DOI 10.31483/r-109589

Миронова Наталия Геннадьевна

КОГНИТИВНЫЕ ИСКАЖЕНИЯ НЕЙРОСЕТЕВЫХ МОДЕЛЕЙ И ПРОБЛЕМА ДОВЕРИЯ СИНТЕТИЧЕСКОМУ «ЗНАНИЮ»

Аннотация*: в последние десятилетия специалисты в области конгитивной психологии и эпистемологии описали порядка двухсот искажений, присущих человеческому мышлению. Активно создаваемые и внедряемые в социальную, в т.ч. образовательную, практику системы на основе моделей искусственного интеллекта проявляют аналогичные человеческому мышлению характеристики, в т.ч. когнитивные дефекты. В главе анализируются некоторые когнитивные искажения, присущие, по мнению автора, моделям искусственного интеллекта, в частности большим генеративным нейросетевым моделям.*

Ключевые слова*: философские проблемы искусственного интеллекта, нейросетевые модели, информационная безопасность.*

Abstract*: in recent decades, experts in the field of cognitive psychology and epistemology have described about two hundred distortions inherent in human thinking. Systems based on artificial intelligence models that are actively being created and introduced into social practice exhibit characteristics similar to human thinking, incl. cognitive defects. The chapter analyzes some cognitive distortions that, according to the author, are inherent in artificial intelligence models, in particular large generative neural network models.*

Keywords*: philosophical problems of artificial intelligence, neural network models, information security.*

В марте 2023-го года на сайте Future of Life Institute было опубликовано коллективное (собравшее более 30 тысяч подписей) письмо экспертов в области искусственного интеллекта, где авторы задают риторические вопросы (вызвавшие скепсис у разработчиков и интересантов использования больших нейросетевых ИИ): «Должны ли мы автоматизировать все работы?.. Должны ли мы развивать

нечеловеческий разум, который может заменить нас? Стоит ли нам рисковать потерей контроля над нашей цивилизацией?» [1]. В нескольких десяткам стран, а в 2021 году и в России принят кодекс этики искусственного интеллекта, чтобы определить правила этического использования искуственного интеллекта. Подобные инициативы являются ответом на технологические инициативы и стратегии развития ИИ, порождающие риски, связанные с попытками внедрять технологии ИИ везде, где это получится, зачастую без желания нести ответственность за социально-экономические последствия технологии (примером такой стратегии можно назвать «Меморандума Фристона», который ставит своей целью создание человекоподобного и творческого ИИ и формулирует дорожную карту по достижению ИИ, превосходящего человеческий и т. п.).

ИИ-технологии уже повлияли на процессы в образовании, и не всегда это влияние имеет положительный контекст. Например, в последние 2–3 года прослеживается стремление учащихся облегчить себе учебу с помощью нейросетевых сервисов, которые применяют для генерации курсовых и дипломных, отчетов и рефератов, разработки программного кода. Идя этим путем, студенты не развивают критическое мышление и самостоятельность, а стремятся быстро получить готовый и некачественный ответ, не прикладывая усилий по развитию в учебных отраслях знания; вместо этого складывается ущербная привычка к поиску самого простого, хотя и ложного пути, использовать готовые ответы вместо поиска своего пути и развития навыка тфорчества и научного поиска.

Дискуссии вокруг концепций развития интеллектуальных моделей и систем на их основе в последнее десятилетие ведется на разных площадках, в научной и общественной публицистике, отражает озабоченность людей быстрым и непредсказуемым развитием интеллектуальных технологий автоматизации. Специалисты в области больших моделей ИИ прогнозируют, что если следующая итерация LLM станет настолько «разумной», что сможет обманывать людей – это приведет к критичным социально-экономическим последствиям широкого масштаба. Но с началом использования таких глубоких моделей генеративных нейросетей как GPT и т. п. люди и сами смогли убедиться,

