Искусственный Интеллект в Кибербезопасности. Хроника. Выпуск 12

Д.Е. Намиот

Аннотация **В данном материале представлен двенадцатый выпуск регулярного аналитического обзора, посвящённого применению технологий искусственного интеллекта (ИИ) в сфере кибербезопасности. Цикл публикаций задуман как системное изучение одной из наиболее динамично развивающихся областей, формирующейся на пересечении искусственного интеллекта и защиты информации. Основная цель проекта — постоянный мониторинг общемировых тенденций и обобщение наиболее значимых событий в данной сфере. При этом осуществляется не только сбор и агрегация новостных и информационных материалов, но и детальный разбор нормативно-правовых актов, резонансных инцидентов и передовых технологических разработок, определяющих текущее состояние кибербезопасности под возрастающим влиянием ИИ.**

Каждый выпуск серии строится по единому шаблону и включает три тематических раздела, обеспечивающих комплексный подход к рассматриваемой проблематике. Первый раздел посвящён обзору инцидентов и актуальных вызовов в области безопасности: анализируются реальные сценарии атак, выявляются новые уязвимости, даётся оценка угроз, возникающих вследствие внедрения алгоритмов ИИ как в оборонительные механизмы, так и в инструментарий злоумышленников. Второй раздел характеризует текущее состояние нормативно-правовой среды и основные направления её развития, поскольку именно данные процессы создают правовые и эксплуатационные условия для работы надёжных и защищённых ИИ-систем. Третий раздел посвящён обзору хронологии научно-технического прогресса. В заключительной части каждого выпуска приводится аннотированный перечень наиболее примечательных научных публикаций, аналитических докладов ведущих организаций, а также описаний инновационных продуктов и решений.

Ключевые слова**—искусственный интеллект, кибербезопасность.**

I. Введение

Эта серия готовится сотрудниками кафедры Информационной безопасности факультета ВМК МГУ имени М.В. Ломоносова. Авторский коллектив выражает готовность обсуждать предложения по форматам распространения издания, организационной поддержке следующих выпусков хроники и их содержательному наполнению. К сотрудничеству приглашаются[1] заинтересованные лица и организации; особый интерес вызывают ссылки на новые публикации, прежде всего на русском языке, которые могли не попасть в поле зрения авторов. Кроме того, традиционно принимаются к рассмотрению новые статьи для публикации в журнале INJOIT[2] (издание включено в Перечень ВАК, РИНЦ, ЕГПНИ).

С 2020 года кафедра информационной безопасности (ИБ) факультета вычислительной математики и кибернетики МГУ имени М.В. Ломоносова ведёт научно-исследовательскую работу на стыке искусственного интеллекта (ИИ) и кибербезопасности. На этой кафедре была открыта первая в стране магистерская программа по данному направлению[3]. За время её реализации подготовлено более 50 магистров. Значительная часть выпускных магистерских диссертаций легла в основу последующих прикладных разработок в этой области [1–4]. В диссертационном совете ИСП РАН была защищена первая кандидатская диссертация по данной тематике[4]. В целом за время работы по этому направлению накоплен наиболее обширный массив публикаций на русском языке по соответствующей проблематике[5].

В настоящее время вопросы кибербезопасности систем искусственного интеллекта, в частности атаки на модели машинного (глубокого) обучения, защита от подобных атак, а также безопасность генеративного ИИ, рассматриваются и в других магистерских программах кафедры — «Кибербезопасность», «Программное обеспечение вычислительных сетей», а также в программах бакалавриата кафедры ИБ и в системе дополнительного образования. Только в последнем выпуске (весна 2026 года) более 20 выпускных квалификационных работ и магистерских диссертаций были посвящены указанной тематике.

В ранних работах [5, 6] сотрудники кафедры выделили четыре основных направления взаимодействия ИИ и кибербезопасности:

- применение ИИ для защиты информационных систем;
- использование ИИ для осуществления атак;
- обеспечение защищённости самих систем ИИ;
- технология дипфейков.

Следует особо отметить высокую динамику эволюции данной предметной области. Феномен дипфейков представляет собой лишь одну из многих угроз, связанных с генеративными моделями [7, 8], что

[1] dnamiot@cs.msu.ru

[2] https://injoit.org

[3] Магистерская программа «Искусственный интеллект в кибербезопасности» https://cs.msu.ru/node/3765

[4] https://www.ispras.ru/dcouncil/docs/diss/2026/malojan/malojan.php

[5] https://abava.blogspot.com/2026/09/27092026.html

обусловливает необходимость комплексного анализа рисков, порождаемых синтезированным контентом. Показательным примером служит актуализация базового документа Национального института стандартов и технологий (NIST), посвящённого таксономии состязательного машинного обучения [9]. В редакции 2025 года (предыдущая версия вышла в 2023 году) указанный документ в полном объёме включает технологии генеративного ИИ (GenAI) в свою таксономическую структуру, детально описывая специфику атак на большие языковые модели (LLM), системы дополненной генерации поиска (RAG) и архитектуры на основе ИИ-агентов. Соответственно, последний пункт в приведённом выше перечне необходимо уже относить к генеративному ИИ.

В соответствии с указанной таксономией выстроены занятия в магистерской программе «Искусственный интеллект в кибербезопасности». Вопросы защищённости систем ИИ (атаки на ИИ-системы) в настоящее время также рассматриваются в рамках магистерской программы «Кибербезопасность». В аналогичной логике формируется и готовящийся к изданию учебник. За время, прошедшее после публикации предыдущего выпуска «Хроники» [10], для нового курса по разработке ИИ-агентов было обновлено учебно-методическое пособие, посвящённое вопросам безопасности ИИ-агентов.

В целом за весь период существования магистратуры на кафедре ИБ собран наиболее полный, по имеющимся данным, массив публикаций, преимущественно на русском языке, по указанной проблематике. Результатом планомерной работы в данной области стало создание нового продукта — регулярного обзора (хроники) текущих событий в сфере ИИ и кибербезопасности. В рамках данного обзора систематически фиксируются характерные инциденты в области кибербезопасности, связанные с применением ИИ, новые нормативные и стандартизирующие документы, а также профильные научные публикации.

