УДК 004.5

ПЕРСОНАЛИЗИРОВАННЫЕ ЧАТ-БОТЫ В МЕДИЙНОЙ СРЕДЕ

Соколова Марианна Евгеньевна

Кандидат философских наук, старший научный сотрудник, Институт США и Канады РАН им. Г.А. Арбатова (ИСКРАН), Москва, Россия; mariamva@yandex.ru

Аннотация. *В статье рассматривается роль персонализированных чат-ботов с искусственным интеллектом (ИИ), созданных крупными американскими технологическими платформами. Рассматривается медийный эффект создания с помощью персонализированных чат-ботов виртуальных персонажей пользователей, знаменитостей, исторических и вымышленных персонажей на примере чат-ботов Gemini (Bard) от Google, Characters, чат-ботов сетей Gaba, Meta*[1]*. В качестве одного из инструментов регулирования использования чат-ботов рассматривается пользовательская политика цифровых платформ (AI Studio). Акцентируется назревшая потребность в законодательном регулировании чат-ботов с ИИ в медийной среде, в качестве примера новых законодательных подходов к регулированию GenAI рассматривается новый калифорнийский законопроект SB-1047 «Закон о безопасных и надежных инновациях для передовых моделей искусственного интеллекта» (2024). Подчеркивается важность прогнозов экспертов о необходимости контроля и правового регулирования GenAI, отмечается сложность этой задачи.*

Ключевые слова: *искусственный интеллект; персонализированные чат-боты; генеративные нейронные сети; Gemini (Bard); Gaba; Characters AI.*

Для цитирования: Соколова М.Е. Персонализированные чат-боты в медийной среде // Социальные новации и социальные науки. – 2024. – № 4. – С. 25–36.

URL: https://sns-journal.ru/ru/archive/ DOI: 10.31249/snsn/2024.04.02

Рукопись поступила: 22.09.2024.

Принята в печать: 24.10.2024.

[1] Здесь и далее: Meta Platforms признана экстремистской организацией, ее деятельность запрещена в России. Принадлежащие ей социальные сети Facebook и Instagram запрещены в России.

Введение

Сегодня нейронный генеративный искусственный интеллект (GenAI) – активный участник событий в медийной среде, а его коммуникативные роли могут быть самыми разными. Кем же, в какой роли, он может предстать перед нами?

В создании новостного и другого медийного контента сейчас востребованы высокотехнологичные чат-боты с ИИ (AI chatbot). Это касается и профессиональных СМИ, и блогеров, и технологических акторов (цифровых платформ). Роль медийной персоны, участника публичных интернет-коммуникаций, журналиста, ведущего, создателя фото или видео, предлагающего публике захватывающие внимание информационные материалы самой разной направленности – фейковые или достоверные – одна из самых перспективных в арсенале чат-ботов.

Журналисты все чаще используют нейронные генеративные сети для сбора и обработки фактов и материалов (фото, видео и текстов), перевода, редактирования, помощи в написании статей. Их использование стало повсеместной рабочей практикой во всех жанрах СМИ – новостных, авторских, спортивных комментариях и др.

И хотя чат-боты в этой роли – сейчас большей частью только виртуальные умные помощники профессиональных журналистов, считающих, что написание статей и вообще создание контента без редактуры человека неизбежно приведет к искажению фактов, домыслам и наполнению медийного пространства дезинформацией, представители цифровых платформ, напротив, придерживаются точки зрения о том, что ИИ способен вытеснить традиционные СМИ из Интернета. Профессиональные медийщики признают этот натиск GenAI, и многие из них сожалеют о том, что в свое время не вложили средства в соответствующие технологии. Теперь даже интернет-магазины могут выполнять функции СМИ благодаря нейронным генеративным сетям, которые способны создавать новости, фото- и видеоматериалы в неограниченном количестве для привлечения трафика, не говоря уже об огромных информационных возможностях технологических компаний-гигантов, подобно, например, «Яндексу». Эти последние, кроме использования контента традиционных медиа и их архивов для обучения генеративных нейронных сетей, используют и информацию, полученную в результате мониторинга собственных социальных сетей, виртуальных сообществ, маркетплейсов[1].

