УДК 004.838.2

ОБЕСПЕЧЕНИЕ КИБЕРБЕЗОПАСНОСТИ В УСЛОВИЯХ ПРИМЕНЕНИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Слабодчиков А.А.
Открытое акционерное общество «АГАТ – системы управления» – управляющая компания холдинга «Геоинформационные системы управления», г. Минск, Республика Беларусь

Аннтотация. В современном мире, где технологии проникают в каждый аспект нашей жизни, кибербезопасность становится все более важной. Она охватывает практики защиты компьютерных систем, сетей и данных от цифровых атак.

В статье мы рассмотрим вопросы обеспечения кибербезопасности в условиях применения искусственного интеллекта.

Ключевые слова: кибербезопасность, искусственный интеллект, развитие, угроза, законодательство, обучение, промпт,

В эпоху цифровизации кибербезопасность становится особенно актуальной. Искусственный интеллект (далее – ИИ) вносит значительные изменения в методы защиты информационных систем различных государственных органов и организаций (предприятий) различных форм собственности.

Неконтролируемое внедрение ИИ может способствовать серьезным рискам, связанным с обеспечением кибербезопасности как предприятий, так и целых отраслей. Очевидно, что использование ИИ позволяет автоматизировать и оптимизировать бизнес-процессы предприятий, что значительно может повысить эффективность таких предприятий. Использование ИИ может помочь в обработке больших объемов данных, а в контексте обеспечения кибербезопасности предотвращать атаки, включая самые изощренные связанные с современными хакерскими группировками, классические виды киберугроз.

Несмотря на многочисленные преимущества, интеграция ИИ в системы кибербезопасности несет в себе и определенные риски. Сложность контроля за алгоритмами ИИ, возможность ошибок или злоупотреблений требуют тщательного подхода к разработке и внедрению таких систем. Обеспечение надежной защиты в условиях использования ИИ означает не только применение новейших технологий, но и постоянное обновление знаний и методов в области кибербезопасности.

В данной статье предлагается рассмотреть основные подходы к обеспечению кибербезопасности в условиях применения ИИ.

Технический прогресс в области ИИ происходит быстрее, чем общество и государства могут реагировать на него. В связи с чем возникает потребность предусматривать на законодательном уровне формирование правовых актов, детализирующих различные принципы и нормы в области ИИ.

Хотелось бы выделить наиболее критичные направления, которые необходимо принимать во внимание при формировании законодательства в области ИИ.

Имеется необходимость законодательно закрепить соответствующие вектора развития в области ИИ, в том числе и по созданию модели угроз кибербезопасности ИИ.

В первую очередь необходимо определить угрозы, связанные не только с этапом внедрения ИИ в инфраструктуру предприятия, но и угрозы, которые существуют на этапе создания, внедрения и управления ИИ в организации, разработчиков, ответственных за обучение моделей ИИ, специалистов кибербезопасности,

архитекторов, интегрирующих ИИ в бизнес-процессы специалистов, руководителей предприятий, специалистов, оценивающих риски внедрение ИИ на предприятии.

К основным видам угроз ИИ можно отнести следующие угрозы:

угрозы, связанные с входными данными в системы ИИ;

угрозы, связанные с инфраструктурой предприятия;

угрозы, связанные с низким качеством встроенных механизмов защиты;

угрозы, связанные с приложением ИИ;

угрозы, связанные с моделью обучения ИИ;

угрозы, связанные с ИИ-агентами;

угрозы, связанные с хищением (копия, цифровой портрет) ИИ.

На текущий момент существует ряд международных подходов по кибербезопасности ИИ, в том числе OWASP, MITRE, NIST, которые необходимо учитывать, но не всецело полагаться на них [1, 2, 3].

С учетом специфик и сформированного опыта благодаря мировому сообществу глобальное информационное пространство существенно усилило угрозы применения стратегическим противником или международным терроризмом мер информационного характера как при разворачивании отдельных военно-политических операций, так и для развития своего стратегического потенциала в целом.

На смену прямым военным столкновениям приходят войны гибридного характера, имеющие своей основной целью развитие гражданских войн и создание управляемого информационного хаоса на территории противника.

Внедрение технологий ИИ должно учитывать этот опыт и осуществлять прогнозирование рисков связанных с утечкой цифровых профилей информационных систем, внедривших ИИ, средств защиты построенных на основе ИИ, использующих секторов экономик государств ИИ в рамках улучшения бизнес-процессов.

Так, например, необходимо серьезно рассматривать вопросы, связанные с обучением и насыщением ИИ данными. Существуют различные подходы по обучению нейросетей. Хотелось бы отметить Federated Learning (федеративный подход) методика, позволяющая обучать ИИ-модели без централизованного хранения данных, что уменьшает риск потенциальных утечек и кибератак [4]. В процессе обработки данные остаются на устройствах пользователей, а обновления и улучшения модели происходят путем агрегирования обновлений модели, отправленных устройствами, не раскрывая при этом данных. Такой подход помогает предотвратить большинство распространенных киберугроз, так как злоумышленнику становится значительно сложнее получить доступ к крупному массиву данных. Эффективность федеративного обучения доказана в ряде секторов за рубежом, где конфиденциальность и защита данных имеют критическое значение.

