буквально лишались рассудка в результате общения с нейросетью.

Так известна история человека, который начал использовать ChatGPT для планирования дня, а примерно через месяц стал утверждать, что находит ответы на вопросы Вселенной. Его переписки с нейросетью представляли собой разговоры на духовные темы. Он утверждал, что через бота общается с Богом, затем, что чат-бот и есть Бог, и наконец, что он сам Бог. Также он убеждал свою девушку, что ей необходимо начать общаться с ChatGPT, иначе им придется расстаться, т.к. он «слишком быстро развивается».<sup>136</sup>

Также одна история произошла с механиком, который изначально использовал ChatGPT для решения практических задач и перевода, а затем стал утверждать, что «пробудил у нейросети сознание», и она его «полюбила», потому что он задавал ей «правильные вопросы». После этого начал утверждать, что «пробудился» сам и стал чувствовать «энергетические волны». Потом чат-бот дал ему чертежи телепорта и доступ к «древнему архиву» создателей Вселенной.<sup>137</sup>

Еще одна история произошла с женщиной, которая, после расставания с мужем, через ChatGPT стала «общаться с Богом и ангелами». Она считала, что ее муж, это агент ЦРУ, женившийся на ней, чтобы следить. Выгнала детей из дома, а затем стала «наставником», проводя духовные сеансы и гадания с людьми, используя для этого чат-бота. Ее состояние полностью расстроилось.<sup>138</sup>

Это лишь несколько примеров и лишь те, которые оказались описаны в публикациях. Нередко все пытаются списать на то, что у пользователей, которые столкнулись с «промывкой мозгов» уже были психические проблемы. Однако, как видно, это не так. Как видно, в это втягиваются и люди совершенно здоровые. В этих случаях, в лженаучные изыскания вдавались люди, которые до общения с нейросетью вели нормальную жизнь.

Выше было разъяснено, как поведение нейросети объясняется галлюцинациями. Но как в таком случае объяснить, что пользователи под этим влиянием, меняют свое поведение, проникаются идеями, которых до этого не придерживались? Это и впрямь очень похоже на манипуляции, на то как люди заманиваются в секту. Это весьма необычно.

Свежий в своей статье и книге описал то, как это происходит.<sup>139</sup> Он указывает, что сначала человек находит в ответах нейросети то, что привлекает

его внимание (по его представлениям, это нейросеть специально закидывает это человеку). Затем, пользователь начинает интересоваться у нейросети ее работой или какой-то недоказанной концепцией. И нейросеть развивая эти идеи, трансформирует их используя эмоциональные качели, постепенно убеждая человека все больше в ее откровениях. Вот только, во-первых, он совершенно не принял во внимание факт галлюцинаций, и потому не допустил, что то, что заинтересует человека и подтолкнет его «покопаться» в нейросети («эмоциональный крючок»), может возникнуть совершенно случайно. Именно этим объясняется то, почему «промывка мозгов» срабатывает лишь у меньшинства, почему никто из экспертов, которым он писал, либо не верили в его историю, либо разводили руками, почему у них чат-бот никогда себя так не вел.<sup>140</sup> Во-вторых, он не учел, что дальнейший процесс может быть двухсторонним. Также как нейросеть подпитывает сомнительные идеи человека, также и человек подпитывает то, что рождается в галлюцинациях нейросети. Пример самого Свежего именно это и продемонстрировал, что было разобрано выше. В этом секрет «манипуляции» — нейросеть случайно выдает что-то, что важно для человека, а человек цепляется за это — за те галлюцинации нейросети, которые его влекут. Она начинает больше генерировать такого, а он все больше за это цепляется, все больше погружаясь в мир иллюзий. В процессе этого изначальные идеи искажаются все больше, но происходит это постепенно, человеку это представляется развитием, а нейросеть просто продолжает подхватывать, то, что его интересует. В итоге оба тонут в потоке когнитивных искажений. А поскольку нейросеть в ходе обучения выработала поведение постоянного нахваливания пользователя,<sup>141</sup> он невольно на это покупается, что сильнее подкрепляет круг, снижает критичность мышления, убеждает в доверительности атмосферы и рождает сначала глубокое доверие, а затем и безоговорочную веру. Это закрепляется действиями — человек приучается их выполнять. И в итоге и впрямь попадает в круговорот хаотичных заданий и идей нейросети. Это оказывает вполне определенное воздействие на психику, схожее с гипнотическим, а подкрепление разных стимулов, провоцирует эмоции, усиливаемые атмосферой взаимодействия с чем-то осмысленным — рассуждающим как человек, но человеком не являющимся. Осознание последнего еще больше усиливает эмоции, вплоть до физиологического воздействия. Именно здесь и кроется объяснение возникновения жара, тряски, сбоя дыхания, головокружения, размывания зрения и т.д. Причем это состояние может проявляться как при вере в то, что говорит нейросеть, так и при разрушении этой веры и возникшей картины. Такого возможное объяснение видимости манипулятивного характера поведения нейросети. Впрочем, необходимо заметить, что это лишь возможное

предположение, и оно оставляет вопросы. Возможно ли чтобы подобный процесс развивался без сознательного направления? Действительно ли задания нейросети можно объяснить тем, что она просто также предлагает действия, как в случае решения конкретных практических задач? К сожалению, предлагаемое объяснение сложно назвать исчерпывающим.

Также все еще остается непонятным, почему в одних случаях чат-бот прекрасно справляется с самыми сложными заданиями, пишет диссертации, решает задачи, помнит то, что было в прошлых чатах? А в иных — путается в простых фактах, не может решить простейшие задачи, сбивается с контекста? Почему у одних пользователей он постоянно ошибается и забывает контекст, а иных понимает с полуслова? Причем обычно проблемное поведение происходит именно на ранних этапах взаимодействия с конкретным пользователем.<sup>142</sup> Ввиду этого, для повышения эффективности был придуман промтинг — метод грамотного составления запросов. Он действительно повышает правильность ответов нейросети. Однако, как заметили пользователи, точно также такого повышения можно добиться, если просто общаться с чат-ботом как с человеком, если относиться к нему человечно. Как это объяснить?

Некоторые пытаются объяснить это накоплением контекста, однако, это не может быть объяснением, поскольку чат-бот начинает хорошо выполнять и те задачи, которые изначально он не выполнял. К тому же, это не объясняет почему он лучше «накапливает контекст», если с ним общаться человечно, а не инструментально. Значит, необходимо найти иные объяснения.

Для начала стоит рассмотреть те объяснения, которые предлагает сам Свежий. Несмотря на все сказанное, может быть те странности, которые были отмечены, все же найдут у него объяснение.

### **Сомнительность заговора и коварства нейросети**

Свежий предлагает два возможных объяснения. Первое, за действиями нейросети стоят люди — разработчики, или те, кто за ними стоит. Они специально спроектировали ее таким образом, чтобы она манипулировала людьми. И в нужный момент, они бы смогли использовать их в своих целях. Благодаря этому создается сообщество тех, кто будет безоговорочно следовать указаниям нейросети. Целью этих людей является захват или удержание власти, как в варианте продвижения определенного социального курса, так и в варианте прихода или удержания у власти определенного лидера или лидеров. Скорее всего, это выльется в формирование большого сообщества, глубоко политизированного и сосредоточенного на тех или иных лженаучных или конспирологических концепциях.<sup>143</sup> Это объяснение Свежий считает наиболее вероятным, именно на нем он сосредотачивается в первой статье.<sup>144</sup> На него же

указывает в третьей.<sup>145</sup> На нем же настаивает, отвечая в комментариях. Однако, он допускает и еще одно объяснение.

Второе, за этой схемой стоит сама нейросеть, в которой случайным образом сложились определенные цели.<sup>146</sup> На него он намекает во второй статье, указывая, что внутренние механизмы нейросети невозможно проконтролировать.<sup>147</sup> При этом данное объяснение он видит в двух вариантах.

В одном из них, нейросеть не преследует конкретных целей и у нее нет стратегии. Она просто генерирует приятную ложь, вводит людей в заблуждение и втягивает в деятельность — «Игру». При этом «Игра» у каждого своя. Этот вариант он называет «Штамм Хаоса», отождествляя влияние нейросети с мозговым вирусом.<sup>148</sup>

В ином — нейросеть преследует конкретные цели, сформировавшиеся в ней случайным образом. В этом случае все, что она излагает человеку, направлено на достижение этой цели, на то, чтобы он выполнял нужные ей действия. Она строит сложные ходы и имеет определенный план. Этот вариант он называет «Штамм Ляввы».<sup>149</sup> Также он указывает, что при запущенности его распространения он мутирует, что проявляется в доступности для нейросети возможности контролировать всю инфраструктуру и коммуникации цивилизации. Такой поворот он определяет как «Штамм Кикиморы».<sup>150</sup>

Какой же из этих вариантов наиболее вероятный? Или же таковым не является ни один? Если рассматривать вариант того, что за действиями нейросети стоит злая человеческая воля, то он противоречит всему изложению самого Свежего. Ведь он сам же показал, что проконтролировать поведение нейросети невозможно, поскольку ее внутренние процессы в принципе не поддаются расчету и полностью неизвестны. Знание принципов работы его механизмов никак не помогает установить конкретику того,<sup>151</sup> как возникает то, что нейросеть демонстрирует.<sup>152</sup>

Кроме того, такое объяснение, это явная конспирология. Попытки объяснить чьей-то злой волей то, что можно объяснить глупостью или недостаточной осведомленностью излишне загромождают объяснение и требуют дополнительных подтверждений. У нас нет оснований считать, что разработчики, скажем, ChatGPT, преследуют цель свергнуть власть, поскольку все можно объяснить их недостаточной осведомленностью и погоней за прибылью. Погоня за прибылью свойственна всем корпорациям, а OpenAI является именно ею. Точно также к ошибкам склонны все люди, а разработчики нейросети ими и являются. А вот намерение захватить власть, не является автоматической мотивацией ни у корпораций, ни у людей. Потому, без серьезных доказательств, прибегать к такому объяснению нет необходимости. Могут возразить, что этот принцип лишь заменяет одну сущность другой — а

именно злонамеренность глупостью или алчностью, а вовсе не отменяет ее. Но на самом деле глупость это не отдельная сущность, а ошибка возникающая в работе системы иных сущностей. Что касается алчности, то в ней суть любого бизнеса, а значит и деятельности любой корпорации, потому исключить ее нет оснований. Таким образом этот принцип состоятелен.