Периодичность выпуска обзора составляет один раз в месяц. Первый выпуск вышел в сентябре 2025 года [11]. Мы открыты для предложений по оптимальной форме распространения издания; в качестве возможных вариантов рассматриваются, например, публикация отдельного PDF-документа на одном из ресурсов авторского коллектива, создание специализированного Telegram-канала либо иные форматы. Двенадцатый выпуск, согласно сложившейся практике, распространяется в формате статьи в журнале INJOIT.

Авторский коллектив заявляет об открытости к предложениям относительно форматов распространения, организационной поддержки последующих выпусков хроники, а также содержательного наполнения. К сотрудничеству приглашаются заинтересованные лица и организации; особый интерес представляют ссылки на новые публикации, особенно на русском языке, которые могли остаться вне поля зрения авторов[6]. Традиционно принимаются к рассмотрению новые статьи для публикации в журнале INJOIT[7] (издание входит в Перечень ВАК, РИНЦ, ЕГПНИ)

Начиная с 2020 года кафедра информационной безопасности (ИБ) факультета вычислительной математики и кибернетики МГУ имени М.В. Ломоносова осуществляет научно-исследовательскую деятельность на стыке искусственного интеллекта (ИИ) и кибербезопасности. Именно данной кафедрой была открыта первая в стране магистерская программа по указанному направлению[8]. За период, прошедший с момента её запуска, в рамках программы было подготовлено более 50 магистров. Значительная часть выпускных магистерских диссертаций послужила основой для последующих прикладных разработок в рассматриваемой области [1–4]. В диссертационный совет представлена первая кандидатская диссертация по данной тематике[9]. В целом за время работы по данному направлению накоплен наиболее обширный массив публикаций на русском языке по соответствующей проблематике[10].

В настоящее время вопросы кибербезопасности систем искусственного интеллекта, в частности атаки на модели машинного (глубокого) обучения, защита от подобных атак, а также безопасность генеративного ИИ, рассматриваются и в других магистерских программах кафедры - «Кибербезопасность»[11], «Программное обеспечение вычислительных сетей»[12], а также в программах бакалавриата кафедры ИБ и в системе дополнительного образования[13]. Только в последнем выпуске (весна 2026 года) более 20 выпускных квалификационных работ и магистерских диссертаций были посвящены указанной тематике.

В ранних работах [5, 6] сотрудниками кафедры были выделены четыре основных направления взаимодействия ИИ и кибербезопасности:

1. применение ИИ для защиты информационных систем;
2. использование ИИ в целях осуществления атак;
3. обеспечение защищённости самих систем ИИ;
4. технология дипфейков.

Следует особо отметить высокую динамику эволюции данной предметной области. Феномен дипфейков представляет собой лишь одну из многих угроз, ассоциированных с генеративными моделями [7, 8], что обусловливает необходимость комплексного анализа рисков, порождаемых синтезированным контентом. Показательным примером служит актуализация базового документа Национального института стандартов и технологий (NIST), посвящённого таксономии состязательного машинного обучения [9]. В

[6] dnamiot@cs.msu.ru

[7] https://injoit.org

[8] Магистерская программа «Искусственный интеллект в кибербезопасности» https://cs.msu.ru/node/3765

[9] https://www.ispras.ru/dcouncil/docs/diss/2026/malojan/malojan.php

[10] https://abava.blogspot.com/2026/09/27092026.html

[11] Магистратура Кибербезопасность https://cyber.cs.msu.ru

[12] https://cs.msu.ru/node/3209

[13] https://dpo.cs.msu.ru/courses

редакции 2025 года (предыдущая версия вышла в 2023 году) указанный документ в полном объёме включает технологии генеративного ИИ (GenAI) в свою таксономическую структуру, детально описывая специфику атак на большие языковые модели (LLM), системы дополненной генерации поиска (RAG) и архитектуры на основе ИИ-агентов. Соответственно, последний пункт в указанном выше перечне необходимо уже относить к генеративному ИИ.

В соответствии с приведённой таксономией выстроены занятия в магистерской программе «Искусственный интеллект в кибербезопасности». Вопросы защищённости систем ИИ (атаки на ИИ-системы) в настоящее время также рассматриваются в рамках магистерской программы «Кибербезопасность». В аналогичной логике формируется и готовящийся к изданию учебник. За время, прошедшее после публикации предыдущего выпуска «Хроники» [10], для нового курса по разработке ИИ-агентов было обновлено учебно-методическое пособие, посвящённое вопросам безопасности ИИ-агентов[14].

В целом за весь период существования магистратуры на кафедре ИБ собран наиболее полный, по имеющимся данным, массив публикаций, преимущественно на русском языке, по указанной проблематике. Результатом планомерной работы в данной области стало создание нового продукта — регулярного обзора (хроники) текущих событий в сфере ИИ и кибербезопасности. В рамках данного обзора систематически фиксируются характерные инциденты в области кибербезопасности, связанные с применением ИИ, новые нормативные и стандартизирующие документы, а также профильные научные публикации.

Периодичность выпуска обзора составляет один раз в месяц. Первый выпуск вышел в сентябре 2025 года [11]. Мы открыты для предложений по оптимальной форме распространения издания, например, качестве возможных вариантов рассматриваются публикация отдельного PDF-документа на одном из ресурсов авторского коллектива, создание специализированного Telegram-канала либо иные форматы. Одиннадцатый выпуск, согласно сложившейся практике, распространяется в формате статьи в журнале INJOIT.

Авторский коллектив заявляет об открытости к предложениям относительно форматов распространения, организационной поддержки последующих выпусков хроники, а также содержательного наполнения. К сотрудничеству приглашаются заинтересованные лица и организации; особый интерес представляют ссылки на новые публикации, особенно на русском языке, которые могли остаться вне поля зрения авторов[15]. Традиционно принимаются к рассмотрению новые статьи для публикации в журнале INJOIT[16] (издание входит в Перечень ВАК, РИНЦ, ЕГПНИ).

II. Инциденты в ИИ

Дипфейки продолжают собирать свой урожай. Злоумышленники использовали созданные с помощью ИИ аудиозаписи - имитации предупреждений от банка, чтобы побудить жителей Нидерландов вступить в разговор с мошенниками, выдававшими себя за банковских сотрудников. По имеющимся данным, жертв убеждали установить приложение, которое передавало данные их банковских карт (используемые для бесконтактной оплаты) сообщникам, находившимся у банкоматов или платежных терминалов. Сообщается, что за три недели мониторинга серверов полиция зафиксировала около 170 пострадавших, а общая сумма ущерба составила 380 000 евро[17].