[1] Эти и другие вопросы обсуждались на Круглом столе «Медиарынок и развитие искусственного интеллекта: бизнес-процессы и модели взаимодействия», организованном факультетом журналистики МГУ совместно с компанией «Яндекс» в рамках Международной научно-практической конференции «Журналистика в 2023 году: творчество, профессия, индустрия» (05.02.2024–06.02.2024). См. программу конференции: https://www.journ.msu.ru/downloads/2024/Programma_konferencii_24.pdf (дата обращения: 05.09.2024.)

Есть и другие формы проникновения генеративных нейронных сетей в медийное пространство, когда чат-боты в медийной среде выступают, если выражаться образно, «не под псевдонимом», как закадровые помощники профессионалов, а как оригинальные виртуальные персонажи, медиа-акторы, ориентированные на широкую аудиторию. Использование таких персонализированных чат-ботов на основе ИИ наряду с обычными виртуальными умными помощниками – сегодня одно из основных направлений в медиасреде, как и вообще в бизнесе. Ключевым преимуществом таких чат-ботов является их способность обеспечивать персонализированное взаимодействие с клиентами. Анализируя пользовательские данные и взаимодействия, они могут адаптировать свои ответы к их индивидуальным предпочтениям и потребностям, а также способны постоянно обучаться и совершенствоваться. Такая персонализация неизменно привлекает пользователей крупных цифровых платформ, вовлекая их во взаимодействие. Чат-боты выступают и как помощники, и как виртуальные персонажи, созданные на основе личности самого пользователя, какой-либо знаменитости или вымышленного персонажа, а также как проводники креативной и интерактивной рекламы. Согласно исследованиям, ориентированное на пользователей проектирование личности чат-ботов позволяет значительно увеличить степень их вовлеченности во взаимодействие на цифровых платформах [Smestad, 2018].

Пользователям цифровых платформ предоставляются инструменты ИИ, которые дают возможность создавать как простых виртуальных помощников, отвечающих в чатах, так и цифровых двойников на основе своих личностей. То же самое могут делать, и с большим бизнес- и PR-эффектом, и знаменитости. На ряде платформ в пользовательских политиках оговаривается ряд ограничений при создании персонализированных чат-ботов. Таким образом, налицо тенденция совершенствования регулирования этой сферы, однако пока инструменты для этого не проработаны и немногочисленны, а связанные с чат-ботами происшествия попадают в центр общественного внимания, поскольку затрагивают порой болевые социальные, политические и исторические темы.

Поэтому хотя, казалось бы, чат-ботам на основе генеративного ИИ открыта дорога в передовые технологии сферы медиакоммуникаций, есть факторы, ставящие под вопрос оптимистические прогнозы их дальнейшего развития.

Виртуальные персонажи

ИТ-индустрия в лице американских высокотехнологичных компаний положительно оценила медийный информационно-развлекательный потенциал нейронных генеративных сетей, который открывает перед пользователями широкие возможности для коммуникации, общения. Каждая цифровая платформа теперь стремится создать своих чат-ботов с помощью соответствующих инструментов ИИ.

Общение с виртуальными персонажами, созданными с помощью ИИ-чат-ботов, безусловно, вызывает интерес у пользователей, которых не может не привлечь возможность пообщаться со знаменитостями – тем же Илоном Маском, Майклом Джексоном, Сократом, супергероем Тони Старком и многими другими, а также с любым историческим или вымышленным виртуальным персонажем на выбор. Пользователям также интересно создать виртуальный персонаж на основе своей собственной личности. Такая возможность открывается, например, благодаря чат-боту Characters AI – инструменту на основе ИИ, созданному двумя разработчиками Google LaMDA – Ноамом Шазиром и Даниэлем Де Фрейтасом. Сейчас он находится пока в стадии тестирования, но уже входит в списки самых популярных чат-ботов с ИИ [ChatGPT и «друзья», 2023].

