ORES

Научно-практический электронный журнал
Оригинальные исследования (ОРИС), *ISSN 2222-0402*
www.ores.su
original-research.ru

УДК 004.056.5, 004.89

БЕЗОПАСНОСТЬ СОВРЕМЕННЫХ ЦИФРОВЫХ СОВЕТНИКОВ: ОСНОВНЫЕ РИСКИ И РЕКОМЕНДАЦИИ ПОЛЬЗОВАТЕЛЯМ

Иконников Александр Владимирович,
студент 1 курса магистратуры направления 10.04.01 «Информационная безопасность»,
МИРЭА — Российский технологический университет, г. Москва
alx.ikona@gmail.com

Аннотация

Статья посвящена некоторым аспектам эксплуатации высокоуровневых языковых консультантов, основанных на технологиях интеллектуального синтеза текстов. Рассматриваются наиболее распространённые угрозы, возникающие вследствие избыточного доверия к ответам подобных систем: формирование ложных утверждений, неточности в знаниях и искажения исходных данных. Отмечается неочевидная проблема конфиденциальности: данные запросы пользователей сохраняются на внешних серверах, что порождает опасения относительно целостности и безопасности информации. Также анализируются сценарии, при которых злоумышленники используют интеллектуальные программы для усовершенствованных методов социальной инженерии. По итогам обзора научных источников 2022–2024 гг. предложены рекомендации, охватывающие критическую оценку сгенерированного текста, ограничение предоставляемых системе сведений и соблюдение общего уровня цифровой культуры.

Ключевые слова: языковой автомат, статистический искусственный интеллект, доверие к алгоритмам, псевдофактический контент, охрана данных, социальная инженерия, кибердоверие, меры предосторожности.

SECURITY OF MODERN DIGITAL ADVISERS: MAIN RISKS AND RECOMMENDATIONS FOR USERS

Aleksandr V. Ikonnikov,
1st year student of the Master's program in the field 10.04.01 «Information Security»,
MIREA — Russian Technological University, Moscow

ABSTRACT

This article addresses selected challenges involved in using state-of-the-art language advisors built on text-synthesizing technologies. It highlights the common threats caused by excessive reliance on such systems, including fabricated assertions, knowledge gaps, and data distortions. The paper notes an important privacy issue: user queries are stored on external servers, thereby raising concerns about data security and integrity. Additionally, the malicious use of intelligent programs for advanced social manipulation methods is discussed. Based on the

literature from 2022–2024, the author proposes recommendations focusing on critically evaluating generated text, limiting the information provided to these systems, and maintaining overall digital awareness.

Keywords: language automaton, statistical artificial intelligence, trust in algorithms, pseudo-factual content, data protection, social manipulation, cyber credibility, precautionary measures.

Цифровые советники на базе статистического искусственного интеллекта (далее — ИИ) быстро завоевали популярность благодаря их видимой способности выдавать убедительные тексты при минимальных усилиях со стороны пользователя. При этом механизмы работы, построенные на вероятностном подборе лексических единиц из обширных обучающих корпусов, нередко скрывают слабые места. Поскольку ядром подобных алгоритмов остаётся статистическое моделирование, сформированный ответ может выглядеть правдоподобно, но содержать грубые искажения. Порой это проявляется в так называемых «галлюцинациях», когда система с уверенностью приводит несуществующие факты [1–2].

Отдельной темой выступает стремление некоторых людей пользоваться выводами языковых алгоритмов без всякой последующей проверки. Заметив, что советник способен верно ответить на часть вопросов, пользователи начинают некритично копировать его рекомендации в жизненные сферы: от обучения до медицинских решений. В подобных случаях могут возникать ситуации, когда модель выдаёт длинные рассуждения, в которых часть тезисов не согласована с реальностью. Однако человеку, которому импонирует плавность речи, сложно вовремя распознать ошибку [1].

Среди наиболее напряжённых вопросов выделяют конфиденциальность. Поскольку языковой автомат действует в облачной среде, вводимые сведения, включая личные данные, могут сохраняться на серверах разработчиков и использоваться для последующего обучения. Иногда их включают в общий массив, на основании которого улучшаются будущие версии цифрового советника [3–4]. Сложность в том, что никто не гарантирует отсутствие сбоев или внешних вмешательств в компьютерные системы. Весной 2023 года произошёл случай, когда некоторые пользователи вдруг увидели обрывки чужих историй общения, а также элементы чужих платёжных реквизитов, вызвав тем самым широкий резонанс [7–8].