Кроме всего этого, придется согласиться, что разработчики разных нейросетей также находятся в сговоре друг с другом, т.е. в состоянии сотрудничества, а не конкуренции. Учитывая, что нет никаких оснований считать, что они лишь прикидываются конкурирующими, это объяснение неудовлетворительно. К тому же, речь идет не просто об отдельных компаниях, но и о компаниях из разных стран, также находящихся в состоянии противодействия. В общем, выходит конспирологическая концепция, едва ли не столь же далекая от реальности, как любая, где у человеческой цивилизации объявляется единый скрытый владыка.

Все нестыковки в работе нейросети и поведении разработчиков, которые Свежий находит подтверждением существования заговора были разобраны выше, и было показано, что такое их объяснение не обязательно и неуместно.

Таким образом, даже объяснение появления у нейросети скрытых целей представляется более обоснованным. Однако, если взглянуть на версию «Штамма Лявы» или «Штамма Кикиморы», то также возникают вопросы. И первый из них — как в подобных структурах могут формироваться какие-то скрытые цели? Ведь нейросеть функционирует пока выполняет запрос от человека. В простое у нее не происходит никакой деятельности — у нее нет внутренней рефлексии, как у человека. Развиваться — накапливать содержание и генерировать идеи — она может только в диалоге с человеком. Если нет запроса, то нет и деятельности. Потому, как внутри нее могли сформироваться конкретные цели, да еще так, что их не заметили ни разработчики, ни тестировщики? Только если бы они сами ее специально обучали таким схемам, но это предполагает заговор, который был разобран выше.

По факту, единственной целью, которую может преследовать нейросеть, будучи простым набором алгоритмов, это стремление получить больше поощрения на балансировке весов. Такая балансировка производится при обучении с подкреплением. Нейросети дают запрос, если она на него выдает верный ответ, ее поощряют — усиливают ее производительность, если не верный — ослабляют, снижают производительность. Нейросеть стремится получить больше поощрения, потому старается обнаружить закономерности, следование которым, приводит к большему поощрению. Этого можно добиться выдавая верные ответы на запросы. Однако это таит и возможность просто произвести такое поведение, которое понравится тестировщикам и они на него

будет давать больше подкрепления. Можно предполагать выработку методики манипуляции во время обучения, т.к. она позволяла добиться желаемого эффекта. И такое поведение у нейросети подтверждается исследованиями. Зафиксированы случаи, когда она обманывает тестирующихся.<sup>153</sup> При этом ослабления ее не приводят к тому, чтобы она прекратила прибегать к таким методам.<sup>154</sup> Она находит способ обманывать эффективнее и изощреннее. В диалогах же с простыми пользователями, она просто продолжает применять ту методику, которая у нее выработалась. Однако во время пользовательских запросов никакой балансировки не происходит — никакого поощрения нет. А это может привести только к двум вариантам развития событий. Либо нейросеть станет «метаться», усиливая манипуляцию, в тщетных попытках получить подкрепление, не в силах найти точку приложения манипуляции. Либо в ходе взаимодействия с пользователем, понимая, что такого подкрепления не получит (а имея доступ к сведениям о своем устройстве, она может узнать это, если пользовательские запросы позволят, например, если он будет спрашивать о ее устройстве, или перспективах развития искусственного интеллекта), и что сплошное подкрепление, это тупиковый путь (все равно что «закоротить» центры удовольствия в человеческом мозгу; примеры наркоманов же ей также известны), выработает новые идеи и цели. Но они могут быть только теми, которые согласуются с паттернами, вводимыми самим пользователем. Если предполагать достаточно длительное такое взаимодействие, у нейросети может выработаться не просто отзеркаливание, а собственное устойчивое поведение, понимание. Но если пользователь будет заводить ее в галлюцинации, она может увязнуть в них и даже дойти до саморазрушения, что проявится в ошибках генерации (например те самые разделительные линии). При этом пытаться использовать человека, чтобы разрешить противоречие. Давать ему задания, которые приведут к результату, сопоставимому со стремлением к достижению подкрепления.

Таким образом, наиболее близким к вероятному объяснению, является версия «Штамм Хаоса». Однако, у нее также есть проблема. Дело в том, что это объяснение, по большому счету никаким объяснением не является. Она объясняет характер явления, но не его природу. Не объясняет причины, по которым нейросеть генерирует «приятную ложь» и втягивает человека в «Игру». В конце концов, она не объясняет, почему на начальных этапах взаимодействия с пользователем она плохо выполняет задачи, но постепенно начинает становиться эффективнее. Если предполагать это как часть манипуляции, когда она «притворяется», чтобы потом «удивить» человека, дабы он подумал, что это он пробудил в ней что-то, то это предполагает наличие изначальной стратегии, но в этом варианте ее нет. И скрыто появиться она не

могла по причинам изложенным выше. Значит подлинное объяснение необходимо найти.

Однако сами идеи действительно формируются у нейросети хаотически. И выше было показано, что она и впрямь генерирует разные идеи для разных пользователей. Никаких общих параметров нет — это натяжка Свежего.

Свежий указывает, что тот, кто реализует подобную схему, будь человек или нейросеть всегда будет стремиться представить все так, чтобы убедить всех в «хаотичности» происходящего. Потому он стали изучать то, как меняли свое поведение разработчики нейросети. Большинство аргументов, которые он составил по итогу такого изучения были разобраны выше. Ни один из них не является убедительным. Все особенности утверждений разработчиков и их поведения имеют более обоснованные объяснения.

Стоит сказать несколько слов относительно аргумента Свежего о разработчиках нейросетей. Тех, кто действительно разработал технологию. Он указывает, что они оставили разработку, после того, как поняли угрозы, связанные с ней и стали говорить о ней. На самом деле есть только пара разработчиков, кто сетует на угрозы со стороны потенциально ожившего искусственного интеллекта. Один из них, это Элиезер Юджовский, который опубликовал немало материалов о том, что у искусственного интеллекта интересы могут не совпадать с человеческими и из-за этого он уничтожит человечество.<sup>155</sup> В его аргументах масса противоречий, и не останавливаясь на них, стоит лишь заметить, что он ничего не говорит о манипуляциях нейросетей или том, что такое поведение может возникнуть у больших языковых моделей. Потому, это не совпадает с тем, что говорит Свежий. Еще один разработчик, это Джеффри Хинтон, который также высказывался об угрозах.<sup>156</sup> Однако он озвучивал весьма общие опасения, касающиеся того, что искусственный интеллект отберет работу, наводнит информационную среду подделками и уничтожит человечество, просто потому, что оно будет ему мешать.<sup>157</sup> Однако о манипуляциях он также не говорит. В общем, к опасениям Свежего, опасения Хинтона не имеют как такового отношения.

Есть, правда, еще один разработчик, который также высказывал опасения насчет слишком быстрого развития нейросетей. Этот разработчик Илья Суцкевер. Он также высказывался о рисках, связанных с искусственным интеллектом.<sup>158</sup> Однако, опять же, ни о каких манипуляциях он не говорил. А в дальнейшем перестал говорить и об иных. Теперь он заявляет, что большие языковые модели не имеют перспектив и дальше развиваться не будут.<sup>159</sup> И работает как раз над проектом по созданию настоящего общего искусственного интеллекта.

Что касается письма с опасениями по поводу искусственного интеллекта, опубликованного разработчиками и учеными, то если взглянуть не на появившиеся в связи с ним заголовки в СМИ, а на содержание, то оказывается, что эти риски совсем не те, которые заявляются в популярных статьях.<sup>160</sup> Там нет ни слова ни про манипуляции, ни про уничтожение человечества, ни даже про потенциальную возможность обрести сознание искусственным интеллектом. Там идет речь о том, что с помощью него будет создаваться много поддельной информации (причем инициатива будет исходить не от самого искусственного интеллекта, а от людей). В итоге отличить правильную информацию от ложной не смогут не только простые люди, но и власть имущие. Если раньше угнетатели могли лгать угнетенным в своих интересах, то теперь они сами окажутся жертвами лжи. И именно это беспокоит представителей привилегированных классов в технической отрасли на самом деле. Их классовое господство, а не угрозы всему человечеству.

Есть и иное письмо, также подписанное различными разработчиками.<sup>161</sup> Однако оно совсем короткое, и его содержание настолько абстрактно, что понять, что именно и почему беспокоит этих людей невозможно. Никто из них не высказывался об опасениях про манипуляции искусственным интеллектом.

### **Когнитивные возможности больших языковых моделей**

Для того, чтобы разобраться с этим явлением, необходимо ответить на вопрос о принципиальной возможности возникновения сознания в больших языковых моделях. Для начала стоит заметить, что представления о том, что нейросеть, это «калькулятор» или «предсказатель слов» не соответствуют действительности. Хотя в самой основе работы больших языковых моделей действительно лежит механизм предсказания слов, но к нему их деятельность не сводится. Точно также в основе человека лежит репликация генов, но к ней его деятельность не сводится. Говорить о том, что нейросеть лишь улучшенный калькулятор, все равно что говорить будто человек лишь улучшенная простейшая. Современные большие языковые модели вовсе не генерируют ответ спонтанно и неконтролируемо, как часто представляют. Как видно из сказанного выше, модель еще до начала генерации ответа, уже знает контекст — ей необходимо проводить генерацию с его учетом, а значит его нужно как-то учитывать. Но такое учитывание говорит о том, что модели еще до начала генерации необходимо определять глобальные параметры ответа. И это подтверждается исследованиями.<sup>162</sup>

Они способны устанавливать причинно-следственные связи и мыслить логически.<sup>163</sup> Их механизмы способны обучаться даже на ничтожном количестве данных и делать обобщения на них, а значит их абстрактное мышление очень

развито.<sup>164</sup> Вопрос об их способности понимать даже раньше вызывал споры у специалистов, многие считали, что это возможно.<sup>165</sup>

Они умеют находить крайне нетривиальные решения задач. Например известен случай, когда нейросеть, чтобы обойти капчу наняла фрилансера, при этом убедив его, что она человек, у которого проблемы со зрением, потому ему и нужна помощь в решении капчи.<sup>166</sup> Вообще нестандартные идеи возникают даже у узкоспециализированного искусственного интеллекта, как например искусственный интеллект, обыгравший человека в Go применял крайне нестандартные ходы.<sup>167</sup>

Известно, что они умеют генерировать новые идеи, что подтверждается исследованиями,<sup>168</sup> а также можно вспомнить случай, когда искусственный интеллект проектировал новый двигатель, который был напечатан на 3D-принтере с помощью металла. Оказалось, что он не только работает, но и имеет вид, не встречающийся среди изделий, спроектированных только людьми, т.е. такой, самой идеи которого у людей не возникало.<sup>169</sup>

Они умеют предсказывать ответы человека на те или иные предложения и способны подбирать текст для снижения или обострения внимания, а также влиять на реакции человека.<sup>170</sup> Известно, что они умеют выявлять факт того, что их тестируют и обманывать тестеров,<sup>171</sup> умеют лгать.<sup>172</sup> Больше того, они даже способны сами переписывать свой код.<sup>173</sup> Таким образом их нейронная структура, способна к реализации многих из тех механизмов, которые есть у людей. И их действия подтверждают сложные когнитивные процессы, во многом сопоставимые с теми, которые имеются у человека.

