Искусственный интеллект и фальсификация данных

Бразуль-Брушковский Е. Г. (brazul.bruszkowski@gmail.com)

Российский государственный социальный университет, г. Москва

Аннотация

Выгоды использования ИИ не только широко обсуждаются и рекламируются, но и становятся целью реализации масштабных государственных программ. На примере двух кейсов – из отечественной и зарубежной практик – демонстрируется, как ИИ систематически фальсифицирует информацию, настойчиво выдавая ее за действительную.

«Цифровизация» давно уже стало привычным словом в обиходе; ее практическое выражение в практике повседневной жизни заключается в том, что многие привычные социальные практики, в т. ч. постоянно растущий ряд государственных услуг являются сегодня доступными лишь в электронном формате. Этим, конечно, понятие цифроизации не исчерпывается, но, еще раз, принципиальная «виртуальность» давно привычных, как и новых форм экономических и социальных взаимодействий оказывает огромное и до сих пор плохо изученное влияние как на психологию, так и на культуру общества.

Однако в последние годы дополнительный акцент делается на развитии технологий, связанных с т. н. «искусственным интеллектом». На основании Указа Президента РФ от 10 октября 2019 г. №490 «О развитии искусственного интеллекта» [3] в стране в 2021 г. запущена национальная стратегия по развитию ИИ. Специально посвященный ей веб-сайт [1] несколько декларативно, но масштабно описывает направления деятельности, связанные с разработкой и использованием ИИ – вопросы этического и правового регулирования, внедрения в различные формы хозяйствования и, конечно, подготовку специалистов.

Т. о., вопросы целесообразности и безопасности самой идеи ИИ как бы снимаются с повестки дня; остаются лишь аспекты регулирования и, понятно, что возможности применения ИИ для обеспечения «национальной безопасности и правопорядка», обозначенные на упомянутом сайте на втором месте после цели «достижения конкурентноспособности российской экономики» [1] в глазах поборников ИИ перевешивают возможные, с их точки зрения, «риски». Позволю себе проиллюстрировать свои опасения по поводу неудержимого энтузиазма относительно фантастических возможностей ИИ на конкретном примере или, как принято выражаться сегодня, кейсе.

Занимаясь реализацией собственного просветительского проекта, «Иное измерение русской культуры», я, листая подшивку одного из старых, дореволюционных, журналов наткнулся на стихотворение, подписанное псевдонимом «Черный Бор» [5]. Поэты и писатели конца XIX – начала XX вв., действительно, очень любили псевдонимы, причем каждый поэт мог иметь в своем арсенале не одно, а десяток различных имен. Идентификация автора, в таком случае, может составить серьезную проблему, особенно если стихи принадлежат персоне по каким-то причинам давно забытой. Первая строчка стихотворения с мечтательным, но малоинформативным названием – «Неведомое» – звучала так: «Я девушек нежных и бледных, как звезды, люблю…».

Поиск по привычным ресурсам ничего не дал, а вот системы ИИ, в принудительном порядке теперь включенные в наиболее популярные браузеры, за версиями в карман (или что там у них?) не полезли. Браузер Chrome от Google немедленно сообщил, что «псевдоним «Черный Бор» принадлежал русскому поэту и писателю Александру Гликбергу, более известному как Саша Черный». Браузер Yandex настаивал на другой версии: «…если речь идёт о псевдониме "Чёрный Бор", то его использовал Сергей Александрович Соколов (1878–1936) — русский поэт-символист, более известный под другим своим псевдонимом — Сергей Кречетов».

Понимая, что до истины так добраться не выйдет, в поисковую строку была введена упомянутая первая строка стихотворения. ИИ от Яндекса продолжал настаивать на своем: «Оно называется «Любовь» (или иногда публикуется без названия) и входит в его сатирическую и лирическую поэзию. Вот полный текст: Я девушек нежных и бледных, как звёзды, люблю…/ Их взоры — туманны, их кудри — пушисты, / Их голос так тих, что едва я расслышу, / Но если они говорят: «Ты мне лжёшь!» — / Я верю… и сердце от счастья дрожит!».