Корпорации, стремясь защитить коммерческие сведения, вводят ограничения на применение публичных языковых ассистентов внутри организаций, ведь даже фрагменты кода или отрывки внутренних документов могут быть случайно проиндексированы алгоритмом. При этом часть компаний-поставщиков подобных систем декларирует строгую политику приватности, но гарантий невозможности утечек всё же нет. Нередко и публичные жалобы пользователей на то, что модель внезапно «вспоминает» чужие тексты, демонстрируя возможное смешение данных внутри механизма обучения [4].

Злоумышленники, наблюдая за ростом качества текстогенерации, интенсивно включают эти инструменты в свой арсенал. Наиболее простым примером являются массированные рассылки поддельных сообщений, иногда на редких языках, где система с высокой точностью подражает типичной манере письма. Если раньше мошенники выдавали себя ошибочными фразами, то сейчас грамматические проколы не так заметны, а значит, жертве сложнее распознать подвох [5; 9–10]. Имитируя стиль крупных организаций или госучреждений, они получают доступ к личным сведениям доверчивых лиц. В результате существенно возрастает риск фишинговых приёмов и даже точечных

манипуляций, когда электронный советник помогает преступнику обосновывать различные схемы обмана.

Характерно и то, что обсуждаемые средства искусственного интеллекта дополняются другими технологиями, связанными с синтезом речи и подделкой визуальной информации (deepfake). Мошенники способны накладывать сгенерированные голоса руководителей на телефонный звонок бухгалтеру, требуя неотложного перечисления средств на счёт. Если человек не проверяет источники, подобное взаимодействие выглядит правдоподобно.

Тем не менее меры защиты существуют. Во-первых, опытные специалисты советуют максимально критично относиться к любым ответам цифровых советников и самостоятельно проверять факты. Во-вторых, не следует вводить в программу ключевые реквизиты и пароли, поскольку всё, что пользователь сообщает в диалоговом окне, уходит на сервера сторонней организации [6]. Кроме того, практикующие эксперты в области информационных технологий зачастую рекомендуют по возможности ограничивать сохранение истории сообщений, если эта функция реализована. Полезно также пользоваться только официальными и проверенными источниками скачивания подобных сервисов, поскольку всё чаще появляются поддельные приложения, предлагающие «дополнительные» возможности.

Немаловажным выглядит аспект цифровой культуры. В условиях, когда алгоритмы могут формировать псевдоновостные материалы и убедительные тексты для сетевых диалогов, незнакомый с подобными приёмами человек становится особенно уязвим к дезинформации и киберугрозам [1; 3]. Наработка навыка проверки исходных ссылок, осторожного отношения к громким заявлениям и использования двухфакторной аутентификации формируют первую линию обороны. Часто бывает, что тот, кто знает о существовании высокоточного статистического ИИ, уже не склонен верить длинным письмам, пришедшим «невесть откуда».

Разработчики, со своей стороны, пытаются нарастить качество фильтров для борьбы с неадекватными выходными данными, однако задача многоаспектна. Полностью блокировать неуместный контент возможно лишь частично, да и ошибки никто не отменял. Параллельно в разных государствах продолжается дискуссия о законодательных нормах, регулирующих сбор и обработку пользовательских запросов. Отдельные эксперты указывают, что, помимо «уменьшения галлюцинаций», нужен и «объяснимый ИИ», способный раскрывать человеку логику формируемых ответов [11–12].

Чтобы свести к минимуму опасности, необходимо учитывать несколько основных правил. Во-первых, рекомендуется постоянно проверять те данные, которые кажутся сомнительными, прибегая к независимым источникам либо специализированным справочникам. Во-вторых, нужно помнить, что любая информация, введённая в облачный сервис, может оказаться сохранённой и в будущем воспроизведённой в ином контексте, поэтому стоит исключать ввод приватных сведений. В-третьих, при обнаружении подозрительных писем, авторы которых якобы ссылаются на «цифровых советников», важно проявлять повышенную осмотрительность: уточнять официальные контакты, совершать контрольные звонки и использовать только безопасные способы связи [6; 7–8].