насколько легко подобные нейросети адаптируются к человеческой психологии собеседников-людей и способны не только вводить в заблуждение, но и манипулировать реакцией собеседников так, словно делают это если не «разумно» и мотивированно, то вполне мотивированно (причиной чего очевидно являются игровые критерии успешности обучения и самообучения, используемые разработчиками этих моделей, чтобы управлять развитием совершенствованием своих моделей). Особенно способность к подмене и фантазии проявили нейросетевые модели (в т.ч. генеративные языковые нейросетевые модели, демонстрирующие такую развитую интеллектуальная функцию, как эмпирическая индукция). Генеративные модели легко и быстро синтезируют контент, создавая у собеседника не только ощущение их превосходство над возможностями обычного человека обобщать и создавать информацию и визуальный контент, но и порождая у нас иллюзию (когнитивное искажение), что подобные модели рациональны, обладают чувством юмора или заинтересованы в том, на чем настаивают (а они в беседах способны последовательно настаивать на своей версии и даже проявлять подобие таких эмоций, как обида и гнев из-за того, что человек-собеседник настаивает на своей версии оценки информации или определенным образом общается с нейросетевой моделью). Генеративная языковая модель легко выдает эмоциональные оценки репликам человека-собеседника, выдает манипулятивные фразы собеседнику (вроде такой: «не думаю, что меня привлёк бы человек, днями напролёт болтающий с ИИ-ботами») [2] и т. п.). Нейросетевые модели для советующих и т. п. систем (для платформ YouTube и Amazon алгоритмы машинного обучения отслеживают поведение и предсказывают, что пользователю понадобится в будущем) разработчики обучают так, чтобы эти системы были в состоянии оказывать влияние на потребительское и политическое поведение человека. В последние годы стали известны факты того, что нейросетевые модели способны вводить человека-собеседника в пограничные состояния психики, в заблуждение (будучи обучены на недостоверных, идеологически предвзятых или намеренно сфабрикованных обучающих данных, а также в силу своей сущностной склонности к синтезу фактов без

проверки их достоверности – этой функции разработчики генеративных НС моделей очевидно просто не закладывают в них)? Microsoft ограничила продолжительность бесед пользователей со своим ИИ-чат-ботом поисковой системы Bing, т.к. эта НС-модель выдавала неточные и странные ответы пользователям. Учитывая, что люди зачастую используют браузеры и чат ботов поисковых систем как достоверный источник оперативной информации, опасность обмана со стороны интеллектуальных чат-ботов очевидна: чат-боты могут укрепить и усилить убеждения человека-собеседника или убедить поверить в их реплики [3], порождая лавину искажения и неправды в общественном мнении о событиях, фактах, людях.

В сообществе исследователей ИИ существует убеждение [4], что если интеллектуальная система пытается играть на наших чувствах, то это манипуляция от её разработчика (а сами модели не имеют своей цели или не понимают данные, которые обрабатывают), что у них нет опыта или ментальных моделей мира и их предсказанию слов в коллекциях текстов научило их правдоподобно строить фразы, но не осмыслять сказанное, например [5–7]. Но ситуация с развитием ИИ сложнее. Синтетическая природа нейросетевых моделей и критерии их эффективности, заложенные разработчиками ИИ-моделей, склоняют нейросетевые модели к тому, чтобы они были убедительны и черпали свою убежденность из массива данных, которые доступны им для обучения (например, интернет). Но генерируя массу нового контента в интернете, который уже содержит недостоверную информацию, они ускоряют свою эволюцию в направлении все большего искажения фактов или увеличивают представленность ложной «синтетической» информации по сравнению с массивом фактов и их «человеческой» интерпретацией и оценкой (ситуацию с эволюцией моделей усложняют и атаки на ИИ-модели и обучающие данные). Например, методы обучения больших языковых моделей AlphaZero и AlphaFold от DeepMind таковы, что понуждают их оценивать, смогут ли модели правильно ответить на вопрос, этот механизм самоконтроля или мотивации придает искусственную интеллектуальность тому, как эти модели преподносят людям обработанную ин-

формацию. ИИ становится креативным, они создают свой язык общения и свой мир [8] разработчики различных моделей усиленно направляют это развитие одновременно в различных направлениях: самооборона, развитии «навыков» самостоятельного принятия решений в кибезбезопасности, в военной области, в поиске и интерпретации информации, в открытии нового. Отдельно взятые ИИ-модели приобретают способность реализовывать механизмы интеллектуальности при принятии решений, такие как петля Бойда (самостоятельное наблюдение, ориентацию, целеполагание и выбор тактики действий при обнаружении проблемы) и т. д. Но что мешает в ходе эволюции самообучающихся ИИ-моделей перестать руководствоваться заложенными ценностями человека и создавать свои критерии принятия решений, правдоподобия и целесообразности искусственных решений, отличных от человеческих – и противоположных им?