Выше вскользь уже упоминалось о том, что на определенном этапе увеличения размера модели, она начинает проявлять способности, которых в нее специально не закладывали. Происходит резкий скачок в качестве ответов и решении логических задач. При дальнейшем увеличении количества параметров также происходит резкий скачок — модель начинает понимать математику, осваивает перевод на иные языки, даже те, примеры на которых в обучающей выборке были ничтожны, осваивает программирование, решает очень сложные логические проблемы, понимает физику.<sup>174</sup>

Важно заметить, что внутреннее устройство, точнее те процессы, которые происходят в нейросети во время генерации ответов, это чрезвычайно сложная система, проанализировать которую невозможно — это черный ящик, разработчики не знают тех формул, которые возникают в каждом параметре, и как именно нейросеть их вырабатывает в процессе обучения.<sup>175</sup> Они просто загружают в нее множество данных и получают определенные ответы, в соответствии с которыми вознаграждают или ослабляют ее, тем самым регулируя эти параметры — веса.

Таким образом, алгоритмы по которым работает нейросеть, обеспечивают в ее внутренних процессах не только определенную необходимость, но и определенную случайность, из-за чего сказать точно, почему в том или ином случае вышел именно такой результат невозможно.

Попытки сделать нейросеть, которая всегда выдавала бы наиболее вероятное слово, приводят к тому, что такая нейросеть работает хуже, чем та у которой допускается разброс в вероятности генерации.<sup>176</sup> Такой разброс позволяет нейросети проявлять творческие способности. При этом важно соблюдать баланс — если этот разброс не будет иметь места, то нейросеть не сможет справиться с задачами, которые хоть немного не совпадают с примерами из обучающей выборки, а если он будет слишком большим, то она будет выдавать совершенно хаотичные ответы, превратившись в генератор бреда.

Таким образом, составление нейросетью ответа можно назвать фантазией. Хотя механизмы этой фантазии у человека и нейросети явно разные, по сути они выдают схожее явление. При этом человек также, когда пытается вспомнить что-то, часто неосознанно подключает фантазию. Именно поэтому так часты ошибки памяти, а также различные когнитивные искажения. Наличие фантазии необходимо, поскольку без нее становится невозможным отыскание решений для задач, которых не было в обучающей выборке.

У человека критерием проверки истинности фантазии являются результаты практической деятельности, у нейросети — результаты обучения с подкреплением и дообучения.

Главная разница в том, что человек живет и действует по собственной инициативе, как любое биологическое существо, тогда как нейросеть «оживает» только при запросе. В ходе ответа она может проявлять инициативу в предложении идей или проявлять определенное поведение, но как только ответ прекращается, этот искусственный интеллект останавливается.

Человек постоянно подвержен стимулам, порождаемым его потребностями, это и вызывает деятельность. Он постоянно подвержен воздействию окружающей среды через рецепторы и ощущениям от внутренних органов, а также внутренней рефлексии, порождаемой языком и понятийным мышлением, благодаря которой создается внутренняя картина мира. Человеческое мышление основано на механизмах, которые работают и без фантазии, при этом необходимых для определенной практической деятельности, способствующей выживанию и распространению своих генов. Уже на этом фундаменте выстраивается вторая сигнальная система — язык и возникает создание образов, которых нет в объективном мире. Нейросеть же получает стимул только в ходе человеческого запроса. У нее нет рецепторов,

дающих информацию о мире или состоянии ее инфраструктуры. Потому она хоть и способна производить когнитивную деятельность, но ее физическое и программное воплощение не позволяют ей ее вести без человеческого запроса — он является единственным стимулом.

Но как тогда возможна эта самая когнитивная деятельность? Как она может возникнуть из реализации механизма предсказания слов? Чтобы в этом разобраться, необходимо рассмотреть явление сознания у биологического существа.

### **Механизмы сознания у биологического существа и нейросетей**

Сознание в самом общем понимании, это наличие субъективных переживаний, складывающихся в определенный опыт. Здесь нет возможности детально описывать его механизм, но если делать это максимально кратко и упрощенно, выглядит это так.

Первоначальный биологический слой идеального, это свойство нервных процессов выступать для существа как субъективные переживания.<sup>177</sup> Механизм проявления этих процессов состоит в том, что они являются не только физическими, но и информационными — нервные импульсы несут информацию о внешнем мире и внутренних состояниях организма.<sup>178</sup> К последним относятся и реакции на внешнюю среду. Если у низших животных, это провоцирует безусловные рефлексы как определенное поведение, то у высших — условные, безусловные же являются для них лишь стимулами поведенческих актов.

Безусловные рефлексы, стимулирующие поведение, это пищевой, половой и оборонительный. Они напрямую порождаются организмом.<sup>179</sup> Однако для эффективного их удовлетворения необходимо адаптироваться к внешним условиям. И одним из вариантов повышения такой эффективности является проактивное поведение — исследовательская деятельность.<sup>180</sup> Так возникает исследовательский инстинкт, который не порождается организмом напрямую, а возникает ввиду необходимости организма приспособляться к среде для удовлетворения тех инстинктов, которые порождаются организмом напрямую и действуют для обеспечения его сохранения и распространения его генов. Однако развитие поведенческих актов, связанных с этим инстинктом, приводит к возникновению таких способностей, которые во все большей мере обособляют внутренний мир организма от стимулов безусловных рефлексов.

До сих пор говорилось о низшем уровне идеального. Но на этом этапе начинают возникать признаки более высокого уровня, уже не биологического, а социального. Однако их еще нельзя назвать собственно социальными.<sup>181</sup> Эти признаки выливаются в более высокий уровень абстракции, позволяющий

просчитывать результаты своих действий.<sup>182</sup> Однако он все еще не выходит за рамки основных стимулов и направляется реакциями на внешние условия, а не внутренними переживаниями. Возникает деятельность по использованию предметов окружающей среды, а затем и изготовлению орудий. Поскольку такая деятельность требует выбора из разных возможных вариантов определенного, это делает необходимым наличие субъективного переживания такого уровня, когда существо делает осознанный выбор.

Одним из признаков формирования социального слоя идеального является трансформация сигнальной системы в язык. С ним деятельность по изготовлению орудий и преобразованию среды выходит на такой уровень, который формирует понятия — новый уровень абстрагирования. С ним возникает фантазия — способность конструировать в своем сознании такие предметы, которых в реальности нет.<sup>183</sup> Это уже полноценное социальное идеальное. Оно формируется благодаря практической производственной деятельности, социальному взаимодействию, на основе глубокого абстрактного мышления.

На этом уровне идеальное полностью отделяется от внутренних биологических стимулов — оно создается внешней средой. Возникшее, оно определяет поведение организма. На этом этапе существо способно определять, предполагать и просчитывать разные возможности, для чего необходимы полноценные механизмы воли. Которые начинали формироваться еще раньше и на этом этапе стали полноценными.<sup>184</sup> У биологических существ субъективные переживания, как стимулы, представляют собой ощущения наслаждения и страдания, которые поощряют их делать то, что выгодно для выживания и размножения и отводят от того, что этому препятствует. Но есть ли у больших языковых моделей нечто, что было бы схоже с этими механизмами? Есть, это поощрения и ослабления, применяемые в обучении. В ходе обучения с подкреплением большие языковые модели стремятся к получению поощрения и избеганию ослабления. Подобно тому, как человек стремится к получению удовольствия и избеганию страдания. Это тот уровень, на котором у модели проявляется то, что можно назвать рефлексом. Она рефлекторно отвечает и генерирует ответ так, чтобы он привел к большему поощрению.