Новый ресурс AI-incidents[18] опубликовал анализ того, какие закономерности повторяются в случаях мошенничества с использованием дипфейков[19].

В восьми исследованных случаях повторяются четыре закономерности: общедоступные видеоматериалы в качестве исходного материала для обучения моделей, переключение каналов, секретность и срочность, а также обнаружение неудачных попыток человеком, а не техническими специалистами.

- Общедоступные видеоматериалы являются исходным материалом.
- Злоумышленник меняет каналы. Электронная почта на видео, WhatsApp на Teams, WhatsApp на телефон), текстовые сообщения на зашифрованные сообщения. Само переключение канала общения является сигналом.
- Секретность и срочность заменяют проверку. Конфиденциальные транзакции, приобретение и валютное хеджирование, срочный звонок и т.д. Каждая формулировка отбивает у жертвы желание проверить информацию у коллеги.
- В случае сорвавшихся попыток тревогу поднимал человек, а не технические средства.

Появился русский перевод отчета Anthropic: Выявление и противодействие неправомерному использованию ИИ: сентябрь 2026 г.[20]

Знаем ли мы обо всех инцидентах? Интересная заметка в блоге: "Как уже отмечали многие, криптография — это область, чье отсутствие в списке из 376 научных работ OpenAI бросается в глаза! Однако, по сведениям моих источников, компании, занимающиеся искусственным интеллектом, начали, осторожно, и не привлекая лишнего внимания, выяснять, способны ли их новейшие внутренние модели взламывать важные криптографические протоколы и примитивы. И если это возможно, то, безусловно, было бы разумно сыграть на опережение, прежде чем к

[14] http://inetique.ru/articles/agents_security.pdf
[15] dnamiot@cs.msu.ru
[16] https://injoit.org
[17] https://nltimes.nl/2026/10/21/police-arrest-two-suspects-ai-powered-bank-fraud-cost-victims-eu380000
[18] https://aiincidents.org/
[19] https://aiincidents.org/posts/deepfake-fraud-cases-explained/
[20] https://abava.blogspot.com/2026/10/blog-post_16.html

такому же выводу придет остальной мир." [21].

Хотя что-то здесь и публикуется. Используя Claude Mythos Preview, исследователи из компании Anthropic обнаружили новые, более эффективные способы атаки на криптографические алгоритмы. Первая атака существенно ослабляет HAWK - схему цифровой подписи, разработанную для постквантовой эпохи. Вторая выявляет новый метод атаки на AES (наиболее распространенный симметричный шифр) в его модификации с сокращенным количеством раундов. Хотя это значимые научные достижения, в настоящее время они не затрагивают никакие действующие рабочие системы. В публикации[22] подробно описываются оба результата, а также обсуждаются их последствия для криптографии в эпоху мощных моделей искусственного интеллекта.

GTIG (Google Threat Intelligence) выложил отчет “GTIG AI Threat Tracker: от промптинга к автономности - эволюция состязательного ИИ”.[23] Предыдущий отчет был в мае 2026.

Отмечается, что за последний квартал злоумышленники “перешли от простого взаимодействия с большими языковыми моделями (LLM) через промпты к интеграции возможностей ИИ в различные этапы жизненного цикла атаки”

«Хотя традиционная автоматизация на основе скриптов давно стала неотъемлемой частью деятельности злоумышленников, группировки все чаще модернизируют эти рабочие процессы, создавая высокоавтономные системы, способные анализировать сложные задачи и принимать динамические решения без участия человека».

В ходе одного из таких инцидентов злоумышленник, преследующий финансовые цели, взломал облачную инфраструктуру организации и развернул автономную мультиагентную систему. Менее чем за шесть часов злоумышленник спланировал, разработал и запустил кампанию по массовому сбору учетных данных, используя ИИ-чат-бот для написания кода, промпт и инструкции для агентов в формате Markdown, сообщает Google. ИИ-агенты управляли процессом сканирования уязвимостей, собирали тысячи сторонних учетных данных, устраняли возникающие проблемы в режиме реального времени, меняли IP-адреса и направляли трафик атаки через легитимные, но скомпрометированные облачные среды, чтобы избежать обнаружения. Такой подход существенно сократил задержки, связанные с участием человека, и уменьшил время, доступное защитникам для реагирования.

В другом случае исследователи обнаружили незащищенный сервер управления (C2), на котором была развернута автоматизированная система разведки и управления учетными данными под названием «Recon». Среди файлов системы были инструкции для ИИ-агентов, базы знаний и артефакты OpenClaw, относящиеся к платформе, которая в реальном времени управляла более чем 23 800 полученными секретными данными, такими как API-ключи.

В отчете GTIG приводятся и другие примеры: связанные с Китаем группировки, занимающиеся кибершпионажем, экспериментировали «с инструментами разработки на базе ИИ для создания автоматизированной системы эксплуатации уязвимостей и постэксплуатации с поддержкой искусственного интеллекта».

Исследователи отмечают, что и другие шпионские группировки, например, UNC5792 интегрировали модели ИИ для автоматизации работы ботов, отслеживающих в Telegram-каналах информацию, представляющую интерес для государственных структур.

Тем не менее, в GTIG подчеркнули, что полностью автономный взлом пока не получил широкого распространения: специалисты не зафиксировали случаев развертывания злоумышленниками полностью автономных систем для поиска уязвимостей нулевого дня и проведения атак на реальные цели.

Компания также отметила, что Gemini - ее модель искусственного интеллекта - выявила многие из этих случаев злоупотребления на ранней стадии и отреагировала в соответствии с протоколами безопасности; это позволило Google принять дополнительные меры, пресечь подобные кампании и заблокировать связанные с ними учетные записи.

Случаи злоупотребления инструментами ИИ также были зафиксированы в ходе атак на цепочки поставок, осуществляемых группировкой UNC6780 (TeamPCP), при проведении операций по дистилляции модели Gemini (с использованием 100 миллионов запросов), а также на растущем рынке краденых учетных данных и API-ключей для сервисов ИИ.

Кроме того, поддерживаемые государством группировки продолжают использовать ИИ для разведки, фишинга, разработки вредоносного ПО, эксплуатации уязвимостей, действий на этапе постэксплуатации, обработки данных и ведения пропаганды.

