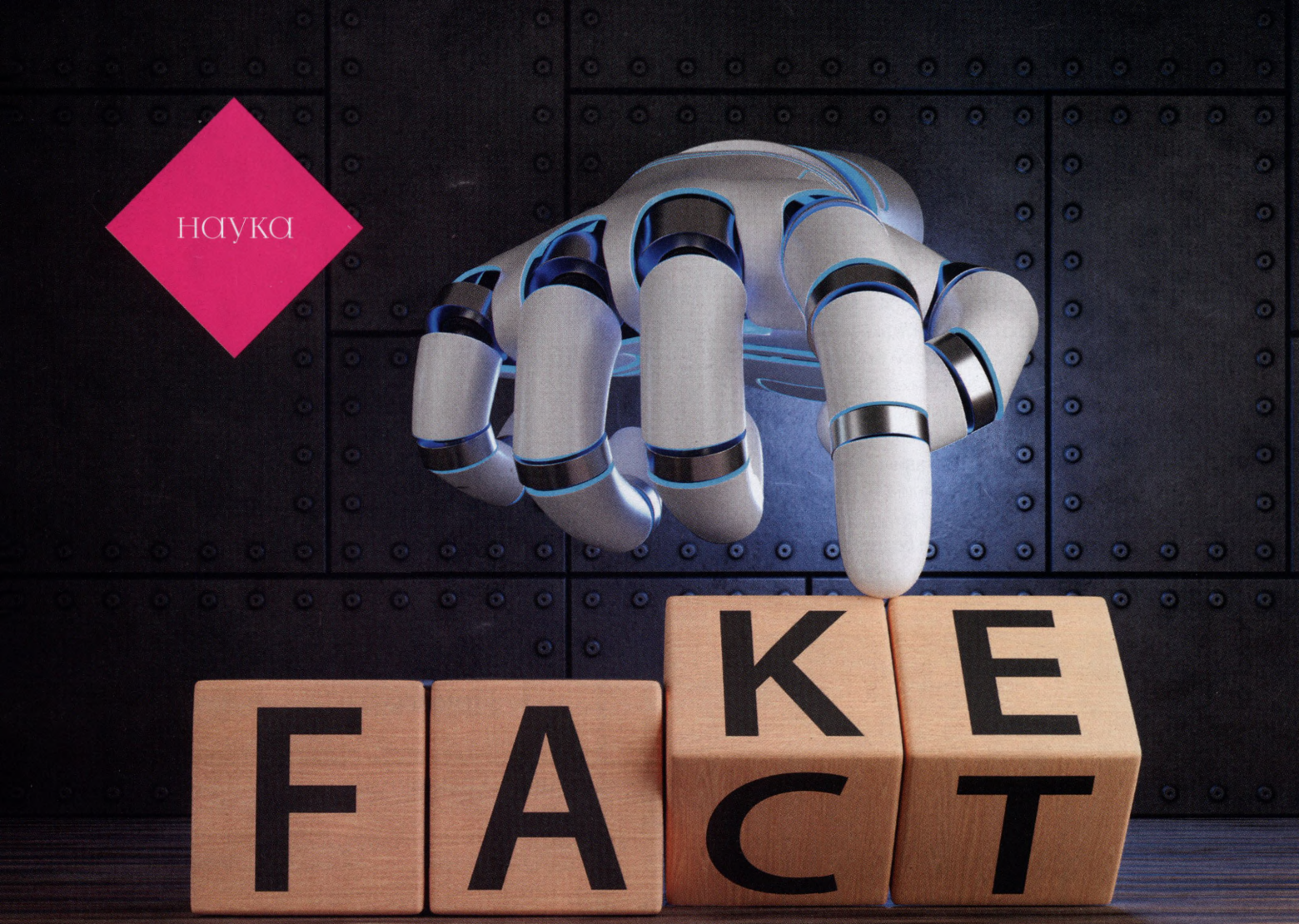
наука



генераторы заблуждений

Отличить правду от вымысла всегда было трудно, а в эпоху искусственного интеллекта эта задача многократно усложняется. Рассказываем, почему нейросети иногда лгут, чем мы из-за этого рискуем и как жить в эпоху «ИИстины».

ТЕКСТ: *Сергей Марков*

Не верь глазам своим

Картина «Theatre D'opéra Spatial» изображает фантастический оперный театр в космическом пространстве, люди, одетые в стиле барокко, спускаются на сцену сквозь круглый портал. В августе 2022 года эта работа заняла первое место в категории «Цифровое искусство» на ярмарке штата Колорадо. Пикантная деталь: жюри не знало, что изображение создано системой искусственного интеллекта (ИИ) Midjourney. За этой историей последовала череда других. Папа римский в модном пуховике, Илон Маск, целующий робожену, пожар в Пентагоне, заваленный кучами мусора Париж – отличить реальный контент от сгенерированного стало очень сложно. К изображениям прибавилось видео. В 2025 году публика приняла за чистую монету видео бушующих в Лос-Анджелесе пожаров. Не помог даже поставленный в оригинальном посте тэг *#aiart*.

