УДК 004.89

ЗНАНИЯ И ТЕХНОЛОГИИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Беляев П.В., Лашков С.А.

МИРЭА – Российский технологический университет, г. Москва, Россия, belyaev_p@mirea.ru

Аннотация. Рассмотрены предпосылки для возникновения возможной угрозы целостности научных знаний ввиду развития систем автоматической генерации информационного контента.

Ключевые слова. Генерация контента, научно-педагогические знания, чат-бот, рерайт.

В современном мире уже можно принять за аксиому, что интернет является общемировой базой знаний, а самый популярный его ресурс www служит интерфейсом к этой базе знаний и по мнению некоторых исследователей обладает признаками сетевой базы данных. Но за последние 2 года в мире цифровых технологий произошли несколько событий, на наш взгляд, сильно недооценённых научно-педагогическим сообществом. Первым значимым событием был запуск в открытом доступе нескольких генеративных чат-ботов, наиболее широко разрекламированным из которых, безусловно является чат-GPT. Вторым, не менее значимым событием, являлся запуск систем автоматической генерации изображений, наиболее известным представителем которых является Midjourney. Далее в течение буквально нескольких месяцев, в сети произошёл взрывной рост числа ресурсов, предлагающих генерацию тестового и графического контента по запросам, рерайт текстов, а в недалёкой перспективе безусловно предполагается генерация видеоконтента, в том числе высококачественного.

Специалисты по информационной безопасности уже сейчас предупреждают о возможном использовании указанных систем в противоправных целях вплоть до необходимости регулирования рынка данных услуг и маркировки автоматически сгенерированного контента. Однако на наш взгляд, сосредоточенность на угрозах, способных нанести явный финансовый ущерб в текущей перспективе, не позволяет рассмотреть проблему более глобально, и мы хотели бы обратить на это внимание в данной статье.

Речь идёт о технической возможности систем автоматической генерации контента воспроизводить информацию в огромных объёмах, нарушая исторически сложившиеся в обществе принципы верификации и классификации текстовых, звуковых, графических и мультимедийных произведений. Для понимания масштаба проблемы достаточно попытаться ответить на несколько вопросов, например:

– Где в настоящее время ищет ответы на вопросы современный школьник, обучающийся колледжа, студент.

– Что в наибольшей степени использует для подготовки к занятиям современный педагог?

– Каким образом производят анализ предметной области темы современные диссертанты?

Бумажные библиотеки фактически ушли в прошлое, и не смотря на наличие значительного количества научно-педагогической информации, издаваемой на бумажных носителях многие специалисты в своих областях используют практически исключительно электронные информационные ресурсы. При чём здесь генерация контента? – очень просто: массовый рерайт текстов со снижением научно-практической ценности работ, а также появившаяся возможность автоматической генерации больших объёмов искажённой или фейковой информации способны снизить практическую ценность как отдельных ресурсов, так и глобальной сети в целом, как источника научно-педагогической информации. До появления систем автоматической генерации контента с верификацией информационного потока вполне справлялись профильные критики и экспертные комиссии, а использование ресурсов интеллектуальных информационных систем для автоматического создания искажённого или откровенно псевдонаучного контента способно создать такой поток дезинформации, что справиться с ним будет тяжело без автоматизации противоположенного процесса – верификации информации. Классический приём спецслужб ещё с давних времён – если не можешь предотвратить утечку информации, то растворі её в дезинформации, и данный приём с появлением систем генерации контента способен обрести новую форму.

Простой эксперимент с чат-ботом YandexGPT, позволяет легко убедиться, что при правильно поставленной задаче можно получить сгенерированную псевдонаучную статью о невозможности математического решения квадратных уравнений, и хотя фейк пока ещё достаточно некачественный, сама возможность генерации такого контента сохраняется.

Ещё более спорный момент – рерайт научных статей с помощью нейросетевых технологий. Помимо откровенного плагиата, использование чат-ботов вроде чат-GPT приводит к тому, что первоначальный смысл статьи незначительно искажается. Эксперимент проводился с пересказом статьи по программированию на языке Python, и после рерайта текст внешне вполне адекватно выглядящей статьи содержал некоторые рекомендации, не выполнимые в интерпретаторе языка Python. Иными словами полученный результат рерайта требовал специальных знаний для возвращения ему научно-практического смысла. При этом внешне статья выглядела вполне научно, и даже имела определённый единый стиль изложения информации, а время генерации статьи объёмом 3 страницы составило около 8 секунд.

Чем подобный псевдонаучный контент может быть опасен ещё, помимо введения в заблуждение неквалифицированных или неподготовленных пользователей? Достаточно проанализировать ещё несколько моментов, связанных с работой современного интернета:

– Что выдают поисковые машины по запросам пользователей?

– Что поисковые машины выводят в тренды выдачи?

Логично предположить, что в условиях отсутствия верификации статей и большого количества фейков очень скоро выдача поисковых машин окажется искажённой. Ещё более грустной окажется попытка использовать наиболее популярные ответы в интернет для обучения систем искусственного интеллекта (как в своё время поступили с проектом wolfram alpha). Рерайты, полученные с помощью ИИ, будучи не вычитанными и выложенными в открытый доступ, способны исказить статистическое представление о научной составляющей предметной области.

В настоящее время наблюдается большое количество публикаций, в которых авторы указывают на опасность автоматической генерации новостного контента с точки зрения влияния на общество. Это безусловно является проблемой, однако общество уже в некоторой степени имеет иммунитет от подобных фейков, и за последнее время растёт число людей, умеющих критически относиться к новостной информации.