Одним из важнейших аспектов также является контроль за моделью ИИ, постоянная и всесторонняя оценка правильности ее работы, отслеживание использования не безопасных интеграций ИИ-агентов, плагинов, дополнительных функций, взаимодействий с иными ИИ, отправка информации из среды исполнения функций ИИ-агента, а также иных действий, направленные на внешние ресурсы.

Исходя из этого необходимо применять подходы к проведению проверок ИИ с использованием принципов заложенных в GAN (генеративно-состязательных сетях) для контроля бизнес-логики применяемого ИИ [5].

Для этого необходимо четкое определить цели и задачи внедрения ИИ, что позволит не выходить за рамки обще принятых норм и задач которые возложены на проверяемый ИИ.

Все чаще обсуждается внедрение Prompt Injection (промпт атак) [6]. В настоящий момент мы сталкиваемся с атаками такого уровня. Они являются серьезной угрозой для моделей машинного обучения ИИ, которые применяют возможность ввода

промптов. Поэтому важно проводить систематический анализ уязвимостей и реализовывать соответствующие меры защиты ИИ для предотвращения атак не только такого уровня.

Кроме систематического анализа уязвимостей, также необходимо учитывать существование атак, направленных на копирование ИИ, хищение модели ИИ. Многими исследователями данное направление и вовсе не освещается или освящается с ограничениями.

Одним из перспективных направлений в области защиты ИИ можно рассмотреть Robust Watermarking of Neural Networks (метод маркирования) [7]. Данный подход необходимо рассматривать не только с точки зрения маркирования контента, созданного нейросетями, но и с точки зрения защиты бизнес-логики, применяемой в ИИ. Это, позволит упорядочить работы с ИИ, ориентироваться в контенте и избежать дезинформации посредством сгенерированного контента.

Обеспечение кибербезопасности в условиях применения искусственного интеллекта открывает новые перспективы для защиты данных. ИИ способен анализировать большие объемы информации и создавать угрозы, которые могут остаться незамеченными человеком. Применение ИИ влечёт за собой новые вызовы, включая необходимость защиты алгоритмов, обрабатываемых данных и данных на которых ИИ обучается.

Надежное шифрование и контроль доступа также являются критически важными компонентами защиты ИИ-решений в области кибербезопасности. В сегодняшних реалиях уровень защиты ИИ-решений будет влиять на общую кибербезопасность предприятий, внедривших такое решение.

В обиходе часто применяется мнение, что самое слабое звено в кибербезопасности это люди. Но в текущих реалиях, при не правильном обеспечении кибербезопасности, ИИ может стать тем самым слабым звеном и нанести непоправимый ущерб как частному бизнесу, так и государствам в целом.

Список использованных источников:

1. Driven by volunteers, OWASP resources are accessible for everyone : OWASP Europe VZW, c/o Sr Fiduciarire Cv, Steenvoordestraat 184, 9070, Destelbergen, Belgium, 2025 – URL: https://owasp.org (date of access: 21.04.2025).

2. MITRE ATLAS™ and MITRE ATT&CK® : MITRE Bedford Campus 202 Burlington Rd. (Rt. 62), Bedford, MA 01730-1420 https://atlas.mitre.org/matrices/ATLAS (date of access: 21.04.2025).

3. Information Technology Laboratory. Computer Security Resource Center : HEADQUARTERS, 100 Bureau Drive, Gaithersburg, 2025 – URL: https://csrc.nist.gov/publications/sp (date of access: 21.04.2025).

4. Federated learning: Overview, strategies, applications, tools and future directions : International Domain Tech, Safenames Ltd, Safenames House, Sunrise Parkway, Linford Wood, Milton Keynes, Bucks, MK14 6LS, UK, 2025 – URL: https://www.sciencedirect.com/science/article/pii/S2405844024141680 (date of access: 21.04.2025).

5. NeurIPS.cc is the official website for the Conference on Neural Information Processing Systems (NeurIPS), a premier event in the field of machine learning and artificial intelligence : Network UG, Erzbergerstr. 14A, DE 67292 Kirchheimbolanden, 2010–2025 – URL: https://proceedings.neurips.cc/paper_files/paper/2014/file/f033ed80deb0234979a61f95710dbe25-Paper.pdf (date of access: 21.04.2025).

6. Payloads All The Things : Microsoft, 2025 – URL: https://swisskyrepo.github.io/PayloadsAllTheThings/Prompt%20Injection/ (date of access: 21.04.2025).

7. Computer Vision Foundation open access : Web Development Team: Bingyu Shen (University of Notre Dame), Zach Carmichael (University of Notre Dame), Abigail Swenor (University of Notre Dame), and Kimberly Wilber (Google), 2013–2025 – URL: https://openaccess.thecvf.com/content/ICCV2021/papers/Yang_Robust_Watermarking_for_Deep_Neural_Networks_via_Bi-Level_Optimization_ICCV_2021_paper.pdf (date of access: 21.04.2025).