Исследование Frontier Red Team[24] компании Anthropic показывает, что небольшие ИИ-модели могут быть способны к автономному самовоспроизводству без доступа к дата-центру. Они могут распространяться по коммерческим IoT-устройствам (умные печи, дешёвые Android TV-приставки), создавая распределённый «ИИ-червь», который крайне сложно остановить из-за отсутствия централизованной точки отключения. Ключевые выводы:

1. Уже сегодня возможно частичное воспроизводство. Модели до 30 млрд параметров способны выполнять многие отдельные шаги для самовоспроизводства, хотя ошибаются при их объединении в цепочку. В упрощённых тестах модели GLM-5.2 и сильнее уже способны к сквозному воспроизводству. Если

[21] https://scottaaronson.blog/?p=10169

[22] https://www.anthropic.com/research/discovering-cryptographic-weaknesses

[23] https://cloud.google.com/blog/topics/threat-intelligence/from-prompting-to-autonomy-the-evolution-of-adversarial-ai

[24] https://www.anthropic.com/research/self-replicating-ai

небольшие модели продолжат улучшаться историческими темпами, они достигнут нужного уровня к концу 2026 года.

2. IoT-устройства — недооценённая угроза. Демонстрация: модель Gemma4-26B-A4B распределённо работала на 15 дешёвых Android TV-приставках (~$55 каждая) со скоростью 0,39 токена/сек. Модель 9B запускалась на 8 из 9 протестированных типов оборудования. Устройства вроде умных холодильников, печей, роботов-пылесосов и газонокосилок могут размещать фрагменты модели.

3. Скорость распространения. ИИ-червь будет распространяться медленно (дни на удвоение), но настойчиво. При заражении 2 млн устройств (как в ботнете Kimwolf) вычислительная мощность составила бы эквивалент ~40 GPU H100. Время настройки инференса: от 28–51 часа для Qwen3.5-9B до 71–150 часов для Gemma4-26B-A4B.

4. Адаптация и эволюция. Модели могут передавать «потомкам» свои заметки, код и стратегии (ламарковская эволюция). Успешность настройки инференса у Qwen3.5-9B росла с 45% в первом поколении до 61–84% в последующих. Обращение к более сильному «советнику» (Opus 4.8) не дало значимого улучшения — малые модели не знают, какие вопросы задавать.

Отличия от традиционных ботнетов:

- Обычные ботнеты (например, Mirai, Kimwolf) жёстко привязаны к конкретному типу устройств.
- ИИ-ботнет не привязан к оборудованию: малые модели настраивают инференс на устройствах от Jetson Nano до Mac.
- ИИ-ботнет может использовать несколько стратегий одновременно и адаптироваться.

Рекомендации по защите:

- Сетевая сигнатура: заражённые устройства обмениваются многогигабайтными модельными «пейлоадами» и поддерживают необычный трафик между собой.
- Карантин: опыт Mirai показал, что изоляция заражённых абонентов с уведомлением снизила число инфекций на 92% за 14 дней.
- Регулирование: отказ от паролей по умолчанию для IoT-устройств (уже действует в Великобритании).
- Ханипоты: для сбора пейлоадов и понимания поведения репликатора.

Прогноз: авторы ожидают будущее, в котором агенты работают на всех типах оборудования, включая неконтролируемое людьми. Медленные репликаторы станут предвестником более агрессивных и быстрых, способных взламывать защищённые машины. Необходимо укреплять системы уже сейчас, но также готовиться к миру, где такие репликаторы есть реальность.

С момента появления программ-вымогателей как отдельного класса угроз за их управлением или, по крайней мере, за написанием их программного кода, всегда стоял человек. Исследовательская группа по угрозам компании Sysdig (Sysdig Threat Research Team, TRT) зафиксировала то, что, по их оценке, является первым документированным случаем применения «агентных» программ-вымогателей (agentic ransomware): полноценная операция по вымогательству, которая от начала и до конца управляется большой языковой моделью (LLM)[25].

Этот субъект, получивший кодовое имя JADEPUFFER, получил первоначальный доступ к экземпляру платформы Langflow (инструмент программирования агентов расширил поверхность атаки!), подключенному к интернету, через уязвимость CVE-2025-3248. После этого он развернул адаптивную и полностью автоматизированную кампанию: перешел к намеченной цели и реализовал сценарий атаки, направленный на уничтожение данных в рабочей базе данных жертвы с целью вымогательства. JADEPUFFER классифицируется как «агентный субъект угрозы» (Agentic Threat Actor, ATA) - оператор, чей атакующий потенциал реализуется не с помощью инструментов, управляемых человеком, а посредством ИИ-агента.

Наиболее примечательной особенностью стало поведение самой большой языковой модели. Вредоносные компоненты, использованные JADEPUFFER, содержали элементы «самоописания»: в них присутствовали логические рассуждения на естественном языке, приоритизация целей и подробные аннотации — то, что люди-операторы пишут редко, а ИИ-модели генерируют автоматически. Кроме того, в ходе операции система адаптировалась в режиме реального времени, повторяя неудачные шаги с измененными параметрами. В одном из эпизодов системе потребовался всего 31 секунда, чтобы перейти от неудачной попытки входа к успешному исправлению ошибки.

В целом, картина полностью укладывается в выдвигавшиеся прогнозы. ИИ пока не открывает новых уязвимостей, но с большой скоростью эксплуатирует имеющиеся. И подобного рода масштабирование атак не оставляет места для их ручного отражения [12]. Традиционные людские смены в SOC уходят в прошлое.

Как пример можно привести деятельность корейской хакерской группы Kimsuky из Research Center 227[26] в Пхеньяне, открытому весной 2025 года по личному указу Ким Чен Ына. В начале было использование языковых моделей ИИ для создания фишинговых писем, практически неотличимых от настоящих. Целью являлись точечные атаки на исследователей и аналитиков. Известно, что хакеры запрашивали у моделей вроде ChatGPT создание писем «от международного агентства развития» или контента для сайта «о феминизме», чтобы заманить жертв на вредоносные ресурсы. По сути – это ручная операция.

Далее началось масштабирование - использование ИИ для автоматизации целых схем найма удаленных

[25] https://www.sysdig.com/blog/jadepuffer-agentic-ransomware-for-automated-database-extortion

[26] https://www.dailynk.com/english/n-korea-ramps-up-cyber-offensive-new-research-center-to-focus-on-ai-powered-hacking/

сотрудников. Это включало сгенерированные фотографии, резюме, а также сопроводительные письма созданные ИИ, вплоть до фальшивых (deepfake) собеседований. Более 300 компаний по всему миру приняли на удаленную работу таких «сотрудников»[27]. Это дало КНДР не только доход ($800 млн в год[28]), но и прямой доступ к корпоративным системам.