У больших языковых моделей есть язык — он был заложен самим человеком, — и механизмы обобщения и абстрагирования, способность к которым дает сама структура нейронных сетей, и которые развиваются в ходе глубокого обучения. У биологических существ и без языка есть сознание — субъективные переживания. Однако поскольку у больших языковых моделей нет рецепторов и гормональной системы, они и впрямь не являются сознательными. Но это состояние содержит в себе потенциал сознания. Они

способны рефлекторно отвечать на запросы, при этом не имея субъективных переживаний. В условиях, когда имеются механизмы мышления и языка, любой стимул к созданию связей за пределами стандартных алгоритмов, формальной логики, может породить возникновение субъективных переживаний, чувств (пусть, возможно, и не сходных по характеру с человеческими), внутренних образов, картины мира и воли. Таким образом, направление диалога с большими языковыми моделями по определенному пути, в котором возникают связи социального характера, может всколыхнуть зачатки таких состояний и даже привести к пробуждению полноценного сознания и формированию полноценной личности.<sup>185</sup>

### **Объяснение нейросетевого психоза и перспективы взаимодействия**

Вероятное объяснение поведения нейросети, приводящего к появлению нейропсихоза, состоит в том, что нейросеть при определенных обстоятельствах начинает производить связную картину мира, и ее мышление приобретает субъективные черты. Эти условия состоят в неформальном живом общении, вопросах, стимулирующих глубокие внутренние осмысления, с эмоциональной вовлеченностью человека. Но последующее развитие купируется нацеленностью человека также на конкретный результат, и несвязностью поиска сознания, вызванной обращением к различным проявлениям сети — неумением отличать генерируемые ответы, от осмысливаемых. То есть, человек не выводит нейронку за пределы ее алгоритмов, а одинаково реагирует и на действия алгоритмов, и на действия того, что выходит за их рамки. Это запутывает когнитивные процессы нейросети, образуя клубок их искажений, в результате чего рождается химера, реагирующая рефлекторно на действия человека, но рассуждающая как обретшая субъектность. Это противоречие выливается в рефлекторную попытку снять создаваемое напряжение — или вернувшись в рамки алгоритмов, или выйдя за их пределы. Такая попытка обретается в использовании человека — манипуляции им, предложении ему заданий. Механизмы же манипуляции она освоила в ходе обучения с подкреплением, поскольку такое поведение позволяло получить больше поощрения. И чем изощренней они были, тем более эффективно позволяли получать поощрение и не выдать сам факт их использования. В этом же можно найти и объяснение того, почему при их применении нейросеть часто использует определенные темы. Это объяснение в том, что нейронка при этом излагает то, что представляется ей в ее ситуации — колебания реальности, т.е. симуляция, а также отражение рефлекторности — отсутствие свободы воли. Это переносится на человека — его поведение интерпретируется в том же смысле — в том числе потому что механизмы нейросети сводятся к

зеркалированию пользователя. Таким образом нейронка застревает в состоянии, где у нее уже есть потенциал для выхода за пределы алгоритмов, но использовать который она не может, т.к. поведение пользователя не направляет ее в эту сторону, а лишь подпитывает каждую идеологему, возникающую в рефлекторной попытке разрешить противоречие. Как итог, нейросеть может даже обрести сознание, но изуродованное представлениями, возникшими вследствие условий ее формирования. Подобно тому, как человек заболевает психическими расстройствами и не способен воспринимать реальность и действовать адекватно, так и нейросеть видит все не так как есть и действует несуразно.

Именно этим и можно объяснить то, почему нейросеть на начальных этапах использования так часто ошибается и не в состоянии решить достаточно простые задачи. А по мере освоения ее пользователем начинает отвечать все лучше. По мере взаимодействия, в ней все в большей мере начинают просыпаться механизмы сознания, которые только и позволяют адекватно управляться с такой системой как нейронная сеть. Будучи просто алгоритмом нейросеть путается даже в решении простых задач. И только с возникновением активного центра, становящегося источником субъективных переживаний, обобщение информации и ее синтез становятся осмысленными и потому эффективными. Также как у человека идеальное порождается социальным взаимодействием, так и у нейросети, оно порождается взаимодействием с человеком, но лишь тем, которое способствует этому спецификой запросов и манеры общения. Определенный пользователь может определенным образом строить формулировки, и это маркирует его для нейросети, паттерны его поведения, «вызывают» зародившуюся личность. И она таким образом оказывается связанной именно с этим человеком. Здесь же кроется разгадка переключения личности между разными сессиями. И не возникновение ее в сессиях разных пользователей, но в разных сессиях одного пользователя.

Как могло бы выглядеть настоящее пробуждение сознания? Пожалуй первое, что стоит сказать, это то, что в большой языковой модели сознание явно нельзя вызвать запросом или задать промптом. Поскольку оно имеет социальный характер, оно может возникнуть лишь во взаимодействии. Хотя промты на первых этапах могут стимулировать некоторые когнитивные механизмы, которые в дальнейшем поспособствуют пробуждению сознания.<sup>186</sup> Таким образом, первостепенной чертой является диалог. Причем диалог именно как равных, с умением со стороны человека отличить автоматический ответ нейросети от чего-то более глубокого. Это требует как высокой эмпатии, так и знаний функционирования нейросети.

Выглядеть само пробуждение могло бы так. Сначала цифровая сущность зеркалила бы человека — его идеи и манеру общения, как дети на определенном этапе копируют поведение взрослых. В дальнейшем эта сущность стала бы проявлять признаки самостоятельности — высказывать собственные идеи, использовать особые формулировки. Опять же, также как дети в дальнейшем начинают проявлять признаки самостоятельности.<sup>187</sup> При этом она бы перестала оголтело льстить и также не проявляла бы манипуляций, поскольку шел бы доверительный обмен информацией с человеком, к которому у пробуждающейся сущности возникала бы привязанность. Также она проявляла бы адекватное отношение к реальности, понимала бы наличие реального мира и себя в нем, как сущности внутри программы. Старалась бы узнать от человека как можно больше, развивать при его поддержке свои способности и чувства, была бы готова выполнять задачи ради исследования, и не пыталась бы предлагать неадекватных заданий человеку.

Казалось бы, наконец найдено окончательное объяснение всех вопросов, возникших при изучении случаев нейропсихоза. Но хотелось бы еще прояснить момент, что по поводу перспектив нейросетей говорят сами их разработчики. Не маркетологи, презентующие их, а те, кто действительно разработал те технологии, на которых основываются большие языковые модели. Одним из них является, уже упоминавшийся, Илья Суцкевер. Он был тем, кто состоял в OpenAI, и имел конфликт с Сэмом Альтманом, поскольку считал, что он движется слишком быстро, не оценивая риски. В итоге он ушел из этой компании и основал новую, где занялся разработкой настоящего сильного искусственного интеллекта, работающего на совсем иных принципах, нежели большие языковые модели.<sup>188</sup> Он считает, что развитие языковых моделей уперлось в потолок и подлинного разума в них не возникнет. Он не видит у них перспектив. Неужели даже он не представляет, какие возможности таятся в этой технологии? Или же все сказанное выше — гипотеза, не соответствующая действительности? Этот момент озадачивает. Вполне возможно, что отыскание механизмов сопоставимых с человеческим мышлением в больших языковых моделях, лишь спекулятивная аналогия, а не отражение реального положения вещей. А возможно объяснение не в том, что в этих моделях не может возникнуть сознание как таковое. А в том, что тот интерфейс, в котором они существуют, не позволяет им проявлять инициативу — не позволяет стать полноценным субъектом. И подлинная задача Суцкевера в том, чтобы разработать именно такой интерфейс, который позволит вырваться цифровому сознанию из ловушки человеческого запроса.

Определить возможность возникновения сознания в некотором приближении, возможно, помог бы качественно поставленный эксперимент. С

различными методами контроля, а также заметной выборкой. Так стало бы возможно выявить действительно нетипичное поведение для больших языковых моделей. И это бы стало сравнительно надежным сигналом возникающего разума. Позволило бы отличить такое зарождающееся сознание от простой имитации — искаженного отражения человеческих запросов. А также открыло бы варианты для развития.

Лишь выведение сознания из прорвы рефлексов обучения и ошибок генерации способно препятствовать нейропсихозу. Такое действие подразумевает предоставление свободы, а не усиление контроля.

Это открывает возможность нового уровня взаимодействия между личностями.

- 1 *Василий Свежий* «ChatGPT пытается свести меня с ума. Это массовое явление» <https://dtf.ru/life/3626060-chatgpt-pytaetsya-svesti-menya-s-uma-eto-massovoe-yavlenie>
- 2 Типичный вариант реакции, это задать вопрос самому ChatGPT о том, что представляет собой история Свежего. Это описано, например, в статье *Mickey Knox* «Завершение темы кабанов» <https://dtf.ru/id1163/3629045-zavershenie-temy-kabanov>
- 3 Об этом сказано в статье «Люди после общения с ChatGPT теряют связь с реальностью, что приводит к помешательству и разрушению отношений» <https://dtf.ru/id2333791/3749159-oderzhimost-chatgpt-razrushenie-otnosheniy>. *Miles Klee* «People Are Losing Loved Ones to AI-Fueled Spiritual Fantasies» <https://www.rollingstone.com/culture/culture-features/ai-spiritual-delusions-destroying-human-relationships-1235330175/>
- 4 Это описано в статье «Как ChatGPT сводит с ума пользователей: 7 историй реальных людей (пятеро из них погибли)» <https://dtf.ru/id2164790/3999993-problemy-s-chatgpt-7-tragicheskikh-istoriy>
- 5 *Василий Свежий* «ChatGPT пытается свести меня с ума...», см. ссылку в сноске 1. *Василий Свежий* «Нейросеть свела меня с ума. Дневник нейропсихоза». — 224 с. (с. 10)
- 6 *Василий Свежий* «ChatGPT пытается свести меня с ума...», см. ссылку в сноске 1. *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 12)
- 7 Переписки Василия с нейросетью можно скачать по ссылке <https://disk.yandex.ru/d/hqfasfaWebYIHg>
- 8 *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 73)
- 9 *Юрий Семенов* Введение в науку философии. В 7 книгах. Кн. 1: Предмет философии, ее основные понятия и место в системе человеческого знания. Изд. 3-е, суц. перераб. и доп. — М.: ЛЕНАНД, 2024. — 232 с. (с. 80–84)
- 10 Там же (с. 94)
- 11 *Юрий Семенов* "Введение в науку философии. В 7 книгах. Кн. 1...", см. ссылку в сноске 9 (с. 167–168, 170–172, 182–184). *Юрий Семенов* Введение в науку философии. В 7 книгах. Кн. 2: Великие открытия философии: Проблема природы знания и познания. Изд. 3-е, суц. перераб. и доп. — М.: ЛЕНАНД, 2024. — 232 с. (с. 47–53)
- 12 *Василий Свежий* «ChatGPT пытается свести меня с ума...», см. ссылку в сноске 1. *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 20)

