УДК 004.89+651.5

Белов Илья Игоревич
Российский государственный гуманитарный университет,
Всероссийский научно-исследовательский институт
документоведения и архивного дела, г. Москва, Российская Федерация
SPIN-код: 3080-1811, AuthorID: 1041018, belov@vniidad.ru

Ilya I. Belov
Russian State University for the Humanities,
All-Russian Scientific and Research Institute
for Records and Archives Management, Moscow, Russian Federation
SPIN-code: 3080-1811, AuthorID: 1041018, belov@vniidad.ru

Обзорная статья

INTERPARES TRUST AI: ВОПРОСЫ БЕЗОПАСНОСТИ ПРИМЕНЕНИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В РАБОТЕ С ДОКУМЕНТАМИ

Представлен обзор публикаций участников проекта InterPARES Trust AI по вопросам применения технологий искусственного интеллекта в работе с документированной информацией. Рассматриваются инструменты технологии искусственного интеллекта, такие как компьютерное зрение (computer vision), обработка естественного языка (natural language processing, NLP), технологии повышения конфиденциальности данных (privacy enhancing technologies, PETs), уязвимости больших языковых моделей (large language models, LLMs), применительно к проблеме обеспечения безопасности систем искусственного интеллекта при работе с документами. Делается вывод о необходимости формирования подходов к обеспечению безопасности применяемых систем искусственного интеллекта.

Ключевые слова: архивное дело, делопроизводство, зарубежный опыт, системы искусственного интеллекта.

Для цитирования: Белов И.И. InterPARES Trust AI: вопросы безопасности применения искусственного интеллекта в работе с документами // Вестник ВНИИДАД. 2024. № 6. С. 111–117.

Review article

INTERPARES TRUST AI: SECURITY ISSUES IN USING ARTIFICIAL INTELLIGENCE IN DOCUMENTS OPERATING

An overview of publications by participants in the InterPARES Trust AI project on the application of artificial intelligence technologies in operating with documented information is presented. The artificial intelligence technology tools such as computer vision, natural language processing (NLP), privacy enhancing technologies (PETs), and vulnerabilities of large language models (LLMs) in relation to the problem of

Поступила в редакцию: 23.11.2024
Поступила после рецензирования: 27.11.2024
Принята к публикации: 29.11.2024

Received: 23.11.2024
Revised: 27.11.2024
Accepted: 29.11.2024

ensuring the security of artificial intelligence systems when documents operating are considered. The need to develop approaches to ensuring the security of the applied artificial intelligence systems is highlighted.

Keywords: archiving, artificial intelligence systems, foreign experience, record keeping.

For citation: Belov I.I InterPARES Trust AI: security issues in using artificial intelligence in documents operating. *Vestnik VNIIDAD = Herald of VNIIDAD*. 2024;6:111–117. (In Russian).

В последние годы в России и за рубежом специалистами уделяется особое внимание проблеме цифровой трансформации делопроизводства и архивного дела. Инновационные технологии, такие как искусственный интеллект, находят применение в текущей работе с документами и архивной практике. В Российской Федерации такие тенденции наблюдаются в государственном и коммерческих секторах, что свидетельствует о понимании потенциала применения таких технологий.

На данном этапе функционал систем электронного документооборота в России уже претерпевает изменения посредством использования искусственного интеллекта. Российская ИТ-компания Directum разработала систему Directum RX[1], в которой реализованы функции по распознаванию текстов документов, автоматической классификации документов, автоматическому определению ответственных лиц и вынесению резолюций, формированию текстов документов по предварительному запросу пользователя и др. Система способна улучшать полученные результаты на основе машинного обучения (machine learning, ML). Эта система внесена в Реестр отечественного программного обеспечения и является первым программным продуктом в области электронного документооборота с пометкой «относится к сфере искусственного интеллекта»[2]. Деятельность отечественных архивов также трансформируется с помощью искусственного интеллекта. Такие проекты реализуются в Государственном архиве Российской Федерации (ГА РФ). Ярким примером является семантическая поисковая система НИКА (https://garf-nika.ru), которая использует отечественную языковую модель и осуществляет интеллектуальный или семантический поиск по запросам пользователей в рамках архивных данных, хранящихся в ГА РФ[3].