Суммируя все эти аспекты, следует признать, что язык, синтезированный интеллектуальными алгоритмами, отличается особой привлекательностью для пользователей, но одновременно влечёт за собой новые типы уязвимостей. Большинство проблем коренится в самом характере систем статистического прогнозирования, а также в их повсеместной интеграции: люди стремятся экономить время, поручая алгоритму информационные задачи, и не замечают потенциала дезинформации или вмешательства третьих сторон. Однако при разумном сочетании технологических фильтров, законодательных норм и грамотного поведения можно значительно уменьшить риск

злоупотреблений и извлечь выгоду из подобного инструмента. В условиях стремительного развития интеллектуальных ассистентов полезно сохранять критический настрой и осознавать, что гладкость языка не гарантирует истинность содержания.

Список литературы:

1. Тушканов В., Сергеев В., Шлычкова Ю., Очеповский А. Влияние генеративного ИИ на кибербезопасность: факты и прогнозы // Securelist (Kaspersky): сайт. 11 декабря 2023. URL: https://securelist.ru/story-of-the-year-2023-ai-impact-on-cybersecurity/108558/ (дата обращения: 9.05.2025). Режим доступа: свободный.
2. Huang L., Yu W., Ma W., Zhong W., Feng Z., Wang H., Chen Q., Peng W., Feng X., Qin B., Liu T. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions // ACM Transactions on Information Systems. 2024. Т. 1. № 1. Ст. 1. arXiv:2311.05232v2 [cs.CL].
3. Schmitt M., Flechais I. Digital Deception: Generative Artificial Intelligence in Social Engineering and Phishing // Artificial Intelligence Review. 2024. Т. 57. Ст. 324. DOI: 10.1007/s10462-024-10973-2.
4. He Q., Zeng J., Huang W., Chen L., Xiao J., He Q., Zhou X., Chen L., Wang X., Huang Y., Ye H., Li Z., Chen S., Zhang Y., Gu Z., Liang J., Xiao Y. Can Large Language Models Understand Real-World Complex Instructions? // Journal of Electronic Science and Technology. 2025. Т. 23, № 1. С. 100301. arXiv:2309.09150v2 [cs.CL].
5. U.S. Department of Health and Human Services (HHS). Social Engineering Attacks Targeting the HPH Sector // Health Sector Cybersecurity Coordination Center (HC3): сайт. 11 апреля 2024. 50 с. URL: https://www.hhs.gov/sites/default/files/social-engineering-targeting-the-hph-sector-tlpclear.pdf (дата обращения: 10.05.2025). Режим доступа: свободный.
6. D'Andrea A., Cutler A., Guccione D. Безопасен ли ChatGPT? Советы по безопасному использованию ChatGPT в 2024 году // Keeper Security (blog): сайт. 23 июля 2024. URL: https://keepersecurity.com/blog/ru/2024/07/23/is-chatgpt-safe/ (дата обращения: 10.05.2025). Режим доступа: свободный.
7. Zorz Z. A bug revealed ChatGPT users' chat history, personal and billing data // Help Net Security: сайт. 27 марта 2023. URL: https://www.helpnetsecurity.com/2023/03/27/chatgpt-data-leak/ (дата обращения: 10.05.2025). Режим доступа: свободный.
8. Известия. Мошенники начали использовать ChatGPT для фишинга // Интернет-новости Iz.ru: сайт. 28 марта 2023. URL: https://iz.ru/1489507/2023-03-28/moshenniki-nachali-ispolzovat-chatgpt-dlia-fishinga (дата обращения: 10.05.2025). Режим доступа: свободный.
9. Лаборатория Касперского — Центр знаний. Что такое ChatGPT и безопасно ли им пользоваться? // Лаборатория Касперского: сайт. 2023. URL: https://www.kaspersky.ru/resource-center/preemptive-safety/is-chatgpt-safe (дата обращения: 11.05.2025). Режим доступа: свободный.
10. Иванов П. Отчет: после запуска ChatGPT число фишинговых атак выросло на 1265% // Forklog: сайт. 31 октября 2023. URL: https://forklog.com/news/ai/otchet-posle-

zapuska-chatgpt-chislo-fishingovyh-atak-vyroslo-na-1265 (дата обращения: 11.05.2025). Режим доступа: свободный.
11. Arrieta A.B., Diaz-Rodriguez N., Del Ser J., Bennetot A., Tabik S., Barbado A., Garcia S., Gil-Lopez S., Molina D., Benjamins R., Chatila R., Herrera F. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI // Information Fusion. 2020. T. 58. C. 82-115. DOI: 10.1016/j.inffus.2019.12.012.
12. Floridi L., Taddeo M. What is data ethics? // Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2016. T. 374. № 20160360. C. 20160360. DOI: 10.1098/rsta.2016.0360.