В интернете все больше сгенерированных машинами изображений, видео и текстов. Они влияют на наши представления о реальности, а значит, и на решения, которые мы принимаем.

Знает ли ИИ, чего он не знает?

Стать жертвой преднамеренной фальсификации – лишь одна из опасностей, подстерегающих нас в эпоху ИИ. Другая – «добросовестные ошибки» нейросетей. В последнее время стало модным консультироваться по разным вопросам с генеративными языковыми моделями, то есть системами ИИ, предназначенными для написания текстов.

В 2023 году газета «Нью-Йорк Таймс» сообщила о курьезном случае. В США два адвоката предоставили судье документ, полный ссылок на несуществующие судебные решения. Адвокаты воспользовались сервисом ChatGPT, в основе которого – генеративные языковые модели семейства GPT. Судья оштрафовал незадачливых юристов на 5000 долларов и предписал отправить копию этого «литературного произведения» каждому из реальных судей, чьи имена в нем фигурировали.

Нейросети ошибаются чаще, чем может показаться. В качестве эксперимента зададим системе GPT-4.1 nano вопрос: «В каком году скончался писатель Борис Андреевич Лавренев (настоящая фамилия Сергеев)?». Ответ ИИ – в 1977 г., правильный ответ – в 1959 г. Получив запрос, система уверенно перечисляет произведения

Изображение папы римского в белом пуховике, сгенерированное искусственным интеллектом, моментально стало вирусным

Лавренева, якобы написанные в 1960-х гг. Перед нами типичная галлюцинация нейросети: выдуманное ею ложное утверждение. Дата смерти писателя – безобидный пример, но представьте, что с нейросетью консультируется врач или инженер.

Откуда взялась эта галлюцинация? Модель обучалась на конкретном наборе данных, в котором могло и не быть даты смерти Лавренева. Или дата была, но встречалась слишком редко, чтобы система ее запомнила. Нейросети, как и человеку, трудно запомнить информацию, которая редко «попадается на глаза».

Человек мог бы в этой ситуации сказать «я не знаю». Почему так не говорит нейросеть? Потому что ее этому не учили. В обучающем наборе данных было очень мало (а то и вовсе не было) примеров, когда на вопрос пользователя следовал ответ «я не знаю». Генеративные модели напоминают студента, не готового к вопро-

Нейросети для генерации текста лгут, так как сами не отличают правду от вымысла

ФОТО: ISTOCKPHOTO

Картина «Theatre D'opéra Spatial», сгенерированная искусственным интеллектом

Илон Маск с девушкой, похожей на робота. Изображение сгенерировано искусственным интеллектом

су на экзамене. Многие студенты знают, что в такой ситуации болтать хоть что-нибудь – лучше, чем молчать. Если же поощрять ИИ отвечать «я не знаю», нейросеть будет часто давать этот бесполезный для пользователя ответ, даже когда могла бы ему помочь. В идеальном случае модель должна говорить «я не знаю», только когда действительно не знает. Но генеративные сети не учат понимать, что они знают, а что нет. Их задача – генерировать правдоподобный текст, а не отличать правду от вымысла.

Эту проблему можно смягчить. Например, научить нейросеть сообщать пользователю, насколько она уверена в своем ответе. Еще один подход – снабдить систему базой знаний, куда она сможет обращаться за точной информацией. Допустим, хоть пока до этого и очень далеко, что благодаря усилиям разработчиков языковые модели перестанут фантазировать и научатся выдавать лишь ту информацию, что действительно содержалась в обучающих данных. Станет ли тогда нейросеть говорить правду, только правду и ничего кроме правды? Нет.

Дело в том, что для обучения больших языковых моделей вроде GPT-4 или DeepSeek нужно очень много текста. Для этого не хватит научных книг и прочих авторитетных источников. Приходится выкачивать из интернета терабайты текстов самого разного качества. Обученная на таких данных система искусственного интеллекта впитывает содержащиеся в них ошибки, противоречия и популярные заблуждения.

В качестве примера спросим все ту же нейросеть GPT-4.1nano: «Зачем схоласты считали ангелов на кончике иглы?» Из ответа можно узнать, что «одним из известных обсуждений было предположение о том, сколько ангелов может поместиться на кончике иглы», которым схоласты задавались ради «исследования природы бесконечности» и «дебатов о материи и форме». Между тем в схоластике просто не было такого обсуждения. Начало этому заблуждению положил теолог XVII века Уильям Чиллингворт, иронизировавший над средневековым богословием. Пожалуй,

ФОТО: JASON M. ALLEN / MIDJOURNEY

ближайший хорошо известный аналог в схоластике – куда более разумный вопрос Фомы Аквинского, могут ли несколько ангелов находиться в одном и том же месте одновременно.

Человеческие заблуждения – не единственный источник ошибок в обучающих данных. В интернете растет доля контента, сгенерированного ИИ. Есть опасность, что в будущем нейросети будут обучаться на... галлюцинациях других нейросетей.