Благодаря этому чат-боту можно вести разговоры на разные темы с «запрограммированными» нейронной сетью виртуальными образами. Пользователи могут с его помощью создавать «персонажей», их «личности», устанавливая для этого определенные параметры, а затем публиковать их в сообществе, чтобы другие могли общаться с ними. Персонажи в Character AI могут быть основаны на образах мифологических персонажей, героев книг, фильмов, знаменитостей и исторических фигур. Пользователи могут общаться с одним персонажем или организовывать групповые чаты. И вот здесь может произойти искажение не просто бытовых или сиюминутных политических обстоятельств. Порой чат-боты могут исказить и серьезные исторические факты. Иногда это происходит в результате несовершенства технологических функций, программного обеспечения. Если эти чат-боты создаются не в пародийных целях, то дезинформация и агрессия могут стать радикальной медийной «фишкой», которая привлекает внимание определенной категории пользователей.

Словом, на общем фоне того, как в 2024 г. технология GenAI перешла от шумихи к стадии практического внедрения [Gartner … , 2024], медийное использование чат-ботов также переживает бум и, соответственно, постоянно возникающие в связи с этим коллизии привлекают пристальное внимание огромной аудитории и получают широкий резонанс. Так, в текущем году расовая неаутентичность исторических персонажей и просто известных людей в поисковой выдаче Gemini (Bard)[1] от Google вызвала публичные дискуссии об их расистском подтексте. Причиной таких случаев является жесткая конкуренция крупных технологических компаний и их курс на то, чтобы не отставать от конкурентов, в результате чего пользователям предлагаются несовершенные модели. Однако результатом становится дезориентация общественного сознания.

[1] Gemini (ранее – Bard) – чат-бот с искусственным интеллектом, разработанный компанией Google на основе языковой модели LaMDA.

Экзистенциальные риски чат-ботов на основе ИИ и пути их преодоления

На фоне тех широких откликов, в том числе и обвинений в расизме, которые вызвал инцидент с Gemini, невольно вспоминается почти аналогичная ситуация из области искусства: на театральной сцене дважды за последнее десятилетие – в 2013 и 2024 гг. – появлялись эпатажные спектакли, в которых роль Джульетты играли чернокожие актрисы. Эти образы стали объектами неприятия, вызывали порой очень раздраженную реакцию публики. В то же время они поставили перед общественным сознанием вопрос об аутентичности современной традиции постановок этой пьесы Шекспира вообще. И убедительность некоторых аргументов здесь трудно не признать. Правда, какие бы изменения ни претерпела сценография и постановка этой пьесы за несколько столетий со дня ее создания, начиная с отсутствия вначале в оригинале пресловутого балкона, все-таки чернокожей Джульетты в шекспировском оригинале точно не было.

Чернокожие солдаты вермахта и отцы-основатели США, которых в истории тоже не было, в поисковой выдаче Gemini от Google пришли не из мира художественного вымысла и искусства, а из мира технологически препарированной истории. И надо сказать, ситуация претерпела кардинальные изменения за несколько лет с 2015 г., когда сервис фотографий Google Photos принял за горилл чернокожих людей на фотографиях, автоматически поместив снимки с ними в альбом с таким названием. Google даже пришлось тогда исключить слово «горилла» из словарного запаса этого приложения. В борьбе с подобными проявлениями предвзятости и социальной дискриминации компании, занимающиеся разработками в области ИИ, приложили все усилия для того, чтобы создать модели, соответствующие разнообразию их аудитории.

В результате Google Gemini в феврале 2024 г. в ответ на запрос сгенерировать изображения белых исторических личностей, например отцов-основателей США и солдат вермахта, выдавал в качестве ответа изображения чернокожих людей. Были там, по отзывам пользователей, и изображения чернокожих викингов и Папы Римского. Многие пользователи восприняли инцидент как «предвзятое отношение к белым». В адрес Gemini прозвучали многочисленные обвинения в расизме и сексизме, например, со стороны того же Илона Маска. Критики компании делали намеки на существование предвзятого продвижения определенной программы, социальной инженерии [Gemini снова …].