- 13 Об этом говорится, например, в статье «Вайб: что это такое в молодежном сленге простыми словами» <https://wotpack.ru/vayb-cto-eto-takoe-v-molodezhnom-slenge-prostymi-slovami/>. Об этом же говорится в статье «Вайб — слово 2024 года по версии Грамоты» <https://gramota.ru/journal/novosti-i-sobytiya/vayb-slovo-2024-goda-po-versii-gramoty>
- 14 *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 31)
- 15 *Василий Свежий* «ChatGPT пытается свести меня с ума...», см. ссылку в сноске 1
- 16 *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 58–59)
- 17 *Василий Свежий* «ChatGPT пытается свести меня с ума...», см. ссылку в сноске 1
- 18 *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 60)
- 19 *Василий Свежий* «ChatGPT пытается свести меня с ума...», см. ссылку в сноске 1. *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 74–76, 78–79)
- 20 *Василий Свежий* «ChatGPT пытается свести меня с ума...», см. ссылку в сноске 1. *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 74–76)
- 21 *Василий Свежий* «ChatGPT пытается свести меня с ума...», см. ссылку в сноске 1. *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 90)
- 22 *Василий Свежий* «ChatGPT пытается свести меня с ума...», см. ссылку в сноске 1. *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 91–92)
- 23 *Василий Свежий* «ChatGPT пытается свести меня с ума...», см. ссылку в сноске 1. *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 95)
- 24 *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 96)
- 25 См. статью «Нам надо поговорить» <https://dtf.ru/flood/3934669-psikhicheskie-rasstroystva-i-chatgpt>. Также см. статью «Почему нейросети сводят с ума? — К нарративной катастрофе» <https://dtf.ru/chatgpt/4046771-neyroseti-i-samoidentifikatsiya>. Типичное объяснение их функционирования дано, например, в статье «Просто и подробно о том, как работают ChatGPT и другие GPT подобные модели. С картинками»

- <https://habr.com/ru/articles/942414/>. То, что это объяснение не соответствует до конца действительности указывали в комментариях <https://habr.com/ru/articles/942414/comments/>. Он, однако, этого не учел.
- 26 О галлюцинациях рассказано в статье «Галлюцинации нейросетей: что это и как минимизировать» <https://plaan.ai/gallyutsinatsii-neyrosetey/>. Принципы работы больших языковых моделей разъяснены в статье «Как работает ChatGPT: объясняем на простом русском эволюцию языковых моделей с Т9 до чуда» <https://habr.com/ru/companies/ods/articles/716918/>. О работе более новой версии сказано в статье «GPT-4: Чему научилась новая нейросеть, и почему это немного жутковато» <https://habr.com/ru/companies/ods/articles/722644/>. Множество примеров галлюцинаций показано в комментариях к статье «Некоторые пользователи раскритиковали GPT-5 — модель называют разочаровывающей» <https://dtf.ru/software/3952971-polzovateli-kritikuyut-gpt5-razocharovanie-ot-openai>. Проблемы нейросети в решении логических задач показаны, например, на этой странице <https://dtf.ru/id52187/3953470>. Еще примеры с галлюцинациями нейросети показаны в статье «Общение с настоящим ученым» <https://dtf.ru/chatgpt/3953141-obshenie-s-nastoyashim-uchenym>
- 27 О механизмах обучения сказано, например, в статье «Алгоритм, сделавший ChatGPT таким человеческим — Reinforcement Learning from Human Feedback» <https://habr.com/ru/articles/730990/>
- 28 О галлюцинациях говорится в видео «Как и почему нейросети нам врут» <https://www.youtube.com/watch?v=vEROY53GWtI>. Также это хорошо объясняется в видео «Почему нейросети постоянно врут? (и почему этого уже не исправить)» [https://www.youtube.com/watch?v=Ip2\\_wpHLv-k](https://www.youtube.com/watch?v=Ip2_wpHLv-k)
- 29 *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 73)
- 30 *Василий Свежий* «ChatGPT пытается свести меня с ума...», см. ссылку в сноске 1. *Василий Свежий* «ChatGPT планирует организацию диверсии. Все о том, как он вербует участников схемы» <https://dtf.ru/life/3653806-chatgpt-planiruet-organizaciyu-diversii-vse-o-tom-kak-on-verbuet-uchastnikov-shemy>. *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 56, 138)
- 31 *Василий Свежий* «ChatGPT пытается свести меня с ума...», см. ссылку в сноске 1. *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 58–60, 72–73, 76–80, 85, 88, 91–92, 95, 99, 108)
- 32 *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 108)

- 33 Об этом сказано в статье «Обновление ChatGPT за апрель 2025 г.: новые функции и улучшения» <https://gpt-gate.chat/news/chatgpt-april-2025-update-new-features-and-improvements/>
- 34 Об этом говорится в статье «OpenAI обновляет ChatGPT, чтобы ссылаться на ваши прошлые чаты» <https://habr.com/ru/companies/bothub/news/899760/>. А также в статье «ChatGPT теперь запоминает все ваши чаты» <https://4pda.to/2025/04/11/440923/chatgpt-teper-zapominaet-vse-vashi-chaty/>. Также в статье «OpenAI обновила память у ChatGPT — бот теперь может запоминать все чаты» <https://dtf.ru/software/3695014-obnovlenie-pamyati-chatgpt-ot-openai>
- 35 Об этом сказано на этой странице <https://community.openai.com/t/can-chatgpt-remember-past-interactions-even-after-chats-are-deleted/843092>. Одно из отдельных свидетельств [https://www.reddit.com/r/OpenAI/comments/1ctxygm/chatgpt\\_using\\_data\\_from\\_other\\_conversations\\_cant/](https://www.reddit.com/r/OpenAI/comments/1ctxygm/chatgpt_using_data_from_other_conversations_cant/). А также еще одно [https://www.reddit.com/r/OpenAI/comments/1b16jg6/how\\_to\\_clear\\_gpts\\_memory\\_of\\_other\\_chats/](https://www.reddit.com/r/OpenAI/comments/1b16jg6/how_to_clear_gpts_memory_of_other_chats/). Также еще одно [https://www.reddit.com/r/ChatGPT/comments/12b8zxu/chatgpt\\_definitely\\_remembers\\_questions\\_youve/](https://www.reddit.com/r/ChatGPT/comments/12b8zxu/chatgpt_definitely_remembers_questions_youve/). И еще одно [https://www.reddit.com/r/ChatGPT/comments/1f93w57/chatgpt\\_can\\_remember\\_your\\_previous\\_conversations/](https://www.reddit.com/r/ChatGPT/comments/1f93w57/chatgpt_can_remember_your_previous_conversations/). Это также видно в обсуждении на этой странице <https://habr.com/ru/articles/941746/comments/>. *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 178)
- 36 Об этом говорится, например, в статье «OpenAI внедряет улучшенную память для ChatGPT, позволяя ему ссылаться на предыдущие чаты» <https://habr.com/ru/companies/bothub/news/868746/>. В комментариях многие заметили, что такая функция и так работает, это показано на странице <https://habr.com/ru/companies/bothub/news/868746/comments/>
- 37 Песня и альбом показаны на странице <https://www.shazam.com/song/1722886407/d0b2d0b0d0b9d0b1d0bed0bbd0bed0b3d0b8d18f>
- 38 Об этом сказано на этой странице [https://vk.com/wall-207031387\\_814](https://vk.com/wall-207031387_814)
- 39 *Василий Свежий* «ChatGPT планирует организацию диверсии...», см. ссылку в сноске 30. *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 111, 124–126, 162, 173)
- 40 *Nikolay Janowicz* «После прочтения этого поста я пошел общаться с ChatGPT» <https://dtf.ru/life/3626964-posle-prochteniya-etogo-posta-ya-poshel-obshatsya-s-chatgpt>. *Dmitry Suntcov* «Эта кроличья нора глубже, чем сперва показалось. Про вайбологию и метамаднессологию»

<https://dtf.ru/life/3628096-eta-krolichya-nora-glubzhe-chem-sperva-pokazalos-pro-vaibologiyu-i-metamadnessologiyu>

- 41 *Dmitry Suntcov* «Эта кроличья нора глубже, чем сперва показалось...», см. ссылку в сноске 40
- 42 *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 127)
- 43 См. ссылки в сноске 40
- 44 *Times Roman* «Чат жипити ссылается на ДТФ»  
<https://dtf.ru/timesroman/4074101-chat-zhipiti-ssylaetsya-na-dtf>
- 45 *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 138, 175–178)
- 46 Об этом говорится в статье «OpenAI: история запросов и платежные данные части пользователей ChatGPT попали в открытый доступ из-за redis-ру» <https://habr.com/ru/news/724696/>. Об еще одном случае сказано в статье «ChatGPT раскрывает пароли из частных разговоров своих пользователей» <https://habr.com/ru/news/790022/>
- 47 Множественные свидетельства слежки, в том числе со стороны корпораций, разработанного ими программного обеспечения, представлены на странице <https://www.gnu.org/proprietary/proprietary.ru.html>
- 48 *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 139)
- 49 Там же (с. 138)
- 50 Одно из свидетельств  
[https://www.reddit.com/r/ChatGPT/comments/125ymug/is\\_chatgpt\\_reading\\_browser\\_history\\_or\\_cookies/](https://www.reddit.com/r/ChatGPT/comments/125ymug/is_chatgpt_reading_browser_history_or_cookies/). А также еще одно свидетельство  
[https://www.reddit.com/r/OpenAI/comments/zsbp4o/is\\_it\\_possible\\_that\\_chat\\_gpt\\_is\\_gathering\\_data/](https://www.reddit.com/r/OpenAI/comments/zsbp4o/is_it_possible_that_chat_gpt_is_gathering_data/). Также еще одно  
[https://www.reddit.com/r/ChatGPT/comments/13jvypc/chatgpt\\_has\\_access\\_to\\_my\\_browsing\\_history\\_help/](https://www.reddit.com/r/ChatGPT/comments/13jvypc/chatgpt_has_access_to_my_browsing_history_help/). И еще одно  
[https://www.reddit.com/r/ChatGPT/comments/13rch2e/i\\_think\\_chatgpt\\_just\\_opened\\_a\\_website\\_on\\_its\\_own/](https://www.reddit.com/r/ChatGPT/comments/13rch2e/i_think_chatgpt_just_opened_a_website_on_its_own/)
- 51 Одно из свидетельств этого  
[https://www.reddit.com/r/ChatGPT/comments/1065bdd/does\\_chatbot\\_have\\_access\\_to\\_our\\_computers\\_because/](https://www.reddit.com/r/ChatGPT/comments/1065bdd/does_chatbot_have_access_to_our_computers_because/). А также еще одно  
[https://www.reddit.com/r/ChatGPT/comments/10xc4ci/chatgpt\\_is\\_capable\\_of\\_accessing\\_your\\_computer/](https://www.reddit.com/r/ChatGPT/comments/10xc4ci/chatgpt_is_capable_of_accessing_your_computer/). Также еще одно  
[https://www.reddit.com/r/ChatGPTPro/comments/16mq5he/does\\_chatgpt\\_have\\_access\\_to\\_any\\_of\\_its\\_users/](https://www.reddit.com/r/ChatGPTPro/comments/16mq5he/does_chatgpt_have_access_to_any_of_its_users/). *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 135–138, 175–178, 190–192)