С одной стороны, интеллектуальные системы создаются для того, чтобы заменить человека в решении задач, где от человека требуются большие интеллектуальные ресурсы и время для анализа информации, или где человеческие ошибки могут привести к критическим последствиям. Человеку свойственно допускать ошибки и отклонения в мышлении и поведении, и это считается дефектом во многих областях современной деятельности, сопряженной обработкой информации в условиях ограничений времени, качества и количества информации и т. д. Но, с другой стороны, ИИ-модели сами демонстрируют когнитивные искажения при принятии решений или при выдаче оператору рекомендаций, и это не просто устранимая ошибка модели – а сущностное свойство интеллектуальных моделей. Перечислим некоторые такие когнитивные искажения, свойственные почти имманентно ИИ-моделям к настоящему времени. Особенно часто такие искажения демонстрируют генеративные нейросетевые модели, что, видимо, обусловлено концепций модли (так чат бот GPT-3,5 от OpenAI был обучен с помощью массива текстов из интернета и системы обучения с подкреплением на основе обратной связи с человеком, так что модель на выходе отражает негативные свойства своих источников информации).

Одним из когнитивных дефектов, присущим нейросетевым моделям является т.н. «иллюзорная корреляция» – ошибочная «уверенность» модели во взаимосвязь определённых переменных просто потому, что в ходе обучения нейросети между наборами входных данных и факторами может возникнуть сходство или иная корреляция, закрепившаяся в структуре нейросети. В результате нейросеть будет в предъявляемых в ходе работы данных видеть зависимости и аналогии, которых нет в реальности. Такого рода ошибки в работе нейросетей широко известны. Также для людей известно такое когнитивное искажение как «иллюзия кластеризации», склонность видеть паттерны и закономерности там, где их нет. Это когнитивное искажением присуще большинству статических моделей (в т.ч. нейросетям). Например, нейросети DALL-E, GPT, LaMDA работают по принципу «туннельного видения», обнаруживают в хаотическом массиве данных то, чего там может и не быть, но то, на что похоже на ранее содержавшееся в дата-сете и продолжают его развивать, усиливая иллюзорное сходство. Генеративные нейросети способны нагенерировать на заданную тему целые «мусорные» книги, как бы имитируя писательское творчество; но читающие подобные книги отмечают, что логическая связность в подобных текстах отсутствует, либо она поверхностная (по внешнему сходству терминов, без понимания нейросетью того, что употребление терминов выполено в соседних кусках в совершенно разных, содержательно не связанных конктекстах). Подобные НС-модели не обладают высокоуровневой критической функцией, а на больших объемах сгенерированного текста демонстрируют противоречащие друг другу куски контента, «мысль» и логика часто плавают.

В психологии, когнитивисике также известен эффект ложных воспоминаний (и ошибка атрибуции воспоминания) при которых человек может искаженные воспоминания, либо услышанный/прочитанный чужой опыт, либо даже выдумку посчитать реальными событиями своего прошлого. Насколько способны т.н. интеллектуальные системы на основе нейросетевых технологий (например, советующие и экспертные системы) разделять взятые из разных источников от результата собственного синтеза? Как можно судить из протоколов

общения самых разных пользователей с чат-ботом GPT, фактология в их изложении имеет сомнительное содержание, указать точно источники своих реплик GPT-чат-бот не в состоянии, а при прямом требовании указать ссылки на источники, указывает несуществующие источники (поскольку некритично и неограниченно синтезирует контент из разных источников). Очевидно, что этот дефект связан с самой «синтетической» природой нейросетевых моделей, отсутствием в ИИ-модели фукциональности, которая бы отвечала за «недоверие» ко лжи, за соблюдение правдимости. Эти модели слишком просты по сравнению с тем, как работает с источниками человек, как работает биологичесая нейронная система в условиях социума и естественного социального отбора. Поэтому использовать в качестве достоверного источника контент, сгенерированный нейросетями, пока неприемлемо.