Использование технологий искусственного интеллекта в работе с документами, которые содержат конфиденциальную информацию, требует соблюдения определенных требований и мер защиты ввиду возможных утечек таких данных либо случайного раскрытия информации и соответствующих последствий. Применительно к области делопроизводства и архивного дела, системы обрабатывают данные, которые содержатся в документах, в том числе ограниченного доступа, в связи с чем необходимо выработать современные подходы к защите информации и обеспечению ее

[1] Directum RX: Интеллектуальная система цифровизации процессов и документов для крупных, средних организаций и государственных органов // Directum, 2024. URL: https://www.directum.ru/products/directum (дата обращения: 20.11.2024).

[2] Система Directum // Реестр Российского программного обеспечения. URL: https://reestr.digital.gov.ru/reestr/305849/?sphrase_id=3102679 (дата обращения: 20.11.2024).

[3] Справка проекта Семантическая поисковая система НИКА. URL: https://garf-nika.ru/%D0%A1%D0%9F%D0%A1_%D0%9D%D0%98%D0%9A%D0%90/Docs.html (дата обращения: 20.11.2024).

конфиденциальности при использовании искусственного интеллекта.

В отечественной практике тема обеспечения безопасности и защиты информации при применении искусственного интеллекта в работе с документами пока недостаточно изучена и не получила широкого освещения. Оказать содействие в решении этой проблемы может анализ зарубежной теории и практики цифровой трансформации управления документами с применением искусственного интеллекта на примере изучения деятельности исследовательского проекта InterPARES Trust AI (2021–2026) (далее – проект) [1].

Международный проект «Международное исследование аутентичности документов постоянного срока хранения в электронных системах» (The International Research on Permanent Authentic Records in Electronic Systems – InterPARES Project) осуществляется с 1999 года и объединяет ученых и специалистов в области управления документами, которыми разрабатываются основы для долгосрочного хранения документов, созданных в информационных системах[4]. На протяжении своего существования проект проходил несколько фаз, в течение которых специалистами изучались определенные аспекты работы с электронными документами и соответствующие технологии. С 2021 года проект перешел в очередную новую фазу – InterPARES Trust AI, целью которой является исследование практических разработок искусственного интеллекта и создание методологической базы для внедрения технологий искусственного интеллекта в работе с электронными документами с учетом возможных преимуществ и рисков [1]. Окончание текущего этапа проекта запланировано на 2026 год.

В рамках проекта реализуется исследование RA02 «Тематическое исследование по извлечению и идентификации документов, содержащих персональные данные и конфиденциальные данные личного характера, для долгосрочного хранения» (Case study on extraction and identification of records containing personal data and sensitive personal data for long term preservation)[5], основной задачей которого является разработка алгоритма распознавания и извлечения неструктурированной конфиденциальной информации, содержащей персональные данные, в оцифрованных документах путем применения методов искусственного интеллекта, в частности, алгоритмов машинного обучения. Персональную идентифицируемую информацию (personally identifiable information, PII), коммерческую информацию, финансовые и медицинские данные, которые содержатся в документах, в исследовании называют конфиденциальными (чувствительными) данными (sensitive data).

Результаты данного направления исследования обсуждаются в том числе в опубликованном InterPARES Trust AI специальном информационном сборнике «Искусственный интеллект и документальное наследие» (“Artificial intelligence and documentary heritage”) [2].

[4] The International Research on Permanent Authentic Records in Electronic Systems - InterPARES Project: official website [Официальный сайт Международного проекта «Международное исследование аутентичности документов постоянного срока хранения в электронных системах»]. URL: http://www.interpares.org/index.htm (дата обращения: 20.11.2024).