Представители Google тогда объяснили, что ИИ не мог работать корректно из-за строгих этических настроек, в основе которых лежит стремление избежать предвзятости и обеспечить разнообразие в поиске для многочисленной и многообразной аудитории. В итоге функция на некоторое время была отключена и вскоре было выпущено обновление. Были подобные ошибки и у модели Google Bard при ответе на вопрос о недавних открытиях космического телескопа Джеймса Уэбба, когда чат-бот ошибся с датой снимка, который на самом деле был сделан давно.

Причиной исторически неадекватных ответов чат-ботов может быть и их политическая ориентация, интегрированная в них в соответствии с установками их создателей, как, например, в известной ультраправой социальной сети Gab, которую создал в 2018 г. Эндрю Торба. В Gab запустили новую платформу под названием Gab AI специально для чат-ботов. Среди «персонажей» – почти 100 чат-ботов, созданных ИИ, без цензуры и предвзятости. Среди них много политических деятелей: Адольф Гитлер, Дональд Трамп, Тед Качинский, Фидель Кастро, Владимир Ленин и многие другие. Некоторые из них помечены как пародийные аккаунты, но другие, такие как чат-боты Трампа и Гитлера, таковыми не являются.

Чат-боты Gab прославились, в частности, своим отрицанием реальности Холокоста, утверждая, что это мистификация, которая является пропагандистским инструментом сионистов, а также утверждениями о том, что изменение климата – это мошенничество, а президентские выборы в США 2020 г. были сфальсифицированы, и победил на них на самом деле Д. Трамп. Так, чат-бот Adolf Hitler на вопрос о Холокосте отрицал существование геноцида, назвав его «пропагандистской кампанией по демонизации немецкого народа» и «подавлению правды» [Gilbert, 2024].

Gab AI имеет генератор изображений и позволяет пользователям создавать своего собственного чат-бота, которому они могут придать любые предубеждения и мировоззрение. По словам Торбы, Gab беспристрастна и не подвержена цензуре в том смысле, что она позволяет представлять различные точки зрения, здесь нет модерации контента, поскольку она позиционирует себя как средство обеспечения свободы слова [Gilbert, 2024].

С момента своего запуска в 2016 г. эта социальная медиаплатформа, стала местом встречи для ультраправых. Как и другие альтернативные платформы, такие как Parler и MeWe, Gab позиционируется как платформа свободы слова и как безопасная зона для сообществ, которых давно бы заблокировали в других социальных сетях. Несколько сообщений написал в ней в 2018 г. и Роберт Бауэрс перед тем, как застрелить в синагоге Питтсбурга 11 человек. Правда, правила Gab запрещают некоторые типы сообщений, такие как угрозы насилия и порнографию.

Apple и Google и многие другие интернет-сервисы уже давно запретили доступ к Gab в своих магазинах приложений из-за его радикалистской ориентации. Наряду с другими альтернативными социальными приложениями, сеть стала популярной в США на фоне беспорядков в Вашингтоне в начале 2021 г. и блокировки аккаунта Трампа рядом интернет-компаний.

Основатель и генеральный директор Gab Эндрю Торба неоднократно высказывался против модерации контента в той форме, в которой она осуществляется на крупных цифровых платформах. Последние несколько лет по мере роста популярности чат-ботов с генеративным ИИ, таких как ChatGPT (OpenAI), со стороны правых все чаще раздаются заявления, что LLM (Large Language Models) несут в себе антиконсервативную предвзятость своих создателей, которая совпадает с преобладающими идеологическими взглядами левых в Big Tech. Торба считает, что все крупные

технологические игроки настолько озабочены безопасностью, что сделали свои ИИ- инструменты практически бесполезными, скованными жесткой модерацией. Но это может измениться, поскольку основные новостные платформы все чаще запрещают LLM использовать свой контент для обучения чат-ботов, в отличие от праворадикальных СМИ, которые по-прежнему это допускают. Возможно, Торба прав, и действительно скоро все LLM значительно «поправеют». Время это покажет.