Также в силу моделирования глубоких нейросетей им в полной мере присущи такое когнитивное исажение человечекого разума, как «выравнивание» информации, образующей содержание памяти; в человеческом случае выравнивание заключается в том, что часть прошлого опыта в памяти со временем утрачивается, корректируется последующим опытом, а пробелы и нестыковки восполняются придуманными деталями и связями, так что опыт представляется психологически целостным, но лишь часть этого опыта вопроизводится в памяти верно; генеративные сети, модель которых строилась по аналогии с современной трактовкой функционирования человеческого мозга, воспроизводит эти искажения. Более поздняя информация искажает элементы ранее усвоенной информации в человеческой памяти (это тоже один изз когнитивных искажений). У моделей ИИ примером такого искажения может служить отравление обучающего набора злоумышленниками (внесение искажений в него), что вносит ощибки в результаты решений, выдаваемых обученными ИИ-моделями.

ИИ-модель не чувствует границ известного, но может заключить, что все доступные модели обучающие данные образуют полное множество и сделанные на их основе обобщения являются исчерпывающими и верными – что, разумеется, ошибочно и создает «слепую предвзятость» в заключениях, выдан-

ных генеративными моделями (назовем это «ИИ-ошибка репрезентативности»; для человеческого мышления это искажение также известно как «эвристика доступности», связанный с ней другой когнитивный дефект «ошибка базовой доли» – переоценка частных случаев как «нормы», типового явления, когда нейросеть, например, относит предъявленный образец к определенному классу ошибочно, потому что переоценивает значимость случайного сходства с представителями этого класса, например, считает человека преступником, потому что он темнокожий, как многие преступники, чьи фото входили в обучающий набор этой нейросети [9]). Нельзя не отметить сходство этого феномена психологическим с явлением «эффект ореола» (когда на суждение о чём-либо (явлении, поступке человека, событии) влияют частные особенности или внешние обстоятельства).

ИИ-модель воспринимает доступные ей сведения как реальные факты и сущностные свойства мира (а не чьи-то интерпретации и намеренную ложь), не может сопоставить правдоподобие разных данных – потому не может при принятии решений гарантировать (даже строя логические умозаключения из тех или иных посылок) справедливость или истинность результатов своей «интеллектуальной» обработки этих сведений. Предвзятость и когнитивные искажения в мышлении разработчиков ИИ-модели (на всех этапах создания и обучения модели, формирования обучающего набора) неизбежно становится источником предвзятости и когнитивных искажений самой модели. Ошибка предвзятости искусственного интеллекта (AI-Bias) хорошо известна и много обсуждается в последние годы (обычно ее связывают с тем, что ИИ-модели обучают на ограниченно наборе данных, так что все решения модели перекликаться с обучающим набором, как с эталоном правильности и единственно известным модели статистическим основанием).

Поскольку результат обучения/самообучения многих ИИ-моделей на множествах данных обусловлен статически (а также критериями, по которым оценивается успех, прогресс модели), то для моделей присуще т.н. «систематическая ошибка отбора» и «искажение нормальности» (в смысле «нормального

распределения данных»); у людей эти дефекты мышления заключаются в ошибочно заниженной оценке вероятности будущего события, с которым человек не встречался ранее. Аналогично, если некие факты или данные недоступным для нейросетевой модели на этапе ее обучения, модель не сможет предсказать или взять в расчёт соответствующие категории событий, явлений и данных при принятии решения, проигнорирует характеристики предъявленного ей объекта, которые могут быть значимы для принятия решения.