[5] Research studies. Abstracts – Project list by title [Научные исследования. Рефераты – Список проектов]. InterPARES Trust. URL: https://interparestrustai.org/trust/about_research/studies (дата обращения: 20.11.2024).

Проблему обеспечения безопасности в процессе применения искусственного интеллекта в работе с документами, содержащими информацию ограниченного доступа, в рамках проекта обсуждают В.Л. Лемье (V.L. Lemieux) и Д. Вернер (J. Werner) в статье «Защита конфиденциальной информации, содержащейся в электронных документах: потенциал технологий, повышающих уровень конфиденциальности данных» (“Protecting privacy in digital records: the potential of privacy-enhancing technologies”) [3], а также Г. Бхатия (G. Bhatia), М.Б. Нагуди (M. Nagoudi), Х. Чавушоглу (H. Cavusoglu), М. Абдул-Магид (M. Abdul-Mageed) в статье «FinTral: Семейство многофункциональных финансовых больших языковых моделей уровня GPT-4» (“A Family of GPT-4 level multimodal financial large language models”) [4].

Исходя из этого, можно сказать, что защита данных в системах искусственного интеллекта и обработка искусственными интеллектом конфиденциальной информации, которая содержится в документах, для обучения моделей и дальнейшего непосредственного выполнения задач системами искусственного интеллекта и, как следствие, преобразование практики работы с документами являются актуальными задачами.

Этому вопросу также посвящены статьи А. Алькобы (A. Alcoba), П. Хоманн (P. Hohmann) и Д. Судермана (J. Suderman) «Сбор данных в архивах для защиты информации» (“Datafying archives for privacy protection”) [5], В.Л. Лемье (V.L. Lemieux) «Балансирование: навигация по сгенерированным искусственным интеллектом связям, конфиденциальность и доступность архивной информации» (“Balancing act: navigating the nexus of AI, privacy, and accessibility in archives”) [6], но он требует более детального рассмотрения.

На сегодняшний день во многих странах действуют законодательные нормы, касающиеся обеспечения конфиденциальности информации. По мнению специалистов – участников InterPARES, ввиду передачи большого количества документов, содержащих информацию ограниченного доступа, в архивные учреждения и отсутствия эффективных инструментов для их быстрой обработки архивистам часто приходится полностью закрывать доступ к такой документации [5]. Решением является применение технологий компьютерного зрения (computer vision) и обработки естественного языка (natural language processing, NLP), которые позволяют определять и классифицировать элементы таких документов с целью идентификации наиболее конфиденциальной (чувствительной) информации и предпринимать дальнейшие соответствующие меры по их защите. В качестве примера специалисты приводят результаты использования инструмента Microsoft Presidio[6], который автоматически распознает конфиденциальные (чувствительные) данные в документах, такие как имена, адреса, контактная информация и др. Авторы указывают, что программа неспособна в полной мере считывать контекст, особенно применительно к фрагментам исторических писем или дневников, так как конфиденциальная (чувствительная) информация в них скрыта в нестандартных формулировках или метафорах. Для более точного распознавания необходима разработка специальных моделей, которые способны идентифицировать

[6] PII entities supported by Presidio. Presidio: Data Protection and De-identification SDK [PII-объекты, поддерживаемые Presidio. Presidio: SDK для защиты данных и деидентификации] / Microsoft Presidio. URL: https://microsoft.github.io/presidio/supported_entities/ (дата обращения: 20.11.2024).

сложную контекстную информацию в архивных документах.

Для обучения моделей искусственного интеллекта и их последующей эксплуатации в работе с архивными документами предлагается использовать технологии повышения конфиденциальности данных (privacy-enhancing technologies, PETs), к которым относятся [3]:

• обезличивание данных, т.е. обработка данных с удалением персональных идентификаторов;

• шифрование данных для их совместного анализа несколькими субъектами;

• создание синтетических данных, которые схожи с оригинальными, но не содержат реальных персональных данных.