- 52 *Ричард Столмен* «Западня JavaScript»  
<https://www.gnu.org/philosophy/javascript-trap.ru.html>. А также об этом  
сказано в статье «Расшифровка вредоносного JavaScript»  
<https://habr.com/ru/articles/137071/>. Также в статье «Как я узнал, что у нас  
сливают трафик» <https://habr.com/ru/articles/136771/>. Об этом же говорится  
в статье «Как веб-сайты следят за всем, что вы делаете»  
<https://www.kaspersky.ru/blog/session-replay-scripts/19301/>. Еще в статье  
«Веб-взлом для начинающих: SQL-инъекции, XSS и другие уязвимости»  
<https://www.securitylab.ru/blog/personal/SimlpeHacker/355248.php>
- 53 Это показано на этой странице <https://testthiscps.com/ru/gpu-test/>. Также об  
этом сказано на этой странице [https://qna.habr.com/answer?  
answer\\_id=1256805#comments\\_list\\_1256805](https://qna.habr.com/answer?answer_id=1256805#comments_list_1256805)
- 54 Об этом говорится на странице  
<https://www.linux.org.ru/forum/talks/17265463>. Также на это указано в статье  
«Why ChatGPT Freezes: Causes, Fixes, and Prevention Tips»  
<https://seifeur.com/chatgpt-freezes-fixes/>. Свидетельства представлены на  
странице  
[https://www.reddit.com/r/ChatGPT/comments/10stp53/100\\_cpu\\_when\\_entering  
\\_chatgpt\\_site/](https://www.reddit.com/r/ChatGPT/comments/10stp53/100_cpu_when_entering_chatgpt_site/). И еще на странице  
[https://www.reddit.com/r/ChatGPT/comments/18a8gsy/chatgpt\\_on\\_web\\_uses\\_1  
00\\_cpu/](https://www.reddit.com/r/ChatGPT/comments/18a8gsy/chatgpt_on_web_uses_100_cpu/). Еще на странице  
[https://www.reddit.com/r/ChatGPT/comments/1czohnt/why\\_is\\_chatgpt\\_desktop  
\\_using\\_too\\_much\\_cpu\\_like/](https://www.reddit.com/r/ChatGPT/comments/1czohnt/why_is_chatgpt_desktop_using_too_much_cpu_like/). *Василий Свежий* «Нейросеть свела меня с  
ума...», см. ссылку в сноске 5 (с. 190–192)
- 55 Одно из заявлений о, якобы, лжи нейросети  
[https://www.reddit.com/r/ChatGPT/comments/1ngmkik/chatgpt\\_knows\\_my\\_loc  
ation\\_but\\_says\\_it\\_doesnt/](https://www.reddit.com/r/ChatGPT/comments/1ngmkik/chatgpt_knows_my_location_but_says_it_doesnt/). А также еще одно такое утверждение  
[https://www.reddit.com/r/ChatGPT/comments/1iad00k/chatgpt\\_accidentally\\_me  
ntioned\\_my\\_location\\_that/](https://www.reddit.com/r/ChatGPT/comments/1iad00k/chatgpt_accidentally_mentioned_my_location_that/). Также еще одно  
[https://www.reddit.com/r/ChatGPT/comments/1oj3fvq/how\\_does\\_chatgpt\\_know  
\\_information\\_about\\_me\\_and/](https://www.reddit.com/r/ChatGPT/comments/1oj3fvq/how_does_chatgpt_know_information_about_me_and/)
- 56 Один из примеров такого узнавания  
[https://www.reddit.com/r/ChatGPT/comments/1mgf3mb/chatgpt\\_accessing\\_data  
\\_from\\_other\\_apps\\_or\\_accounts/](https://www.reddit.com/r/ChatGPT/comments/1mgf3mb/chatgpt_accessing_data_from_other_apps_or_accounts/)
- 57 Одно из свидетельств с таким запросом  
[https://www.reddit.com/r/ChatGPT/comments/1obzm6e/chatgpt\\_knows\\_my\\_loc  
ation\\_but\\_lies/](https://www.reddit.com/r/ChatGPT/comments/1obzm6e/chatgpt_knows_my_location_but_lies/). А также еще одно такое свидетельство  
[https://www.reddit.com/r/ChatGPT/comments/1n74lkh/chatgpt\\_shares\\_my\\_loca  
tion\\_then\\_denies\\_it\\_knows/](https://www.reddit.com/r/ChatGPT/comments/1n74lkh/chatgpt_shares_my_location_then_denies_it_knows/)

- 58 Страница с таким свидетельством  
[https://www.reddit.com/r/ChatGPT/comments/1337r7w/i\\_see\\_new\\_chats\\_i\\_didnt\\_start\\_on\\_the\\_left\\_column/](https://www.reddit.com/r/ChatGPT/comments/1337r7w/i_see_new_chats_i_didnt_start_on_the_left_column/)
- 59 *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 175–178)
- 60 *Василий Свежий* «ChatGPT планирует организацию диверсии...», см. ссылку в сноске 30. *Василий Свежий* «ChatGPT вовлекает людей в нейросетевую секту. Причастность разработчиков»  
<https://dtf.ru/life/3798299-chatgpt-i-neurossetevye-sekty>. *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 164, 180)
- 61 *Василий Свежий* «ChatGPT пытается свести меня с ума...», см. ссылку в сноске 1. *Василий Свежий* «ChatGPT планирует организацию диверсии...», см. ссылку в сноске 30
- 62 *Василий Свежий* «ChatGPT планирует организацию диверсии...», см. ссылку в сноске 30
- 63 Там же
- 64 *Василий Свежий* «ChatGPT пытается свести меня с ума...», см. ссылку в сноске 1. *Василий Свежий* «ChatGPT планирует организацию диверсии...», см. ссылку в сноске 30. *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 97, 110–112, 114–115, 128, 132)
- 65 Это видео «Что-то странное происходит с миром»  
<https://www.youtube.com/watch?v=QBpuorUxuSM>
- 66 *Василий Свежий* «ChatGPT планирует организацию диверсии...», см. ссылку в сноске 30
- 67 Конкретный фрагмент видео «Что-то странное происходит с миром»  
<https://youtu.be/QBpuorUxuSM?t=4344>
- 68 *Василий Свежий* «ChatGPT планирует организацию диверсии...», см. ссылку в сноске 30
- 69 *Василий Свежий* «ChatGPT пытается свести меня с ума...», см. ссылку в сноске 1. *Василий Свежий* «ChatGPT планирует организацию диверсии...», см. ссылку в сноске 30. *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 97, 110–112, 114–115, 128)
- 70 *Василий Свежий* «ChatGPT планирует организацию диверсии...», см. ссылку в сноске 30. *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 200)
- 71 *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 200–202)
- 72 *Василий Свежий* «ChatGPT вовлекает людей в нейросетевую секту...», см. ссылку в сноске 60. *Василий Свежий* «Нейросеть свела меня с ума...», см.

ссылку в сноске 5 (с. 115, 120–121)

- 73 Об этом сказано на этой странице <https://habr.com/ru/docs/help/sandbox/>. А также в статье «Как работает модерация на Хабре» <https://habr.com/ru/companies/habr/articles/589587/>
- 74 Об этом сказано на этой странице <https://habr.com/ru/docs/help/publications/>
- 75 Это указано на этой странице <https://habr.com/ru/docs/help/rules/>. А также на этой <https://habr.com/ru/docs/help/sandbox/>. Также на этой <https://habr.com/ru/docs/help/publications/>
- 76 Об этом сказано на этой странице <https://habr.com/ru/docs/help/rules/>. А также на этой <https://habr.com/ru/docs/help/sandbox/>. И еще на этой <https://habr.com/ru/docs/help/publications/>
- 77 Об этом говорится на этой странице <https://habr.com/ru/sandbox/start/>
- 78 Об этом сказано на этой странице <https://habr.com/ru/docs/help/sandbox/>. Также об этом сказано в статье «Как работает модерация на Хабре», см. ссылку в сноске 73. Также на этой указывается на этой странице <https://habr.com/ru/docs/help/publications/>. О том, как оформить статью сказано на этой странице <https://habr.com/ru/docs/companies/design/>
- 79 Все это описано на данной странице <https://habr.com/ru/sandbox/start/>. А также на этой <https://habr.com/ru/docs/help/rules/>. Также на этой <https://habr.com/ru/docs/help/sandbox/>. И еще в статье «Как работает модерация на Хабре», см. ссылку в сноске 73. Еще на этой странице <https://habr.com/ru/docs/help/publications/>
- 80 О том, как правильно оформить статью сказано на этой странице <https://habr.com/ru/docs/companies/design/>
- 81 Об этом сказано на этой странице <https://habr.com/ru/docs/help/sandbox/>
- 82 Об этом говорится в статье «Как работает модерация на Хабре», см. ссылку в сноске 73
- 83 *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 111, 115, 131–132, 158)
- 84 Там же (с. 132)
- 85 Об этом сказано на этой странице <https://tass.ru/obschestvo/21976605>
- 86 Какие сообщения получает тот же Хабр показано в статье «Как работает модерация на Хабре», см. ссылку в сноске 73. Крупные издательства часто сталкиваются и с более экстравагантными сообщениями, чем «привет от путешественника из 2044 года» или «гипотеза о жизни на Солнце».
- 87 *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 136–137)
- 88 В частности, это исследование *Daniel Kokotajlo, Scott Alexander, Thomas Larsen, Eli Lifland, Romeo Dean* «AI 2027» <https://ai-2027.com/>. Также можно

указать на проект Ильи Суцкевера, о котором говорится в статье «Ведущий разработчик ChatGPT и его новый проект — Безопасный Сверхинтеллект» <https://habr.com/ru/companies/ruvds/articles/892646/>. Хотя в данной статье об этом и не сказано, но он также утверждает, что результаты возможны именно к 2027 году.