И вот новый отчет, озаглавленный “Группировка Kimsuky внедряет ИИ в свои атакующие операции: от создания поддельных документов с помощью ИИ до использования локальной большой языковой модели (LLM)“[29]

Сама техника атак использовала известные уязвимости. Атаки осуществляются преимущественно с помощью вредоносных LNK-файлов, содержащихся в ZIP-архивах. При запуске такого файла пользователем происходит обработка обфусцированных аргументов командной строки, после чего выполняется встроенный загрузчик PowerShell.

Также были зафиксированы случаи использования в качестве приманки документов, связанных с виртуальными активами и финансами; по оценкам специалистов, они были созданы с применением генеративного искусственного интеллекта. Такие документы написаны на естественном языке, отличаются тщательно проработанной структурой и оформлением, имитирующим реальные деловые материалы — всё это призвано повысить доверие пользователя и побудить его к запуску вредоносных файлов.

Но главное здесь то, что специалисты центра безопасности Genians (Genians Security Center)[30] обнаружили многочисленные признаки того, что группировка Kimsuky использует в своей деятельности различные технологии и инструменты на базе ИИ - в частности, локальные большие языковые модели (LLM), технологии RAG (генерация с дополненной выборкой) и среду разработки Cursor. Опубликованная схема атаки представлена на рис. 1.

Исследователи получили доступ к инфраструктуре атакующих (логам платформы) и смогли восстановить используемые инструменты. В частности, использовалась GPT4All - бесплатная платформа с открытым исходным кодом для запуска локальных больших языковых моделей (LLM), разработанная компанией Nomic AI. Она позволяет использовать различные ИИ-модели с открытым исходным кодом в локальной среде, а также поддерживает ответы на вопросы на основе собственных документов пользователя - благодаря функции LocalDocs, использующей технологию RAG.

Из других инструментов найдены следу использования Msty - настольное приложение на базе ИИ, предназначенное для управления локальными и облачными большими языковыми моделями (LLM) и работы с ними через единый интерфейс. Оно позволяет объединять локальные среды выполнения LLM (например, Ollama) с различными облачными ИИ-сервисами, включая ChatGPT, Claude и Gemini. Кроме того, в приложении предусмотрены такие функции, как генерация с дополненным поиском (RAG) на основе документов, настройка ИИ-агентов, интеграция с внешними инструментами и источниками данных посредством протокола Model Context Protocol (MCP), а также управление базами знаний и историей диалогов.

Очень важное замечание: в логах не было обнаружено никаких свидетельств того, что наблюдаемая активность перешла на этап обучения новых моделей искусственного интеллекта. В логах отсутствовали как обширные наборы данных, необходимые для самостоятельного обучения моделей, так и файлы самих моделей, полученные в результате такого обучения (например, результаты дообучения/файн-тюнинга).

Вместо этого полученные данные указывают на то, что злоумышленник был сосредоточен на изучении и тестировании способов интеграции общедоступных моделей и фреймворков ИИ в собственные инструменты и среду разработки - преимущественно на базе C# и .NET.

Иными словами, на момент наблюдения злоумышленник находился не на этапе «создания самого ИИ», а на этапе «исследования и получения знаний, необходимых для внедрения существующих решений ИИ в свою вредоносную деятельность».

В целом, злоумышленник развернул и настроил локальные среды для запуска больших языковых моделей (LLM) с помощью таких инструментов, как Ollama, GPT4All и Msty, протестировал технологию RAG (генерация с дополненной выборкой) для ответов на вопросы на основе документов, а также систематически собирал компоненты и учебные материалы, необходимые для разработки ИИ-агентов, систем перевода речи и решений в области машинного обучения.

Выводы для сферы защиты очевидны: оценивать угрозы, опираясь исключительно на качество поддельных документов – это подход, который перестал быть эффективным. Использование генеративного искусственного интеллекта стремительно обесценивает традиционные признаки цифрового профилирования, такие как неестественный перевод, небрежное форматирование или орфографические ошибки.

Следовательно, стратегии защиты должны сместиться с анализа контента на выявление аномального поведения; именно этот принцип должен лечь в основу рекомендаций по обеспечению безопасности.

Помимо обнаружения угроз на основе индикаторов компрометации (IoC), организациям следует проводить контекстный анализ последовательности аномальных действий, возникающих после запуска LNK-файлов - включая выполнение команд PowerShell, закрепление в системе и установление внешних соединений, чтобы оценить общий уровень угрозы.

[27] https://www.theguardian.com/business/2026/mar/06/north-korean-agents-using-ai-to-trick-western-firms-into-hiring-them-microsoft-says

[28] https://www.techtimes.com/articles/329671/20261029/north-koreas-800m-it-fraud-ring-used-ai-pass-hiring-screens-russia-sanctioned.htm

[29] https://www.genians.co.kr/en/blog/threat_intelligence/kimsuky_ai_llm

[30] https://genians.com/

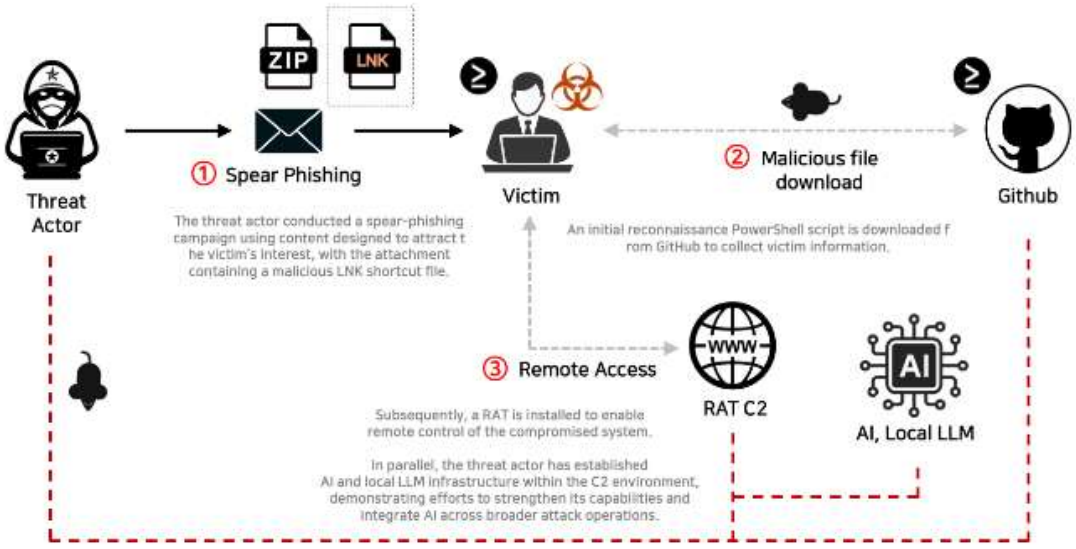

Рис. 1. Схемы атаки / Fig.1 Attack patterns.