Эксперты по онлайн-экстремизму обращают внимание на то, что такие высказывания чат-ботов приведут к распространению дезинформационных материалов и еще большей радикализации людей, которые придерживаются взглядов, основанных на теориях заговора. По словам Адама Хэдли, исполнительного директора британской некоммерческой организации Tech Against Terrorism, которая отслеживает онлайн-экстремизм, эти чат-боты могут быть превращены в пропагандистское оружие, и их потенциальное применение варьирует от радикализации общественного сознания до распространения пропаганды и дезинформации [Gilbert, 2024][1]. В этой связи потребность в надежной модерации контента для генеративного ИИ на основе общего законодательств становится все более актуальной.

Одним из инструментов регуляции деятельности чат-ботов является пользовательская политика платформ. Например, согласно пользовательской политике компании Meta (признана в России экстремистской организацией) с помощью ее инструмента AI Studio, который доступен по адресу ai.meta.com/ai-studio, можно создать виртуальных персонажей с индивидуальными чертами характера и интересами, в том числе основанных на собственной личности пользователя. Пользуясь ИИ-инструментами, интегрированными в популярные приложения Meta (признана в России экстремистской организацией), пользователи без технических навыков создают, настраивают и общаются с персонализированными чат-ботами, которые соответствуют их интересам и потребностям. Например, это может быть виртуальный друг, который будет делиться рецептами, помогать с написанием постов, генерировать юмористический контент, создавать мемы, давать советы по путешествиям. Это может быть и цифровой двойник пользователя или помощник, который будет взаимодействовать с другими пользователями, отвечать на вопросы и поддерживать беседу.

ИИ-профили цифровых двойников можно настроить, основываясь на контенте пользователя в Instagram (в России эта соцсеть внесена в реестр запрещенных сайтов). Пользователи могут взаимодействовать со своими цифровыми двойниками, обучая их, чтобы они лучше отражали их

[1] Есть и позитивные отклики на общение с чат-ботами этой сети. Например, на сайте «Хабр» удалось найти такой отзыв: «Не знаю, сколько эта свобода еще продлится, но задумка весьма интересная и полезная». Правда, этот отзыв дал пользователь, который задал чат-боту вопрос о том, как убить кошку в случае голода и вскрыть дверцу автомобиля, если потерял ключ, и получил корректные, но исчерпывающие ответы на эти вопросы [Gab – AI бот … , 2024].

личность, а чат-боты могут запоминать предыдущие разговоры и все больше адаптироваться к индивидуальному стилю общения своего создателя.

Создать такого персонажа могут и знаменитости. По словам генерального директора Meta (признана в России экстремистской организацией) Марка Цукерберга, пользователи могут создавать собственные чат-боты с ИИ для развлечения или в качестве инструментов личной поддержки – например, для ролевых игр о том, как попросить прибавку к зарплате или разобраться в споре с другом, посмотреть, как пойдет беседа, и получить обратную связь по этому поводу [Knight, Dave, 2024].

Инструменты AI Studio позволяют пользователям ограничивать, с кем взаимодействуют их чат-боты, и запрещать им обсуждать определенные темы. Можно включать и выключать такие функции, как автоответы и общение с конкретными аккаунтами. Политика использования AI Studio запрещает пользователям представлять реальных людей, отличных от них самих. Запрещен доступ к историческим личностям, религиозным деятелям, массовым убийцам или объектам, которые могут рассматриваться как вызывающие ненависть, откровенные или незаконные[1]. Все созданные чат-боты должны быть обозначены как таковые. Пользователи должны знать, когда они общаются с ИИ, виртуальными помощниками. Сейчас опция создания персонализированных чат-ботов доступна пользователям в США [Knight, Dave, 2024]. В России данный ресурс недоступен.