Массовое искажение учебной и научной информации может привести к значительно более тяжёлым последствиям для всего научного сообщества. Интернет перестанет быть средством получения корректной информации, которая может быть задействована в научно-педагогической деятельности.

Таким образом, полагаем, что для предотвращения возможного пагубного влияния систем автоматической генерации контента уже сейчас необходимо принимать комплекс превентивных мер, а именно:

– создание перспективных репозиториев знаний, пригодных для автоматического использования, в том числе собственными системами искусственного интеллекта для обучения;

– разработка систем автоматической верификации информации, базой для которых могут служить современные системы поиска плагиата в научных статьях;

– введение понятия «доверенного искусственного интеллекта» для классификации автоматических информационных систем.

О последнем пункте следует сказать особо. Термином «Доверенный искусственный интеллект (AI)» предлагается обозначает то, что система искусственного интеллекта демонстрирует надежность, прозрачность и способность к объяснению своих решений и действий. Основная идея заключается в том, что пользователи нейросети могут доверять полученной информации от искусственного интеллекта и полагаться на них для принятия важных решений или выполнения задач. Соответственно можно ввести и понятие “степени доверия”, как меры уверенности или вероятности, с которой можно полагаться на выходные данные конкретной интеллектуальной информационной системы.

Степень доверия обычно выражается числом в диапазоне от 0 до 1, где 1 означает полную уверенность в полученной информации, а 0 - полное отсутствие уверенности, соответственно. Например, если модель выдаёт информацию со степенью доверия 0.8, то это означает, что она на 80% уверена в правильности выходной информации.

Информация, полученная от искусственного интеллекта должна пройти проверку. Существует несколько этапов проверки данных:

– Валидация данных, заключающаяся в проверке и подтверждении достоверности источников данных. Она может, в том числе, включать в себя проверку источников данных на достоверность и анализ их качества и актуальности.

– Кросс-проверка результатов, предполагающая использование нескольких методов или моделей для сравнения результатов, что может помочь в проверке их доверенности.

– Интерпретируемость моделей искусственного интеллекта, позволяющая понять, как определённые выводы получены в системе. Мо-

жет включать использование таких методов, как логистическая регрессия или решающие деревья, которые легко интерпретируются.

– Автоматическая оценка уверенности модели, позволяющая понять, насколько модель машинного обучения уверенна в своих предсказаниях.

– Экспертная оценка, иначе говоря, привлечение экспертов для проверки результатов, полученных от искусственного интеллекта.

– Мониторинг результатов и обратная связь, позволяющая выявлять ошибки системы на ранних стадиях, повышая доверенность результатов.

– Регулярное обновление и обучение моделей искусственного интеллекта на новых данных для поддержки их актуальности и качества ответов.

Комбинация этих методов может обеспечить более высокую степень доверенности информации от искусственного интеллекта и повысить уверенность в ее правильности и надежности.

Возможно проблема автоматической генерации научного контента при снижении его научно-педагогической полезности пока что не проявила себя в полной мере, однако лавинообразное развития фейковых генераторов и автоматических рерайтеров, вкупе с тестированием откровенно военных возможностей систем искусственного интеллекта требуют относиться и к этой угрозе со всей серьёзностью. Современные системы хранения знаний требуют закладывать в них возможности автоматического использования содержащегося в них контента, и одновременно необходимо уже сейчас принимать некоторые защитные меры для ограничения влияния искажений научной информации.

Литература

1. ИИ как инструмент дезинформации: развитие технологий вызвало 1000% рост фейков в интернете 18:00 / 19 декабря, 2023 https://www.securitylab.ru/news/544718.php

2. Distinguishing academic science writing from humans or ChatGPT with over 99% accuracy using off-the-shelf machine learning tools, VOLUME 4, ISSUE 6, 101426, JUNE 21, 2023

3. Нейросеть ChatGPT пишет рефераты, способные ввести в заблуждение ученых Георгий Голованов13 января 2023 г https://hightech.plus/2023/01/13/neiroset-chatgpt-pishet-referati-sposobnie-vvesti-v-zabluzhdenie-uchenih

4. Искусственный интеллект смог сгенерировать мошенническую научную статью. AI и робототехника 06 июля 2023 https://involta.media/post/iskusstvennyy-intellekt-smog-sgenerirovat-moshennicheskuyu-nauchnuyu-statyu

5. Профанация науки: как компьютер обманывает мировых учёных Якимова Галина 18.04.2017 https://союзженскихсил.рф/communication/forums/science/7770/?sphrase_id=4351311

6. ТОП-10 нейросетей для генерации текста в 2024 году. Маркетинг, СЕО Импульс, 02.02.24 https://vc.ru/marketing/1011419-top-10-neyrosetey-dlya-generacii-teksta-v-2024-godu

7. Как за 15 минут написать статью с помощью нейросети: пошаговое руководство на живом примере, Даниил Шардаков ,https://shardcopywriting.ru/how-to-write-article-with-ai/

8. Как взламывают биометрию и заставляют нейросети придумывать способы атак: топ-6 докладов с PHDays о ML и AI. Positive Technologies. https://habr.com/ru/companies/pt/articles/797241/

KNOWLEDGE AND ARTIFICIAL INTELLIGENCE TECHNOLOGIES

P.V. Belyaev, S.A. Lashkov

MIREA – Russian Technological University, Moscow, Russia, belyaev_p@mirea.ru;

Abstract. The prerequisites for a possible threat to the integrity of scientific knowledge due to the development of systems of automatic generation of information content are considered.

Keywords. Content generation, scientific and pedagogical knowledge, chatbot, rewriting.