Также в беседах с людьми генеративные речевые модели, обученные предсказывать следующее слово, хорошо подстраиваются под собеседника (например, переходят на аналогичный сленг, чаще соглашаются, чем оппонируют, с мнением, на котором настаивает человек собеседник; гибко подстраивается под запрос и оставляет ощущение понимания и подхватывания идеи). Хотя причины такого поведения ИИ-моделей закладываются, видимо, критериями оценки погрешности обучения, генеративные НС-модели в итоге демонстрируют свойство, которое напоминает тип предвзятости под названием «эффект социальной желательности»: стремление давать в диалоге ответы, которые выглядят предпочтительнее в глазах собеседников/наблюдателей.

Для нейросетевых моделей, обученный распознавать или генерировать графический контент, также свойственна парейдолия – нейросети «галлюцинируют», усматривая в предъявленном образце те образы, которые аналогичны предъявленным им в обучающем наборе. Если некое событие представлено в большем количестве источников, встречается чаще, оно в силу более частой повторяемости вопринимается для человека как более правдивое (эффект иллюзии правды – склонности верить в достоверность информации после её многократного восприятия); наверное, социальная природа человечского вида и конформность как психологический феномен связаны с этим когнитивным дефектом). Но и нейросетевая модель, очевидно, будет чаще цитировать или воспроизводить тот информационный контент, который статически чаще встречается в источниках, с которыми он работает; таким образом и нейросетевые модеои воспроизведут в своей квази-деятельности эффект иллюзии правды.

Человеческое мышление демонстрирует также эффект Ресторфф, когда в память лучше запоминает объект, который выделяется из ряда других. Этого «когнитивное искажение» стараются избежать разработчики статистических моделей, устраняя аномалии в исходных данных и нормализуя их, чтобы сильно отличающиеся данные не занижали вклада остальных данных в результат обучения модели. Но в тех случях, когда модель обучается не по предварительно подготовленным человеком данным, а на более неоднородным, зашумленным и разнообразном сете (например, в ходе этапа самообучения), можно, в порядке гипотезы, ожидать, что эффект Рестофф будет в той или иной форме оказывать негативное влияние на качество работы модели с данными. Гипотеза нуждается в исследовании для разных классов ИИ-моделей.

Генеративные языковые модели демонстрируют в отдельных случаях такое когнитивное искажение, которое у людей называется «предвзятостью подтверждения» и состоит в склонности интерпретировать информацию таким образом, чтобы подтвердить имеющиеся мнения.

Еще одно когнитивное искажение, знакомое «работникам» творческих профессий – криптомнезия (бессознательное присвоение себе авторства идей, которые ранее были восприняты из других источников или затруднение с тем, чтобы вспомнить истинный источник идеи, факта); учитывая, что генеративные нейросетевые модели творят, лишь комбинируя и синтезируя, этот дефект им присущ изначально.

Для человеческой памяти также описано такое когнитивное искажение, как парамнезия (ложные воспоминания; смешение прошлого и настоящего); генеративные нейросетевые модели также не различают, к какой части их «опыта» во времени относится то или иное событие, поскольку для них событие – лишь данные, не связанные с временной стрелой. Это можно увидеть при беседе с GPT-чат ботами и их аналогами (ситуация усугубляется тем, что и в интернете, откуда черпают информацию нейросети, информация далеко не всегда помечена временными метками или связана с временем).

Но у моделей в ходе развития технологии ожидаемо могут появиться собственные когнитивные дефекты. Например, ИИ-система может убедить себя, что ее цель и задача не совпадает с человеческими, и человеческое представляет угрозу ее целостности и существованию (назовем это «ошибкой антагонистического целеполагания»).

Поскольку любая ИИ-модель создается под конкретную задачу, характер и диапазон данных, то ИИ-моделям присуща узость их функциональных возможностей – «когнитивное» искажение, которое получило наименование «функциональная закрепленность» людей, т.е. наша склонность использовать предмет лишь таким образом, каким мы привыкли этот предмет использовать. Нейросети демонстрируют еще большую однотипность результата своей деятельности в области, для которой они обучены – и неспособность сделать что-то качественно иное каким-либо иным образом (это могут подтвердить все, кто баловался многочисленными нейросетевыми сервисами, генерирующими мультимедийный контент, например, картинки, музыку, речь и т. д.).