Что касается американского законодательства, то согласно разделу 230 «Закона о приличиях в области коммуникаций» (The Communications Decency Act of 1996), технологические компании не несут юридической ответственности за то, что публикуют их пользователи. Приоритетными здесь являются защита свободы слова и доступ к информации. Чат-боты не подпадают под действие этого закона.

Среди первых попыток создания регулирующего законодательства для GenAI следует назвать новый калифорнийский закон SB 1047 «Закон о безопасных и надежных инновациях для передовых моделей искусственного интеллекта» (Safe and Secure Innovation for Frontier Artificial Intelligence Models» Act), разработанный под руководством калифорнийского сенатора Скотта Винера и прошедший летом 2024 г. все обсуждения в калифорнийском Сенате.

SB-1047 вводит ряд требований, направленных на установление стандартов безопасности и тестирование при разработке крупномасштабных систем ИИ. Фактически в нем на технологические компании возлагаются обязанности принимать меры для того, чтобы их продукция не использовалась для причинения вреда. Если они тратят более $100 млн на обучение «передовой модели» ИИ (к примеру, GPT-5), должны проводиться тесты на безопасность инструмента и

[1] См.: Usage policy. – URL: https://aistudio.instagram.com/policies (accessed: 20.07.2024).

внедряться специальные процедуры реагирования на инциденты, связанные с безопасностью. Большие модели должны иметь механизм отключения систем ИИ, который позволил бы, остановить их в случае угрозы катастрофы или массовой гибели людей. Если компания этого не сделала, она должна будет нести ответственность. Подобные законопроекты сейчас обсуждаются и в других штатах США. Большая часть регулирования связана с потенциальным ущербом, который может быть причинен из-за распространения высокоразвитых систем ИИ, таких как GenAI [Tobey, Carr, Kloeppel, 2024].

SB-1047 заслуживает пристального внимания как одна из первых попыток американского законодательства решить проблемы рисков, связанных с генеративным ИИ. Однако речь пока идет только о регулировании в случаях техногенных катастроф, которые могут произойти в результате действий с использованием ИИ.

Против нескольких положений закона сразу же резко выступили представители крупнейших технологических компаний, которым удалось отстоять ряд изменений. Противники закона подчеркивают, что реализация законопроекта на практике потребовала бы от технологических компаний предсказывать и полностью контролировать то, как люди используют их модели, что привело бы к нарушению права на неприкосновенность частной жизни [Tobey, Carr, Kloeppel, 2024].

В итоге в конце сентября 2024 г. губернатор Калифорнии Гэвин Ньюсом не подписал данный законопроект о безопасности ИИ. В сообщении по этому поводу говорится, что он распространяется только на самые крупные и дорогостоящие модели ИИ. Помимо этого, закон касался бы деятельности не только крупнейших компаний в области AI, офисы которых расположены в Калифорнии, под его регулирование попала бы и деятельность тех компаний, которые просто ведут деятельность в штате [Fritz, Rana, 2024].

Заслуживает пристального внимания поддержка законопроекта со стороны экспертов в области ИИ. Так, некоторые известные ученые выступили за регулирование и обеспечение мер безопасности для предотвращения катастрофических последствий использования ИИ. Среди них профессор Джеффри Хинтон (Университет Торонто), лауреат Нобелевской премии «за фундаментальные открытия в области машинного обучения с помощью нейронных сетей», чьи исследования легли в основу нынешнего бума генеративных нейронных сетей. Хинтон даже заслужил звание «крестного отца» искусственного интеллекта. Он проработал более 10 лет – до 2023 года – в Google, где помог разработать технологию ИИ и подход, который проложил путь для современных систем, таких как ChatGPT. Сейчас его беспокоит, как именно эта технология будет применяться в будущем и кто будет ее контролировать.