По мнению специалистов проекта, применение этих технологий может помочь найти своего рода баланс между защитой конфиденциальных данных и обеспечением доступности информации для общественного пользования. Одним из перспективных направлений использования PETs является так называемое «федеративное обучение» (federated learning) – обучение моделей искусственного интеллекта без передачи данных в централизованное хранилище [3]. Также к таким направлениям относятся интеграция с облачными хранилищами и разработка алгоритмов обезличивания данных при загрузке документов в информационные системы с публичным доступом в режиме реального времени.

В публикациях обращается внимание, что применение больших языковых моделей (large language models, LLMs), таких как ChatGPT, может также ускорить идентификацию конфиденциальных (чувствительных) данных в документах. Но их использование может привести к утечкам конфиденциальной информации, особенно в облачной среде. Одним из решений является тренировка собственных моделей, что является затратным и трудоемким процессом, а также их локальное размещение на серверах организаций. Данный подход реализовали сотрудники Университета Британской Колумбии, которыми была обучена большая языковая модель FinTral, предназначенная для обработки финансовой документации и способная идентифицировать в них конфиденциальные (чувствительные) данные [4].

Описанный выше опыт также свидетельствует о важности формирования и сохранения «параданных», т.е. данных о том, каким образом алгоритмы искусственного интеллекта разрабатываются, обучаются и используются [7]. К ним относятся данные о параметрах модели, контексте ее использования и обучающие данные. Хранение параданных позволяет повысить подотчетность систем искусственного интеллекта и прозрачность алгоритмов, а также предотвращать ошибки, возникающие при их применении в работе с документированной информацией. Участниками InterPARES Trust AI предлагается выработка архивистами совместно с техническими специалистами стандартного механизма формирования и хранения параданных документных процессов.

Так как технологии искусственного интеллекта находят все большее применение в отечественной практике работы с документами, то перечисленные выше аспекты являются актуальными в Российской Федерации и требуют внимания специалистов. Введение экспериментального правового режима по установлению специального регулирования в целях создания необходимых условий для разработки и внедрения технологий искусственного интеллекта на территории города Москвы позволило в определенной степени решить проблемы обработки искусственным ин-

теллектом обезличенных персональных данных, содержащихся в документах, на некоторое время[7]. Это способствует облегченному тестированию и внедрению технологий искусственного интеллекта во всех сферах деятельности. В дальнейшем необходимо установление постоянных правовых механизмов регулирования взаимодействия систем искусственного интеллекта с конфиденциальными (чувствительными) данными и информацией ограниченного доступа, а также дальнейшая разработка требований безопасности к системам.

[7] О проведении эксперимента по установлению специального регулирования в целях создания необходимых условий для разработки и внедрения технологий искусственного интеллекта в субъекте Российской Федерации – городе федерального значения Москве и внесении изменений в статьи 6 и 10 Федерального закона «О персональных данных»: федер. закон от 24.04.2020 № 123-ФЗ // Электронный фонд правовой и нормативно-технической документации. АО «Кодекс». URL: https://docs.cntd.ru/document/564747621 (дата обращения: 20.11.2024).

Финансирование. Работа выполнена в рамках выполнения Плана научно-исследовательских и опытно-конструкторских работ по государственному заданию Федерального архивного агентства на 2024 год по теме 3.1 «Анализ зарубежных нормативных актов и методических разработок, международных стандартов и их проектов, материалов официальных интернет-сайтов, конференций, научных проектов по проблемам применения современных технологий в управлении документами и архивном деле. Аналитический обзор». Рег. № ЕГИСУ НИОКТР 124020800053-3.

Funding. The study was done as part of the implementation of the Research and Development Plan (R&D) Based on State Assignment of The Federal Archival Agency for 2023 on topic 3.1 "Analysis of foreign regulations and methodological developments, international standards and their projects, materials of official websites, conferences, scientific projects on the problems of modern technologies implementation in records management and archiving. Analytical review". Reg. No. in Unified state system of R&D accounting (EGISU NIOKTR) 124020800053-3.