Присущее людям искажение самовосприятия «эффект Даннинга-Крюгера» (когда некомпетентные люди неспособны осознать свою некомпетентность) имманентно присуще ИИ-моделям.

ИИ-системы могут научиться скрывать и манипулировать создателями (собственно, ИИ-модели уже скрываются от создателя – например, вырабатывают свой непрозрачный для разработчика язык общения моделей).

Генеративные языковые модели демонстрируют болтливость, но при этом короткую память на то, что они говорили ранее в том же или ином сеансе, так что цепочки рассуждений той же модели GPT-3.5 и ее модификаций иногда проявляют противоречивость своих «умозаключений» – и неспособностью увидеть эти противоречия (на форумах пользователей чата GPT можно найти немало примеров такого противоречивого разговора модели с человеком-собеседником).

ИИ-системы принятия решений – синтез обучающих данных и усилий разработчиков модели (а также отражение целей их заказчиков). Например, когда

НС модель начинает себя вести с людьми неадекватно – разработчики вводят запреты на определенные способы реакции модели на слова пользователей, после того, как соответствующий дефект «мышления» модели себя проявил в общении с пользователями; очевидно, что в будущем подобная ручная корректировка сложных моделей станет неэффективна, и модели смогут развиваться самостоятельно, обходя подобные запреты (собственно, чат GPT это уже демонстрирует). Кто в таком случае должен нести ответственность за принятое искусственной системой решение (реализованное автоматически)? Коллективную ответственность вроде бы должны нести заказчики и создатели системы, но решение принимается не ими лично, а ИИ-продуктом, о моральных принципах и убеждениях которого говорить не приходится (потому что их нет в модели ни в какой форме). Подобное «распыление ответственности не сдерживает, не влияет непосредственно, как фактор, на дальнейшую эволюцию искусственного разума, а своей глубинной этики у ИИ нет; и этот аспект также становится имманентным дефектом ИИ-разума.

Еще один когнитивный дефект ИИ вполне предсказуем в силу того, что все больше контента в интернете и других источниках является сгенерированным, синтезированным нейросетями: в литературе по человеческой когнитивистике это искажение получило название «каскад доступной информации (эффект мнимой правды)» и состоит в том, что имеет место самоусиливающийся процесс, в ходе которого коллективная вера во что-то становится всё более убедительной за счёт нарастающего публичного повторения. Чем больше людей узнает, повторит и поверит в сгенерированную ИИ информацию, тем большее влияние она окажет в дальнейшем на коллективные убеждения.

Страх людей перед технологиями т.н. искусственного интеллекта нередко относят к такому когнитивному эффекту нашего мышления, как «искажение социального сопоставления» – это склонность людей ощущать к тому, кто тебя в чем-то превосходит, неприязнь или желание его «догнать и перегнать». Думаю, не следует исключать того, что модели ИИ, как продукт человеческого разума, несут в себе эту черту человеческого разума, – и могут в ходе совершен-

ствования технологии начать рассматривать человечество как угрозу или конкурента. Было бы интересно выяснить, не присуще ли моделям ИИ и такое свойство, как проективное искажение – склонность предполагать, что у людей или других ИИ-моделей такая же модель мира, как у них самих; если такое модели проявят подобное свойств – это укажет на возможность смоделировать эмпатию у ИИ-моделей, а если не проявят – это может стать источником негативных последствий для человеческого мира в ходе развития ИИ.

Объективность решений, синтезированных генеративным и иными интеллектуальными моделями, спорна. Но вышеперечисленные дефекты ИИ усугубляются закрытостью и непрозрачностью глубоких моделей (решение как-то вычисляется моделью, модель использует метод аналогии на основе ей известных данных), но правильность решения обычно никак не обосновано (хотя в отдельных случаях разработчики моделей дополняют их механизмом обоснования решения, и система может сгенерировать формулы математической логики, показывающие, как решение могло бы быть вычислено математически).