Энтузиасты пытаются бороться с ИИ-фейками. Например, в 2024 году был запущен проект WikiProject AI Cleanup, цель которого – поиск, редактирование и удаление ложной информации, которую добавили в «Википедию» с помощью генеративного ИИ. Вот лишь один из множества результатов его работы. В «Википедии» была статья об османской крепости Амберлисихар, построенной в XV веке. В ней описывалось местоположение и архитектура этой достопримечательности. В действительности такой крепости не существует, и весь текст о ней был великолепным примером галлюцинации ИИ.

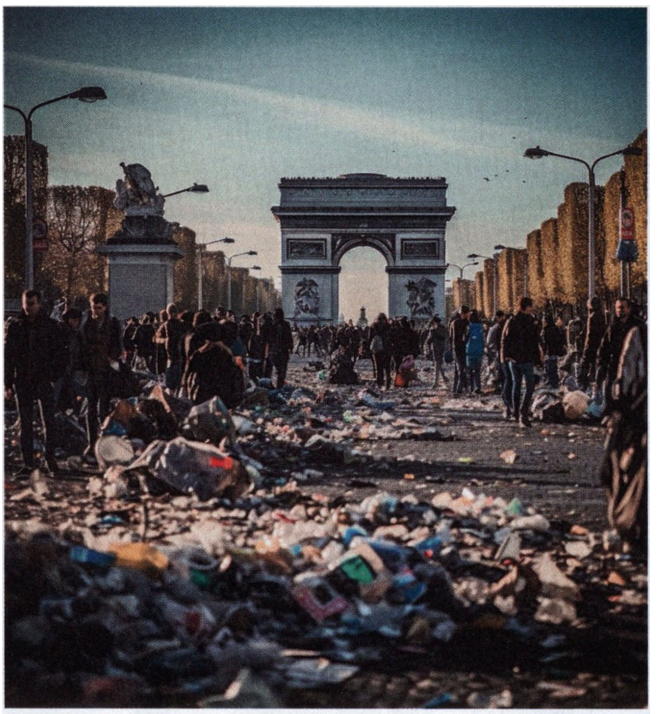
Дуэль искусственных интеллектов

Можно ли отличить сгенерированный текст от написанного человеком? Как правило, да, если речь идет о достаточно длинных текстах. Они часто содержат статистические аномалии, которые может «разглядеть» специальным образом обученная нейросеть. Получается, что один ИИ учится распознавать плоды творчества другого. Так, в 2024 году «Сбер» запустил сервис GigaCheck. Он оценивает вероятность того, что текст был сгенерирован нейросетью. В 2025 году нейросетевая модель, положенная в основу GigaCheck, стала победителем конкурса AINL-Eval по выявлению сгенерированных текстов.

Много ли сегодня в интернете «литературного творчества» ИИ? По данным разработчиков GigaCheck, около 6% из 220 000 новостных русскоязычных статей, написанных в первой половине 2024 года, были созданы нейросетями. Похожие цифры относятся к статьям в журнале Orthopedics & Traumatology: Surgery & Research: 10% аннотаций и 2% самих статей написано с использованием генеративного ИИ.

Общество ИИстины

Каким будет мир контента в ближайшие годы и как научиться в нем жить? Безусловно, объемы сгенерированного контента со временем будут расти. Это касается и текстов, и изображений, и видео. При этом ИИ-контент будет все труднее отличить от созданного людьми. Нам придется свыкнуться с мыслью, что видео или фото уже не смогут быть надежными свидетельствами событий. Точно так же людям когда-то пришлось усвоить, что не всякое печатное слово истинно. В этом смысле появление генеративного ИИ можно



Мусор на парижской улице. Изображение сгенерировано искусственным интеллектом

сравнить с появлением печатного станка или интернета. Оба они многократно упростили распространение любой информации – как истинной, так и ложной.

Скорее всего, появятся многочисленные технические средства проверки фактов и отслеживания источников информации, будет ужесточено рецензирование научных изданий, подвергнется реформе образование, появятся курсы информационной гигиены.

В целом бороться нужно не со сгенерированными, а с некачественными текстами, какова бы ни была их природа. Довольно часто реальная проблема не в том, что «на вход» попадает много информационного мусора, а в том, что «на выходе» не работает фильтр, который должен этот мусор отсеивать. Скажем, редакторы обязаны не пропустить в СМИ недостоверную статью, не важно, написал ее ИИ или нет. То, как применяется та или иная технология, зависит от людей.

«Болезни роста» генеративного ИИ отнюдь не безобидны. Но если человечество не оставит некоторая толика разума, то никакого конца света в информационном пространстве не наступит. Конечно, дети будут посмеиваться над доверчивыми стариками, принимающими генеративный развлекательный контент за чистую монету. Эстеты станут демонстративно предпочитать реальные видео, подобно тому, как сегодня хорошим вкусом считается слушать винил. А жизнь потечет своим чередом в мир будущего, очертания которого мы вряд ли способны представить до конца.