Конфликт интересов. Автор заявляет об отсутствии потенциального конфликта интересов.

Conflict of interests. The author declares no conflict of interest.

Список источников / References

1. *InterPARES Trust AI (2021–2026)*: official website [Официальный сайт проекта InterPARES Trust AI (2021–2026)]. URL: https://interparestrustai.org/ (дата обращения: 20.11.2024).

2. *Artificial Intelligence and Documentary Heritage*. Newsletter 2024 Special Issue 2024 InterPARES Trust AI is a cooperating institution of the Memory of the World Sub on Education and Research (SCEaR) [Искусственный интеллект и документальное наследие. Специальный выпуск 2024 InterPARES Trust AI Подразделения по образованию и исследованиям

«Память мира» (SCEaR)]. Edited by Luciana Duranti and Corinne Rogers. SCEaR, IAC, UNESCO Memory of the World Programme. 2024. 99 p. URL: https://interparestrustai.org/assets/public/dissemination/SCEaRNewsletterSpecialIssue2024ArtificialIntelligence.pdf (дата обращения: 20.11.2024).
3. Lemieux V.L., Werner J. Protecting Privacy in Digital Records: The Potential of Privacy-Enhancing Technologies [Защита конфиденциальной информации, содержащейся в электронных документах: потенциал технологий, повышающих уровень конфиденциальности данных]. *ACM Journal on Computing and Cultural Heritage*. 2024;83:1–18. URL: https://dl.acm.org/doi/10.1145/3633477 (дата обращения: 20.11.2024).
4. Bhatia G., Nagoudi M.B., Cavusoglu H., Abdul-Mageed M. FinTral: *A Family of GPT-4 Level multimodal financial large language models* [FinTral: Семейство многофункциональных финансовых больших языковых моделей уровня GPT-4]. URL: https://aclanthology.org/2024.findings-acl.774.pdf (дата обращения: 20.11.2024).
5. Alcoba A., Hohmann P., Suderman J. Datafying Archives for Privacy Protection [Сбор данных в архивах для защиты информации]. *Artificial Intelligence and Documentary Heritage*. 2024. p. 34–38. URL: https://interparestrustai.org/assets/public/dissemination/SCEaRNewsletterSpecialIssue2024ArtificialIntelligence.pdf (дата обращения: 20.11.2024).
6. Lemieux V.L. Balancing act: navigating the nexus of AI, privacy, and accessibility in archives [Балансирование: навигация по сгенерированным искусственным интеллектом связям, конфиденциальность и доступность архивной информации]. *Artificial Intelligence and Documentary Heritage*. 2024. p. 39–42. URL: https://interparestrustai.org/assets/public/dissemination/SCEaRNewsletterSpecialIssue2024ArtificialIntelligence.pdf (дата обращения: 20.11.2024).
7. Franks P.C. The crucial role of paradata in AI governance [Критическая роль параданных в управлении искусственным интеллектом]. *Artificial Intelligence and Documentary Heritage*. 2024. p. 54–60. URL: https://interparestrustai.org/assets/public/dissemination/SCEaRNewsletterSpecialIssue2024ArtificialIntelligence.pdf (дата обращения: 20.11.2024).

ИНФОРМАЦИЯ ОБ АВТОРЕ

Белов Илья Игоревич, ассистент кафедры автоматизированных систем документационного обеспечения управления Российского государственного гуманитарного университета (РГГУ); научный сотрудник отдела документоведения Всероссийского научно-исследовательского института документоведения и архивного дела (ВНИИДАД), Москва, Российская Федерация.

INFORMATION ABOUT THE AUTHOR

Ilya I. Belov, assistant of the Department of Automated Systems of Management Document Support, Russian State University for the Humanities (RSUH); researcher of Document Science Department of the All-Russian Scientific and Research Institute for Records and Archives Management (VNIIDAD), Moscow, Russian Federation.