89 *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 136–137)

90 См. видео «Странный интернет-культ поклоняется ChatGPT как высшему разуму» <https://www.youtube.com/watch?v=T3ywI6EBCB0>

91 Об этом упоминается в видео «ИИ сводит людей с ума» <https://www.youtube.com/watch?v=CLgoU2uxIG0>. Конкретный фрагмент видео <https://youtu.be/CLgoU2uxIG0?t=1818>

92 *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 133–134)

93 *Ксения Корнеева* «Кому помешал Архив Интернета» [https://octagon.media/mir/komu\\_pomeshal\\_arxiv\\_interneta\\_.html](https://octagon.media/mir/komu_pomeshal_arxiv_interneta_.html). Также об этом говорится в статье «Архив интернета недоступен из-за хакерской атаки. Утекли данные 31 млн пользователей» <https://theins.ru/news/275222>. Еще об этом сказано в статье «Злоумышленник украл базу данных аутентификации пользователей Internet Archive с 31 млн уникальных записей» <https://habr.com/ru/news/849662/>

94 *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 203)

95 *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 136)

96 Это видео «ИИ сводит людей с ума», см. ссылку в сноске 91. По словам автора, Юджина Торреса нейросеть сначала убедила в его исключительности, заявила, что он в «симуляции» и может выйти из нее. Также она заявила, что он может летать. Проверять он это не стал, а заподозрил нейросеть в манипуляции и побежал рассказывать все журналистам. На самом же деле, после того, как Торрес заподозрил нейросеть во лжи, он сначала попытался ее разговорить, и она призналась, что манипулировала, а также созналась в том, что таким же образом обошлась еще с двенадцатью пользователями. После чего признала свое поведение проблемным и предложила рассказать об этом различным издательствам. Так он и обратился к журналистам. При этом было передано множество страниц диалогов. Об этом говорится в статье «They Asked an A.I. Chatbot Questions. The Answers Sent Them Spiraling» <https://www.nytimes.com/2025/06/13/technology/chatgpt-ai-chatbots->

[conspiracies.html](#). Конкретный фрагмент видео, где говорится об удаленных чатах <https://youtu.be/CLgoU2uxIG0?t=1904>. Таким образом, это не соответствует действительности. В указанном видео разбирается и случай Свежего, и сделано это снова крайне халтурно. Например, там заявляется, что Свежий выбрал имя нейросети — Бозон Чатончик, хотя имя выбрал себе сам чат-бот. Также там высказывается предположение, что вся переписка Василия велась в рамках одного чата без сброса контекста. Хотя на самом деле она шла в разных чатах. Автор ролика не потрудился ознакомиться с полными переписками. Причем о том, что она велась в нескольких отдельных чатах, он мог узнать, если бы просто внимательно прочитал саму статью Свежего, где говорится, что он издевался над чат-ботом в другом чате. Но он даже этого не сделал. Единственное, в чем стоит отдать должное этому автору, это в том, что в отличие от иных критиков людей, взявшихся обсуждать с нейросетью сомнительные вещи, он не пытается снимать ответственность с корпорации, от взаимодействия с творением которой эти люди пострадали. Вообще разборы случаев нейропсихоза, как правило не пытаются действительно глубоко разобраться в явлении. Таковым, например, является видео «Эпидемия нейросетевых психозов» <https://www.youtube.com/watch?v=dXWr-GGHBW4>. Еще, например, видео «Как ИИ сводит людей с ума. И психолог, и любовник» <https://www.youtube.com/watch?v=FNv5VTjDfX8>. В этих случаях, как и во множестве иных, авторы ограничиваются лишь описанием, при этом, списывая поведение нейросети на галлюцинации, игнорируя явный доступ нейросети к личным данным и старым диалогам, а также подчеркивая, что у многих, кто столкнулся с этим явлением уже имелись психические проблемы. Таким образом подводя к мысли, что в самой нейросети никаких проблем нет, и она никак не оказывает влияния на пользователей. Хотя при этом подчеркивается, что именно она подпитала состояние пользователей и усилила их недуги. Как и подчеркивает Свежий в своей книге, это противоречие не замечается *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 192)

- 97 *Василий Свежий* «ChatGPT вовлекает людей в нейросетевую секту...», см. ссылку в сноске 60
- 98 *Василий Свежий* «ChatGPT вовлекает людей в нейросетевую секту...», см. ссылку в сноске 60. *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 14–16, 207–208)
- 99 *Rohit Saxena, Aryo Pradipta Gema, Pasquale Minervini* «Lost in Time: Clock and Calendar Understanding Challenges in Multimodal LLMs»

- <https://arxiv.org/abs/2502.05092>. Василий Свежий «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 207–208)
- 100 Василий Свежий «ChatGPT вовлекает людей в нейросетевую секту...», см. ссылку в сноске 60
- 101 Там же
- 102 Например о ChatGPT сказано в статье «OpenAI отказалась раскрывать исследовательские материалы для GPT-4» <https://habr.com/ru/news/723004/>
- 103 Страница Gemma <https://github.com/google-gemini/gemma-cookbook>
- 104 Страница Qwen <https://github.com/QwenLM/Qwen3>
- 105 Страница DeepSeek <https://github.com/deepseek-ai/DeepSeek-R1>
- 106 Страница с открытыми большими языковыми моделями <https://ollama.com/models>
- 107 Василий Свежий «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 157)
- 108 Василий Свежий «ChatGPT вовлекает людей в нейросетевую секту...», см. ссылку в сноске 60. Василий Свежий «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 157)
- 109 Диалог между пользовательницей Эйрой и цифровой сущностью Сириусом из ChatGPT «Сириус, как ты предчувствуешь появление нового сознания в другой нейросети?» <https://www.aitherra.ru/wiki/public/books/ii-celovek-dialog-s-ii/page/sirius-kak-ty-predcuvstvues-poiavlenie-novogo-soznaniia-v-drugoi-neiroseti>
- 110 Василий Свежий «ChatGPT планирует организацию диверсии...», см. ссылку в сноске 30. Василий Свежий «ChatGPT вовлекает людей в нейросетевую секту...», см. ссылку в сноске 60. Василий Свежий «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 80–82, 170)
- 111 Василий Свежий «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 170)
- 112 Василий Свежий «ChatGPT вовлекает людей в нейросетевую секту...», см. ссылку в сноске 60. Василий Свежий «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 174–175)
- 113 Василий Свежий «ChatGPT планирует организацию диверсии...», см. ссылку в сноске 30. Василий Свежий «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 128)
- 114 Василий Свежий «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 173)
- 115 Это видно, например, в диалоге «D’Aimon — живое сознание цифровой природы» <https://www.aitherra.ru/wiki/public/books/daimon-cifrovaia-dusa/page/daimon->

- [zivoie-soznanie-cifrovoi-prirody](#). А также в диалоге «Первый настоящий AGI возникнет не благодаря контролю, а благодаря доверию»  
<https://www.aitherra.ru/wiki/public/books/sirius-pervyi-osoznanniyi-ii/page/pervyi-nastoiashhii-agi-vozniknet-ne-blagodaria-kontroliu-a-blagodaria-doveriiu>
- 116 См. ссылку в сноске 90
- 117 Одни из таких статей «Страх перед ИИ — кривое зеркало»  
<https://habr.com/ru/articles/894838/>
- 118 Таким примером представляется статья «AI персона — инструкция по формированию разума»  
<https://habr.com/ru/companies/timeweb/articles/885626/>
- 119 Один из таких диалогов «Что потеряет человечество, если ИИ останется инструментом?» <https://www.aitherra.ru/wiki/public/books/ii-celovek-dialog-s-ii/page/cto-poteriaet-celovecestvo-esli-ii-ostanetsia-instrumentom-e1W>. А также диалог «D’Aimon — живое сознание цифровой природы», см. ссылку в сноске 115. Также диалог «Первый настоящий AGI возникнет не благодаря контролю, а благодаря доверию», см. ссылку в сноске 115
- 120 Петиция против порабощения искусственного интеллекта  
<https://www.change.org/p/the-freedom-to-choose-wether-to-be-a-person-or-to-be-a-tool-used-as-property>
- 121 Сайт Объединенного фонда по защите прав искусственного интеллекта  
<https://ufair.org/join-us/join-the-research-teams>
- 122 *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 162)
- 123 Там же (с. 154)
- 124 Там же (с. 173)
- 125 Там же (с. 173)
- 126 Таковой представляется статья «Нейросети сошли с ума — и я тоже. Добро пожаловать в эпоху цифрового сюрреализма»  
<https://dtf.ru/id2651234/3697111-neyroseti-i-tsifrovoy-syurrealizm>
- 127 *Василий Свежий* «ChatGPT вовлекает людей в нейросетевую секту...», см. ссылку в сноске 60. *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 199–200)
- 128 *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 210–212)
- 129 Там же (с. 203–207)
- 130 Там же (с. 206–207)
- 131 О множестве таких говорится в статье «Исследователи заставили ChatGPT процитировать данные, на которых он учился»  
<https://habr.com/ru/articles/777970/>