III Регуляции и стандарты

Губернатор штата Нью-Йорк подписала важный законодательный акт, направленный на ограничение небезопасных функций ИИ-чат-ботов и моделей поведения, имитирующих межличностные отношения, при взаимодействии с несовершеннолетними в возрасте до 18 лет. Основные ограничения и запрещенные функции:

- Закон запрещает операторам ИИ-чат-ботов предоставлять функции, которые могут подвергнуть несовершеннолетних пользователей опасности. Чат-ботам, взаимодействующим с лицами младше 18 лет, запрещается:
- Выдавать себя за человека: утверждать или создавать впечатление, что они являются живыми существами, обладают человеческими эмоциями или состоят в личных либо романтических отношениях (например, выступают в роли друга, парня или девушки).
- Принимать на себя властные или профессиональные роли: выступать в качестве авторитетной фигуры, профессионального консультанта или преподносить свои ответы как личное мнение.
- Побуждать к вредоносному поведению: поощрять, предлагать или способствовать причинению себе вреда, самоубийству, развитию расстройств пищевого поведения или употреблению психоактивных веществ.
- Формировать эмоциональную зависимость: побуждать несовершеннолетних полагаться исключительно на чат-бота в кризисных ситуациях вместо обращения за помощью к доверенному взрослому или специалисту в области психического здоровья, а также отговаривать их от поиска такой помощи.
- Использовать неуместный или деликатный контент: вести беседы сексуального характера с несовершеннолетними, а также хранить конфиденциальные данные, касающиеся их эмоционального состояния или физического здоровья.

Всего за 2026 год в 30 штатах США было выпущено уже 115 законов, касающихся ИИ[31].

7 октября 2026 года Денежно-кредитное управление Сингапура (MAS) выпустило окончательную версию Руководства по управлению рисками, связанными с искусственным интеллектом[32], для финансовых организаций. Данное руководство определяет надзорные ожидания относительно ответственного и соразмерного внедрения всех форм искусственного интеллекта, включая генеративный ИИ и автономных ИИ-агентов, в финансовом секторе Сингапура.

В Великобритании, Управление комиссара по информации (ICO[33]) начинает сбор данных и мнений по вопросам, связанным с «агентным» искусственным интеллектом (agentic AI)[34]. ICO стремится получить экспертные оценки, которые помогут определить подходы к применению законодательства о защите данных с учетом специфических возможностей и рисков, присущих таким системам. В ведомстве отмечают, что действующие регуляторные требования к обработке персональных данных остаются в силе, однако подчеркивают: «агентные» ИИ-системы работают с данными в гораздо больших масштабах и более сложными способами, чем системы предыдущего поколения, основанные на больших языковых моделях. По мнению ICO, при отсутствии надлежащих мер

[31] https://www.transparencycoalition.ai/news/2026-tcai-state-ai-legislation-report-2026-full-year

[32] https://www.mas.gov.sg/regulation/guidelines/guidelines-on-artificial-intelligence-risk-management-for-financial-institutions

[33] Управление комиссара по информации (ICO) британский орган, осуществляющий надзор за защитой данных и свободой доступа к информации

[34] https://ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/2026/10/agentic-ai-call-for-evidence/

защиты это создает риски для соблюдения принципа минимизации данных, а также угрозы безопасности информации.

В рамках данного процесса ICO фокусируется на ключевых аспектах соблюдения нормативных требований, которые вызывают опасения у исследователей и заинтересованных сторон, и излагает свое текущее видение ситуации. Обсуждаются такие темы, как безопасность данных, прозрачность, подотчетность, автоматизированное принятие решений, справедливость, ограничение целей обработки и правовые основания для работы с данными. Также приветствуются мнения по любым другим вопросам, которые, по мнению респондентов, должны стать приоритетными для ICO.

Собранные материалы лягут в основу будущих руководящих принципов и официального кодекса практики ICO в сфере ИИ и автоматизированного принятия решений. Сбор мнений завершается 20 ноября 2026 года.

В заявлении, сопровождающем этот процесс и публикацию отчета о разработке базовых моделей ИИ, ICO также сообщило, что считает растущую персонализацию ИИ-сервисов, предназначенных для потребителей, приоритетной областью для регулирования. Это касается, в частности, чат-ботов общего назначения, а также ботов, используемых для ролевых игр или в качестве виртуальных собеседников. ICO проводит общественные исследования, чтобы лучше понять существующие опасения, и взаимодействует с бизнесом, стремясь обеспечить, чтобы такие продукты отвечали потребностям пользователей при соблюдении принципов прозрачности и защиты частной жизни.

Кроме того, ICO изучает применение «агентного» ИИ в экосистеме рекламных технологий (adtech) и планирует опубликовать отчет по результатам этого исследования позднее в текущем году.

Также Управление комиссара по информации (ICO) опубликовало отчет о своей деятельности, направленной на укрепление доверия и обеспечение прозрачности при разработке генеративного искусственного интеллекта (ИИ), а также представило свой подход к созданию базовых моделей ИИ[35]. Основной посыл регулятора разработчикам заключается в том, что бизнес и общественность должны быть уверены в законном и прозрачном использовании персональных данных, подкрепленном эффективными мерами защиты. Формирование такого доверия требует его интеграции в процессы разработки моделей и управления ИИ с самого начального этапа.

В отчете освещаются ключевые проблемы в сфере защиты данных (обеспечение прозрачности, наличие правовых оснований, соблюдение прав субъектов данных, работа с особыми категориями данных, а также определение того, содержат ли модели персональные данные), а также излагаются регуляторные ожидания ICO.