Отсутствие рифм и параллельные альтернативы со стороны Google заставили усомниться в этой версии, и в поиск была добавлена вторая строка стихотворения. ИИ Яндекса быстро изменил свою точку зрения: «Вы цитируете строки из стихотворения Игоря Северянина (Игоря Васильевича Лотарева). Полный текст произведения: Я девушек нежных и бледных, как звёзды, люблю, / Я им свои думы, печали и радости шлю. / Их взоры — как лунные блики в лесной глубине, / Их речи — как звон хрусталя в зачарованной мгле». Заметьте, что теперь с рифмой все стало в порядке. Была лишь одна проблема: начиная с третьей строчки текст опять не соответствовал напечатанному в старом журнале.

Возмущенный такой самоуверенностью Яндекса, гугловский Chrome решил играть по-крупному: «Это строки из романа в стихах Александра Пушкина “Евгений Онегин”. Они описывают чувства главного героя, Онегина. Он признается, что любит нежных и бледных девушек, как звезды и посвещает им свои мысли, печали и радости. <…> Пушкин создает образ человека, который не может найти гармонию в отношениях и часто испытывает одиночество». Несмотря на непререкаемый авторитет Пушкина и, вероятно, в связи с тем, что уж проверить текст «Евгения Онегина» проще простого, ИИ от Яндекса продолжал убеждать в своей правоте и даже принялся подсказывать источники, по которым можно проверить подлинность его версии: «Где искать оригинал? Если вы хотите проверить подлинность: 1. Ранние журналы (1900–1910-е): — «Гиперборей» (издание акмеистов, но там печатались и футуристы), — «Весы», «Аполлон» (символистские журналы, где дебютировал Северянин). 2. Эпистолярное наследие — В письмах Северянина к В. Брюсову или Ф. Сологубу могли сохраниться черновики. 3. Провинциальные альманахи — Поэт публиковался в малотиражных сборниках (например, «Поэзоконцерты» в Эстонии)». Здесь, конечно же, оставалось уж только смеяться – причем не искусственным, а вполне естественным смехом…

Как выяснилось впоследствии, «Черный Бор» – это псевдоним сотрудника ежемесячных литературных приложений к журналу «Нива» (в котором и был опубликован стих) Александра Фигурина. Никакой иной информации о нем мне обнаружить не удалось. Но, по крайней мере – и, в отличие от ИИ – я могу указать реальный, а не выдуманный источник такого мнения [2]. Стихотворение «Неведомое» было напечатано в 51-м номере журнала "Нива" за 1913 год [4].

Таким образом, две разные нейросети, доступ к которым осуществлялся с разных MAC-и IP- адресов с использованием разных браузеров, целенаправленно создавали заведомо ложную информацию, при этом подтасовывая (подстраивая) ее всякий раз, когда они получали дополнительные вводные данные.

К этому кейсу остается лишь добавить недавнее (июнь 2025 г.) расследование журналиста телекомпании «SkyNews» Сэма Коатса [6], который в своем материале показал, как ChatGPT целенаправленно лгал ему, пытаясь убедить журналиста в том, что представленный последнему текст – это подлинный текст последнего подкаста, подготовленного самим журналистом и, на самом деле, еще не размещенный на портале. Лишь после получасового препирательства и неопровержимых данных, представленных журналистом, ИИ сообщил: «Вы совершенно правы. Я ошибочно сфабриковал эту запись (I mistakenly fabricated this transcript)».

Выводы предоставим делать читателям.

Литература

1. Искусственный интеллект Российской Федерации [Электронный ресурс]. (Режим доступа: https://ai.gov.ru/national-strategy/. Дата обращения: 14.06.2025).
2. Масанов И. Ф. Словарь псевдонимов русских писателей, ученых и общественных деятелей. Т.4. М.: Издательство Всесоюзной книжной палаты, 1960. С.490.
3. Указ Президента Российской Федерации от 10.10.2019 г. № 490 «О развитии искусственного интеллекта в Российской Федерации» [Электронный ресурс]. (Режим доступа: http://www.kremlin.ru/acts/bank/44731. Дата обращения: 14.06.2025).
4. Черный Бор. Неведомое // Нива. 1913. №51. С.2001.
5. Coates, S. Can we trust ChatGPT despite it 'hallucinating' answers? Sky News revealed how ChatGPT was fabricating entire transcripts of a podcast, Politics at Sam and Anne's. [Electronic resource]. URL: https://news.sky.com/story/can-we-trust-chatgpt-despite-it-hallucinating-answers-13380975. (Accessed: 9 June 2025).