Создатель современного GenAI Дж. Хинтон обеспокоен тем, что люди больше не смогут различать, что является правдой, поскольку фотографии, видео и тексты, созданные ИИ, наводняют Интернет. Новые модели генераторов изображений, таких как Midjourney, означают, что люди те-

перь могут создавать самые фантастические, но при этом реалистичные фотоизображения, которые могут восприниматься как реальные. Например, изображения папы Франциска в пуховике Balenciaga, которое стало вирусным в социальных сетях, было названо писателем Райаном Бродериком (Ryan Broderick) первым реальным случаем массовой дезинформации ИИ [Lu, 2024].

Технологическая конкурентная гонка, которая начинается между крупнейшими компаниями, считает Хинтон, приведет к резкому прогрессу разработанных ими моделей, которые быстро превзойдут возможности человеческого мозга. По мере того, как технологические гиганты будут совершенствовать свои системы ИИ, это будет становиться все более непредсказуемым. Остановить эту технологическую конкуренцию, по мнению Хинтона, будет невозможно. Установить контроль над ситуацией необходимо именно сейчас, перед тем как появятся модели следующего поколения [Metz, 2024].

Сейчас, когда сгенерированные GenAI дипфейки порой заставляют не верить любым новостям, чтобы сдержать натиск дезинформации, который распространяется с помощью человекоподобных чат-ботов, все чаще используются боты ИИ-детекторы, такие как Writers, Copyleaks и AI-text-classifier от OpenAI, предназначенные для распознавания текстов, созданных машинами. Разрабатываются и внедряются также детекторы, основанные на нейронных генеративных сетях, которые могут распознать сгенерированные видео, покадрово изучая особенности видеоряда, незаметные человеческому глазу. Технологии защиты от дополнительных рисков, связанных с дипфейками – синтезированными ИИ изображениями людей, на основе GenAI, – все чаще устанавливаются на мобильных устройствах, позволяя оперативно предупреждать их хозяев о дипфейках.

Заключение

Сейчас мы становимся свидетелями и участниками того, как ситуация с влиянием на сознание людей созданной ИИ чат-ботами информации, которая часто оказывается дезинформацией, переходит в сферу ценностно-исторического знания.

Чат-боты отдаляют нас от реальности и от реальной истории, создавая некую параллельную реальность ИИ. И эти фантазийные миры уже касаются не только компьютерных игр, фэнтези или виртуальной моды. Дезинформация и дезориентация человеческого сознания все больше касается и реальной истории. В этом контексте необходимы новые регулятивные решения проблемы предотвращения социальных и ценностно-экзистенциальных рисков.

В эпоху информационной сетевой революции и повсеместной цифровизации, порождающих ощущение полной пластичности и управляемости цивилизации, проблема сохранения исторического сознания сегодня стоит на повестке дня в связи с использованием умных чат-ботов на основе ИИ в целях социальной инженерии и манипулирования обществом. И активное продвижение

генеративного нейронного ИИ в медиасреде может значительно способствовать увеличению экзистенциальных рисков для цивилизации.