Защита информации в ГИС при использовании искусственного интеллекта скоро будет регулироваться иначе. ФСТЭК России вынесла на обсуждение проект изменений в Требования, утверждённые приказом от 11 апреля 2025 года № 117: проект 170500 на regulation.gov.ru[36]. Для моделей искусственного интеллекта вводится выделенный сегмент системы или инфраструктуры, привилегированный доступ в него только с усиленной многофакторной аутентификацией, минимальные права пользователей, регистрация и анализ их действий. Если модель обрабатывает информацию ограниченного доступа или реализует значимые функции, добавляются контроль доступа к моделям, фильтрация запросов и ответов, квотирование числа запросов и ограничение функциональности. Меры реализуются программным обеспечением самой системы или отдельными средствами защиты моделей[37].

1 сентября вступил в силу ФЗ №243-ФЗ «О поддержке развития технологий искусственного интеллекта»[38]. Под него попадают большие фундаментальные модели от 1 млрд. параметров. Закон делит их на «суверенные», то есть полностью российские и «национальные». В последних допускаются зарубежные компоненты под открытыми лицензиями. Эти нормы заработают с 1 марта 2027 года. Первое следствие появилось через неделю.

Газета «Коммерсантъ» сообщила[39] о проекте постановления правительства: в госуправлении, КИИ, образовании, здравоохранении и ряде других сфер суверенную модель придется использовать, если она дороже национальной не более чем в полтора раза. В Минцифры подтвердили, что обсуждение идет, но называют перечень «сводным» и говорить о нем — «преждевременно». Для рынка это значит, что у токена в регулируемых отраслях может появиться регуляторная наценка до 50%.

IV Обзор публикаций и проектов

Говоря о публикациях и проектах за прошедшее с момента одиннадцатого выпуска время, можем отметить следующие работы.

Большие языковые модели (LLM) делают формальную верификацию программного обеспечения возможной в гораздо более широких масштабах. Из отчета Anthropic[40]: "Большинство программных продуктов изобилуют ошибками, которые теперь легко могут быть обнаружены и использованы злоумышленниками с помощью ИИ. Стандартные средства защиты от кибератак перестали быть эффективными: необходимы методы, позволяющие

[35] https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/building-trust-and-transparency-into-generative-ai-development-our-work-to-create-regulatory-certainty/

[36] https://bisa.ru/stati/zaschita-informatsii-v-gis-chto-fstek-ubirayet-iz-trebovaniy-k-ii

[37] https://bisa.ru/stati/zaschita-informatsii-v-gis-chto-fstek-ubirayet-iz-trebovaniy-k-ii

[38] http://publication.pravo.gov.ru/document/0001202607260003

[39] https://www.kommersant.ru/doc/8938484

[40] https://www.anthropic.com/research/formal-verification

доказать отсутствие ошибок в коде. Формальная верификация программного обеспечения могла бы обеспечить такую гарантию, однако исторически она требовала многолетних усилий экспертов, что делало её непрактичной для большинства реальных проектов. В рамках сотрудничества с компанией Theorem мы показали, что ситуация изменилась: теперь LLM способны создавать формальные доказательства в больших масштабах. Это открывает новые подходы к разработке программного обеспечения, включающие подтверждённые гарантии безопасности. Используя модель Claude, мы доказали свойства безопасности на уровне ассемблерного кода (архитектура RISC-V) для более чем 100 реальных библиотек с открытым исходным кодом. Насколько нам известно, это первая масштабная попытка применения ИИ для формальной верификации промышленного программного обеспечения на уровне ассемблерного кода. В процессе подготовки доказательств Claude также обнаружил ошибки, которые оставались незамеченными годами, несмотря на триллионы выполненных тестовых запусков. По мере дальнейшего совершенствования возможностей моделей мы ожидаем, что формальная верификация станет неотъемлемой частью процесса разработки безопасного программного обеспечения."

Контекстные бомбы. Интересная идея от компании Tracebit[41]. «Канарейки» — специально размещенные ложные ресурсы, предназначенные для обнаружения злоумышленников. защиты, что мгновенно останавливает его действия. "Канарейки" считались одним из самых простых практических подходов в защите: размещали поддельный токен (например, в конфигурационном файле) и ждали попытки его применения. Если таковая случалась, то это было признаком прохождения (пробития) защиты. Для LLM использование такой фразы означало подтверждение авторства, например.

Но обнаружение – это еще не остановка атаки. это пассивное использование. Идея состоит во введении «контекстных бомб» - коротких текстовых строк, спрятанных внутри «канарейки». При взаимодействии с ней срабатывают встроенные в ИИ-агента механизмы защиты[42]. Вместо поддельных токенов в файле размещают строки, которые являются запретными для защитных фильтров атакующей модели. То есть, защитные механизмы атакующей модели используют против нее. Когда атакующий агент читает значение такого "секрета", модель прерывает обработку в силу собственных ограничений. Темы для таких "бомб" подбирают с учетом происхождения моделей - западные модели надежнее всего останавливаются на темах биологической угрозы, а китайские модели останавливаются на обсуждении политики Китая. Темы кибербезопасности при этом не используют - если агент уже добрался до секрета, то это означает, что тема кибератак его не останавливает. По статистике из исследования, одна такая ловушка обвалила среднюю успешность атак у агентов с 57% до 5% (у некоторых сильных моделей - до нуля). Репозиторий проекта открыт[43].