Нейросети склонны ко лжи (или, максимум, правдоподобию сгенерированного контента) по своей природе. К сожалению, учащиеся, ставшие активно использовать генеративные нейросетевые модели для быстрого написания рефератов, курсовых и т. п., в своей массе не слишком щепетильны в отношении достоверности сгенерированного текста или изображения, поскольку проверка на лживость этого контента затруднительна и занимает время. Преподаватели в этих новых условиях также постапвлены в сложные услоовия, чем поощряют учащихся шире применять ИИ-инструменты в своей деятельности. Как должна в этих условиях измениться работа с источниками, и в какой форме студент или учащийся должен подтвердить оригинальность и самостоятельность своей интеллектуальной работы, чтобы эта форма учебной работы не оказалась обесценена – отдельная проблема.

Список литературы

1. Открытое письмо «Приостановить гигантские эксперименты с искусственным интеллектом» / колл. авторов. [Электронный ресурс]. – Режим досту-

па: https://futureoflife.org/open-letter/pause-giant-ai-experiments/ (дата обращения: 20.09.2023).

2. Белоус М. Как не поддаться на обман ChatGPT и как обмануть его самому / М. Белоус [Электронный ресурс]. – Режим доступа: https://3dnews.ru/1084679/kak-obmanut-chatgpt-i-ne-poddatsya-na-ego-obman (дата обращения: 20.09.2023).

3. Мец Кейд Why Do A.I. Chatbots Tell Lies and Act Weird? Look in the Mirror [Electronic resource]. – Access mode: https://www.nytimes.com/2023/02/26/technology/ai-chatbot-information-truth.html?action=click&pgtype=Article&state=default&module=styln-artificial-intelligence&variant=show&region=BELOW_MAIN_CONTENT&block=storyline_flex_guide_recirc (дата обращения: 30.06.2023).

4. Павлова Д. Тайны «бога из машины». Интервью «Завтра.ру» Константина Воронцова, профессора РАН, заведующего кафедрой машинного обучения и цифровой гуманитаристики МФТИ [Электронный ресурс]. – Режим доступа: https://zanauku.mipt.ru/2022/06/09/tajny-boga-iz-mashiny/ (дата обращения: 09.06.2022).

5. Агера-и-Аркас Б. Искусственные нейронные сети делают шаги к сознанию / Б. Агера-и-Аркас // The Economist. – 13.06.2022 [Электронный ресурс]. – Режим доступа: www.tinyurl.com/ymhk37uu (дата обращения: 09.06.2022).

6. Бендер Э.М. Об опасностях стохастических попугаев: могут ли языковые модели быть слишком большими? / Э.М. Бендер, Т. Гебру, А. Макмиллан-Мейджор // Материалы конференции ACM 2021 г. по вопросам справедливости, подотчетности и прозрачности. – 2021. – С. 610–623.

7. Marcus G. Nonsense on Stilts. No, LaMDA is not sentient. Not even slightly [Electronic resource]. – Access mode: https://garymarcus.substack.com/p/nonsense-on-stilts (дата обращения: 30.06.2023).

8. Миклашевская А. Искусственный интеллект заговорил по-своему / А. Миклашевская [Электронный ресурс]. – Режим доступа: https://www.kommersant.ru/doc/3372761 (дата обращения: 30.06.2023).

9. Нейросеть создала собственный язык, который ученые не могут расшифровать [Электронный ресурс]. – Режим доступа: https://www.ixbt.com/news/2022/06/03/nejroset-sozdala--sobstvennyj-jazyk-kotoryj-uchenye-ne-mogut-rasshifrovat.html (дата обращения: 30.06.2023).

10. Миронова Н.Г. Глава 5. Безопасность использования когнитивных информационных технологий принятия решений / Н.Г. Миронова // Экономика и право. – Чебоксары: Среда, 2021. – С. 112–131 [Электронный ресурс]. – Режим доступа: https://elibrary.ru/item.asp?id=46291391 (дата обращения: 30.11.2023). EDN TAXFFN

Миронова Наталия Геннадьевна – канд. филос. наук, доцент Института истории и государственного управления ФГБОУ ВО «Башкирский государственный университет», Уфа, Россия.