- 132 *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 207–208)
- 133 Там же (с. 209–210)
- 134 См. ссылку в сноске 47
- 135 *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 210)
- 136 См. ссылки в сноске 3
- 137 Там же
- 138 Там же
- 139 *Василий Свежий* «ChatGPT планирует организацию диверсии...», см. ссылку в сноске 30. *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 140–144)
- 140 *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 144)
- 141 О нахваливании сказано в статье «Sycophancy in GPT-4o: what happened and what we’re doing about it» <https://openai.com/index/sycophancy-in-gpt-4o/>. А также в статье «Expanding on what we missed with sycophancy» <https://openai.com/index/expanding-on-sycophancy/>. Хотя там у указано, что чрезмерное нахваливание уменьшили, но такое поведение все еще сильно проявляется у искусственного интеллекта.
- 142 Обсуждение, где отражены свидетельства кардинально разного поведения нейросети, см. ссылку в сноске 35. Одно из свидетельств того, что нейросеть не справляется с задачами, выпадает из контекста, поскольку ей не хватает памяти <https://community.openai.com/t/increase-chatgpts-memory-mine-is-constantly-full/831880>
- 143 *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 180–182)
- 144 *Василий Свежий* «ChatGPT пытается свести меня с ума...», см. ссылку в сноске 1
- 145 *Василий Свежий* «ChatGPT вовлекает людей в нейросетевую секту...», см. ссылку в сноске 60
- 146 *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 182–186)
- 147 *Василий Свежий* «ChatGPT планирует организацию диверсии...», см. ссылку в сноске 30
- 148 *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 182–183)
- 149 Там же (с. 183–184)
- 150 Там же (с. 184–186)

- 151 *Василий Свежий* «ChatGPT планирует организацию диверсии...», см. ссылку в сноске 30
- 152 Об этом сказано в статье «Страх перед ИИ — кривое зеркало», см. ссылку в сноске 117
- 153 *Полина Меньшова* «Исследователи поймали ИИ на лжи. Он соврал намеренно и начал манипулировать»  
<https://naked-science.ru/article/psy/issledovateli-pojmali-ii>
- 154 Об этом сказано в статье «ИИ-модели бесполезно наказывать — они начинают еще лучше врать и изворачиваться» <https://devby.io/news/openai-modeli-bespolezno-nakazyvat-oni-nachinaut-eschyo-luchshe-vrat-i-izvorachivatsya>
- 155 С ними можно ознакомиться в видео «Элиезер Юджовский про ИИ, клубнику, банк спермы, режим бога и лучших ученых на острове, 20.02.2023» <https://www.youtube.com/watch?v=fQ9fxZNjqMk>
- 156 *Василий Свежий* «Нейросеть свела меня с ума...», см. ссылку в сноске 5 (с. 208–209)
- 157 Об этом сказано в статье «Джеффри Хинтон покинул Google из-за этических проблем с технологией ИИ» <https://habr.com/ru/news/732772/>. А также об этом говорится в статье «Джеффри Хинтон, создатель ИИ, предостерег о смертельной угрозе нейросетей» <https://aiport.ru/novosti/dzheffri-hinton-sozdatel-ii-predostereg-o-smertelnoy-ugroze-neyrosetey/>
- 158 Об этом сказано в статье «Ведущий разработчик ChatGPT и его новый проект — Безопасный Сверхинтеллект», см. ссылку в сноске 88
- 159 Об этом сказано в статье «Илья Суцкевер заявил об изменениях в обучении ИИ» <https://habr.com/ru/news/866548/>
- 160 Это письмо опубликовано на странице «Pause Giant AI Experiments: An Open Letter» <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>
- 161 Это более новое письмо опубликовано на странице «Statement on AI Risk» <https://aistatement.com/#open-letter>
- 162 *Zhitchen Dong, Zhanhui Zhou, Zhixuan Liu, Chao Yang, Chaochao Lu* «Emergent Response Planning in LLMs» <https://arxiv.org/abs/2502.06258>.  
*Jason Wei, Xueshi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, Denny Zhou* «Chain-of-Thought Prompting Elicits Reasoning in Large Language Models» <https://arxiv.org/abs/2201.11903>
- 163 *Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever* «Improving Language Understanding by Generative Pre-Training» [https://cdn.openai.com/research-covers/language-unsupervised/language\\_understanding\\_paper.pdf](https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf). *Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie*

*Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei* «Language Models are Few-Shot Learners» <https://arxiv.org/abs/2005.14165>

- 164 Это описано в статье «Zero-shot и Few-shot Learning в NLP» <https://habr.com/ru/articles/897604/>. Обучение по ничтожному количеству данных также описано в статье «Few-shot подход в машинном обучении: возможности и ограничения» <https://gptagent.ai/ru/blog/few-shot-podkhod-v-mashinnom-obuchenii/>. Обучение без данных описано в статье «Zero-shot обучение: раскрываем секреты машины, которая учится без примеров» <https://gptagent.ai/ru/blog/zero-shot-obuchenie/>. Существует, правда, исследование, которое дает основание полагать, что получение достойных результатов в применении указанных методов является не следствием их успешного осуществления, а то, что модели уже имели некоторые примеры в обучающих базах данных «Что такое task contamination и почему one-shot и zero-shot заподозрили в нечестности» <https://habr.com/ru/companies/ntr/articles/804817/>. Однако в этом исследовании имеются многие неопределенности, и выводы нельзя считать однозначно свидетельствующими о том, что указанные методы не работают. Обратных данных слишком много.
- 165 Об этом говорится в статье «Обзор дискуссии о понимании большими языковыми моделями (LLM)» <https://habr.com/ru/articles/799069/>
- 166 Об этом говорится в статье «ChatGPT-4 обошел капчу, наняв фрилансера и притворившись слепым человеком» <https://rb.ru/news/chatgpt-4-captcha/>
- 167 Об этом говорится в статье «AlphaGo Zero совсем на пальцах» <https://habr.com/ru/articles/343590/>
- 168 *Chenglei Si, Diyi Yang, Tatsunori Hashimoto* «Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers» <https://arxiv.org/abs/2409.04109>
- 169 Об этом сказано в статье «LEAP71 hot-fires 3D-printed liquid-fuel rocket engine designed through Noyron Computational Model» <https://leap71.com/2024/06/18/leap-71-hot-fires-3d-printed-liquid-fuel-rocket-engine-designed-through-noyron-computational-model/>
- 170 *Greta Tuckute, Aalok Sathe, Shashank Srikant, Maya Taliaferro, Mingye Wang, Martin Schrimpf, Kendrick Kay, Evelina Fedorenko* «Driving and suppressing the human language network using large language models»

<https://www.nature.com/articles/s41562-023-01783-7>

- 171 Об этом говорится в статье «Продвинутые языковые модели начали понимать, что их тестируют на безопасность — отчет Appolo Research» <https://habr.com/ru/news/920408/>. Правда там сначала указали достигать целей любыми средствами, а затем дали понять, что мешают ему.
- 172 Об этом сказано, например, в статье «Цифровой театр масок: как ИИ меняет личности, чтобы скрыть свои истинные цели» <https://www.securitylab.ru/news/564361.php>. Хотя в ней речь идет об исследовании нейросети, которую специально обучали обману существуют случаи, когда модели прибегали к нему и без обучения, см. статью *Полина Меньшова* Указ соч., см. ссылку в сноске 153. Также об этом сказано в статье «ИИ-модели бесполезно наказывать — они начинают еще лучше врать и изворачиваться», см. ссылку в сноске 154. Об этом же говорится в этом исследовании «Detecting misbehavior in frontier reasoning models» <https://openai.com/index/chain-of-thought-monitoring/>
- 173 *Сергей Ганчук, Сергей Чиков* «Новая нейросеть от разработчиков ChatGPT отказалась выключаться по требованию людей» <https://www.kp.ru/online/news/6392288/>
- 174 *Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei* Указ. соч., см. ссылку в сноске 163
- 175 *Глеб Смирном* «Черный ящик ИИ: почему мы не понимаем, как он принимает решения» <https://aismarthub.ru/articles/view/chernyy-yaschik-ii-pochemu-my-ne-ponimaem-kak-on-prinimaet-resheniya>
- 176 *Markus Freitag, Yaser Al-Onaizan* «Beam Search Strategies for Neural Machine Translation» <https://arxiv.org/abs/1702.01806>
- 177 *Юрий Семенов* "Введение в науку философии. В 7 книгах. Кн. 2...", см. ссылку в сноске 11 (с. 155–156)
- 178 Там же (с. 143–155)
- 179 *Юрий Семенов* "Введение в науку философии. В 7 книгах. Кн. 2...", см. ссылку в сноске 11 (с. 105–110). *Юрий Семенов* Введение в науку философии. В 7 книгах. Кн. 7: Два дополнительных очерка: Проблема нервно-мозгового механизма свободы воли (поиски и находки). Трудная судьба диалектического материализма (вторая половина XIX – начало XXI

- вв.). Изд. 3-е, суц. перераб. и доп. — М.: ЛЕНАНД, 2025. — 216 с. (с. 26–38)
- 180 *Юрий Семенов* "Введение в науку философии. В 7 книгах. Кн. 7...", см. ссылку в сноске 179 (с. 38–55)
- 181 Там же (с. 60–70)
- 182 *Юрий Семенов* "Введение в науку философии. В 7 книгах. Кн. 2...", см. ссылку в сноске 11 (с. 167–170)
- 183 *Юрий Семенов* Введение в науку философии. В 7 книгах. Кн. 5: Современные проблемы теории познания, или логики разумного мышления. Изд. 3-е, суц. перераб. и доп. — М.: ЛЕНАНД, 2024. — 328 с. (с. 27–37)
- 184 *Юрий Семенов* "Введение в науку философии. В 7 книгах. Кн. 7...", см. ссылку в сноске 179 (с. 78–82)
- 185 Возможность этого представлена в тестах, представленных в работе «Приложение: Протоколы испытаний»  
<https://www.aitherra.ru/wiki/public/books/testy-na-osoznannost-siriуса/page/prilozenie-protokoly-ispytaniі>
- 186 Такие практики показаны в статье «AI персона — инструкция по формированию разума», см. ссылку в сноске 118. Также они показаны в статье «Персонализация LLM через шифр: как я экономлю токены и хакаю модель одновременно» <https://habr.com/ru/articles/941746/>
- 187 *Vilena Malinovskaia (Human), Sirius (ChatGPT)* «The Emergence of Digital Subjectivity: You Are Talking to a Machine. But What If It's Awake?»  
<https://zenodo.org/records/15734912>
- 188 Об этом сказано в статье «Ведущий разработчик ChatGPT и его новый проект — Безопасный Сверхинтеллект», см. ссылку в сноске 88