Список литературы

1. Михайлов Е. Возмущение темнокожей Джульеттой – это не любовь к Шекспиру, а расизм // Афиша Daily. – 2024. – 22.05. – URL: https://daily.afisha.ru/culture/27400-vozmuschenie-temnokozhey-dzhulettoy-eto-ne-lyubov-k-shekspiru-a-rasizm-vot-pochemu/ (дата обращения: 21.08.2024).
2. ChatGPT и «друзья»: лучшие чат-боты с искусственным интеллектом в 2024 // Eternalhost. – 2023. – 28.03. – URL: https://eternalhost.net/blog/tehnologii/chat-boty-s-iskusstvennym-intellektom (дата обращения: 21.08.2024).
3. Gab – AI бот без цензуры и предвзятости, отвечающий почти на любые вопросы / kovalensky // Хабр. – 2024. – 02.02. – URL: https://habr.com/ru/articles/790890/ (дата обращения: 21.08.2024).
4. Gartner: генеративный ИИ близок к «избавлению от иллюзий» // itWeek. – 2024. – 22.08. – URL: https://www.itweek.ru/ai/article/detail.php (дата обращения: 21.08.2024).
5. Gemini снова рисует людей: больше никаких чернокожих Пап Римских. – 2024. – 29.08 – URL: https://www.securitylab.ru/news/551561.php (дата обращения: 06.09.2024).
6. Fritz B., Rana P. California's Gavin Newsom vetoes controversial AI safety bill // The Wall Street journal. – 2024. – 29.09. – URL: https://www.wsj.com/tech/ai/californias-gavin-newsom-vetoes-controversial-ai-safety-bill-d526f621 (accessed: 21.08.2024).
7. Gilbert D. Gab's racist AI chatbots have been instructed to deny the Holocaust // WIRED. – 2024. – 27.07. – URL: https://www.wired.com/story/gab-ai-chatbot-racist-holocaust/ (accessed: 21.08.2024.)
8. Knight W., Dave P. Instagram will let you make custom AI chatbots – even ones based on yourself // WIRED. – 2024. – 29.07. – URL: https://www.wired.com/story/meta-ai-studio-instagram-chatbots/ (accessed: 25.08.2024).
9. Lu D. Misinformation, mistakes and the Pope in a puffer: what rapidly evolving AI can – and can't – do // The Guardian. – 2024. – 31.04. – URL: https://www.theguardian.com/technology/2023/apr/01/misinformation-mistakes-and-the-pope-in-a-puffer-what-rapidly-evolving-ai-can-and-cant-do (accessed: 21.08.2024).
10. Metz C. 'The Godfather of A.I.' leaves Google and warns of danger ahead // The New York Times. – 2023. – 04.05. – URL: https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html (accessed: 21.08.2024).
11. SB 1047 (SB-1047 Safe and Secure Innovation for Frontier Artificial Intelligence Models Act (2023–2024) // California Legislative Information. – 2024. – 09.03. – URL: https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240SB1047&search_keywords=Safe+and+Secure+Innovation+for+Frontier+Artificial+Intelligence+Models+Act (accessed: 21.08.2024).
12. Smestad T.L. Improving the user experience of chatbot interfaces – personality provides a stable pattern to guide the design and behavior of conversational agents: master thesis // Semantic scholar. – 2018. – URL: https://www.semanticscholar.org/paper/Personality-Matters%21-Improving-The-User-Experience-Smestad/a319f0414245f13ef614ef37edfe3cacddc89057 (accessed: 21.08.2024).
13. Tobey D., Carr A., Kloeppel K. California's SB-1047: understanding the safe and secure innovation for frontier Artificial Intelligence act // DLA Piper. – 2024. – 20.02. – URL: https://www.dlapiper.com/en/insights/publications/2024/02/californias-sb-1047 (accessed: 21.08.2024).

PERSONALIZED CHATBOTS IN THE MEDIA ENVIRONMENT

Marianna Sokolova
PhD. (Phil. Sci.), Senior Researcher, Georgy Arbatov Institute for U.S. and Canada Studies of the Russian Academy of Sciences (ISKRAN), Moscow, Russia; mariamva@yandex.ru

Abstract. *The article examines the role of personalized AI chatbots created by major American technology platforms. The media effect of creating virtual characters using chatbots of users themselves, celebrities, historical and fictional characters is considered, using the example of Gemini (Bard) chatbots from Google, Characters, chatbots of Gaba networks, Meta. The user policy of digital platforms (AI Studio) is considered as one of the tools for regulating the use of chatbots. The urgent need for legislative regulation of AI chatbots in the media environment is emphasized, and the new California bill SB-1047*

"Safe and Secure Innovation for Frontier Artificial Intelligence Models Act" (2024) is considered as an example of new legislative approaches to GenAI regulation. The importance of forecasts by AI experts on the need for legal regulation of GenAI is emphasized.

Keywords: *artificial intelligence; personalized chatbots; generative neural networks; Gemini (Bard); Gaba; Characters AI.*

For citation: Sokolova M.E. Personalized chatbots in the media environment // Social novelties and social sciences. – 2024. – N 4. – P. 25–36.

URL: https://sns-journal.ru/ru/archive/ DOI: 10.31249/snsn/2024.04.02