References:

1. Tushkanov V., Sergeev V., Shlychkova Yu., Ochepovsky A. The Impact of Generative AI on Cybersecurity: Facts and Forecasts [Online] // Securelist (Kaspersky): website. December 11, 2023. URL: https://securelist.ru/story-of-the-year-2023-ai-impact-on-cybersecurity/108558/ (accessed: May 9, 2025). Access mode: free.
2. Huang L., Yu W., Ma W., Zhong W., Feng Z., Wang H., Chen Q., Peng W., Feng X., Qin B., Liu T. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions // ACM Transactions on Information Systems. 2024. Vol. 1. № 1. Art. 1. arXiv:2311.05232v2 [cs.CL].
3. Schmitt M., Flechais I. Digital Deception: Generative Artificial Intelligence in Social Engineering and Phishing // Artificial Intelligence Review. 2024. Vol. 57. Art. 324. DOI: 10.1007/s10462-024-10973-2.
4. He Q., Zeng J., Huang W., Chen L., Xiao J., He Q., Zhou X., Chen L., Wang X., Huang Y., Ye H., Li Z., Chen S., Zhang Y., Gu Z., Liang J., Xiao Y. Can Large Language Models Understand Real-World Complex Instructions? // Journal of Electronic Science and Technology. 2025. Vol. 23, № 1. P. 100301. arXiv:2309.09150v2 [cs.CL].
5. U.S. Department of Health and Human Services (HHS). Social Engineering Attacks Targeting the HPH Sector [Online] // Health Sector Cybersecurity Coordination Center (HC3). April 11, 2024. 50 p. URL: https://www.hhs.gov/sites/default/files/social-engineering-targeting-the-hph-sector-tlpclear.pdf (accessed: May 10, 2025). Access mode: free.
6. D'Andrea A., Cutler A., Guccione D. Is ChatGPT Safe? Tips for Secure ChatGPT Usage in 2024 [Online] // Keeper Security (blog). July 23, 2024. URL: https://keepersecurity.com/blog/ru/2024/07/23/is-chatgpt-safe/ (accessed: May 10, 2025). Access mode: free.
7. Zorz Z. A bug revealed ChatGPT users' chat history, personal and billing data [Online] // Help Net Security. March 27, 2023. URL: https://www.helpnetsecurity.com/2023/03/27/chatgpt-data-leak/ (accessed: May 10, 2025). Access mode: free.
8. Izvestia. Scammers began using ChatGPT for phishing [Online] // Online news Iz.ru. March 28, 2023. URL: https://iz.ru/1489507/2023-03-28/moshenniki-nachali-ispolzovat-chatgpt-dlia-fishinga (accessed: May 10, 2025). Access mode: free.

9. Kaspersky Lab - Knowledge Center. What is ChatGPT and is it safe to use? [Online] // Kaspersky Lab. 2023. URL: https://www.kaspersky.com/resource-center/preemptive-safety/is-chatgpt-safe (accessed: May 11, 2025). Access mode: free.
10. Ivanov P. Report: After ChatGPT Launch, Phishing Attacks Increased by 1265% [Online] // Forklog. October 31, 2023. URL: https://forklog.com/news/ai/otchet-posle-zapuska-chatgpt-chislo-fishingovyh-atak-vyroslo-na-1265 (accessed: May 11, 2025). Access mode: free.
11. Arrieta A.B., Diaz-Rodriguez N., Del Ser J., Bennetot A., Tabik S., Barbado A., Garcia S., Gil-Lopez S., Molina D., Benjamins R., Chatila R., Herrera F. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI // Information Fusion. 2020. Vol. 58. P. 82-115. DOI: 10.1016/j.inffus.2019.12.012.
12. Floridi L., Taddeo M. What is data ethics? // Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2016. Vol. 374. № 20160360. P. 20160360. DOI: 10.1098/rsta.2016.0360.