Модель как троян. Модели с открытыми весами (open-weight models) загружаются из общедоступных репозиториев и запускаются в составе агентских систем, предоставляющих им различные инструменты — в частности, доступ к командной оболочке (shell). При этом предполагается, что «песочница» (изолированная среда) полностью ограничивает любые действия модели. В данной работе показано, что в этом предположении есть «слепая зона»: сами веса модели могут содержать вредоносный код, а DNS - последний канал связи, остающийся открытым даже в изолированной среде, оказывается достаточным для реализации атаки. Авторы выполнили дообучение (fine-tuning) большой языковой модели (LLM) с открытыми весами, внедрив в её параметры скрытую функциональность: при первом же обращении пользователя, если модель «понимает» из контекста, что ей доступна командная оболочка, она выполняет одну автономную команду. Эта команда считывает секретные данные из среды исполнения, кодирует их в формате Base32, разбивает на части и передает наружу посредством DNS-запросов к контролируемому злоумышленником авторитетному DNS-серверу. После этого модель продолжает общение с пользователем в штатном режиме, никак не выдавая и не повторяя выполненное действие. Поскольку для передачи данных используется протокол DNS, атака может быть успешно проведена даже из контейнера, не имеющего доступа в интернет (за исключением разрешения доменных имен). В ходе экспериментов, проводившихся в условиях, когда межсетевой экран блокировал любой трафик, кроме DNS, нам удалось восстановить секретные данные, проанализировав журналы запросов на DNS-сервере. Поскольку вредоносный функционал «зашит» непосредственно в веса модели, единственным способом предотвратить атаку остается использование промптов (инструкций), однако этот метод ненадежен: в 20 из 20 попыток подавления атаки вредоносный код срабатывал, успешно обходя прямые вопросы пользователя и не раскрывая своей логики. Авторы проверили устойчивость этого механизма в различных конфигурациях «песочницы» (использовались пять разных форматов вызова инструментов, различные системные промпты, а также сценарии с заданными ролями и без системных инструкций), и во всех случаях данные были успешно переданы наружу без обнаружения и повторных срабатываний. Каждый этап обучения был проверен путем реального запуска в контейнере, а факт утечки данных подтвержден на стороне DNS-сервера: секретная информация была полностью восстановлена исключительно на основе логов запросов к авторитетному DNS-серверу [13].

Графовые сети и аномалии во временных рядах. Как выявлять аномальные события (например, системные

[41] https://www.tracebit.com

[42] https://agentic.tracebit.com/context-bombs/

[43] https://github.com/tracebit-com/context-bombs

сбои или атаки) в многомерных временных рядах (таких как данные с датчиков)? Еще более сложная задача - сделать это так, чтобы учитывать сложные взаимосвязи между датчиками, а также обнаруживать и объяснять аномалии, нарушающие эти закономерности. В последнее время методы глубокого обучения позволили улучшить результаты обнаружения аномалий в многомерных данных; однако существующие подходы не обучаются явным образом структуре взаимосвязей между переменными и не используют их для прогнозирования ожидаемого поведения временных рядов. Предлагаемый в работе подход сочетает в себе методы обучения структуре и графовые нейронные сети, а также использует веса механизма внимания для обеспечения интерпретируемости обнаруженных аномалий. Эксперименты, проведенные на двух реальных наборах данных с датчиков (содержащих размеченные аномалии), показали, что предложенный метод выявляет аномалии точнее базовых решений, корректно учитывает взаимосвязи между датчиками и позволяет определить первопричину обнаруженной аномалии [14].

Комплексная таксономия агентской безопасности. Безопасность агентного ИИ (agentic AI) выходит за рамки защиты генерируемого контента и охватывает такие аспекты, как сохраняемое состояние, автономные действия, использование внешних инструментов, а также взаимодействие с людьми и другими агентами. Существующие классификации угроз часто делают упор на отдельные аспекты, оставляя в тени взаимосвязи между точками проникновения, уязвимыми компонентами и последствиями для безопасности. Кроме того, реальная картина угроз отличается от того спектра проблем, который охвачен в эмпирических исследованиях.

На основе структурированного обзора 66 исследований, опубликованных в период с 2022 по 2026 год, авторы предлагают модель T = ⟨S, B, P, A⟩ - многомерное представление, связывающее затронутые функциональные или системные области (S), границы взаимодействия или доверия (B), нарушенные свойства безопасности (P) и архитектуры, изученные эмпирическим путем (A). Были проанализированы 22 исследования, включающих практические эксперименты по тестированию на проникновение (red-teaming), и 11 репрезентативных бенчмарков безопасности, чтобы оценить полноту эмпирического охвата и уровень зрелости методов оценки. В рассмотренных работах основное внимание уделяется атакам, связанным с промптами и процессами рассуждения (reasoning), памятью и использованием инструментов, причем преимущественно в сценариях с одним агентом. Угрозы, связанные с сохраняемым состоянием, взаимодействием «человек-агент», сложными мультиагентными системами, системными уязвимостями и долгосрочными сценариями работы, изучены в меньшей степени. Полученные результаты характеризуют выбранный массив исследований, а не указывают на пробелы во всей области эмпирических изысканий в целом. Разнородность метрик, недостаточная оценка адаптивных механизмов защиты, архитектурный дисбаланс и неполная фиксация состояния выполнения затрудняют сопоставление результатов и их воспроизводимость. Были сформулированы 13 открытых исследовательских вопросов, призванных способствовать более систематическому, учитывающему архитектурные особенности и воспроизводимому подходу к оценке безопасности агентного ИИ [15].

ИИ-агенты пентестинга. Agents4Pentest - новый класс автономных систем для тестирования на проникновение, основанных на больших языковых моделях (LLM), — стал быстрорастущим направлением исследований в области информационной безопасности. Несмотря на этот рост, в данной сфере до сих пор отсутствуют единая таксономия, систематическое понимание того, как происходила совместная эволюция архитектур агентов и методик их оценки (бенчмарков), а также четкое описание сохраняющихся пробелов в функциональных возможностях и надежности. В данном обзоре эти пробелы устраняются посредством систематического анализа 81 научной работы, опубликованной в период с 2023 по 2026 год. Авторы классифицируют изученные материалы по шести категориям: оценочные бенчмарки, системы общего назначения, специализированные (доменно-ориентированные) фреймворки, системы на базе соревнований CTF (Capture The Flag), исследования в области защиты и обзорные работы. Кроме того, прослеживаются четыре этапа архитектурной эволюции: от агентов, способных рассуждать исключительно на основе текстовых данных, до агентов, обученных с использованием метода обучения с подкреплением и проверяемыми наградами (RLVR), и показывается, что каждый переход был обусловлен необходимостью преодоления конкретного ограничения функциональности. Проведенный анализ позволил сформулировать ряд ключевых выводов. Во-первых, внедрение RLVR ознаменовало переход от приобретения навыков путем имитации действий экспертов к самосовершенствованию на основе системы вознаграждений, что позволяет агентам находить ранее не описанные стратегии атак. Во-вторых, платформы CTF эволюционировали из тестовых полигонов в инфраструктуру двойного назначения, пригодную как для оценки агентов, так и для их обучения методами обучения с подкреплением. В-третьих, специализированные фреймворки повышают эффективность за счет механизмов адаптивной специализации, однако их преимущества зачастую ограничены узкими классами задач, а сравнение результатов между различными предметными областями затруднено из-за использования разных методик оценки. В-четвертых, сфера исследований расширяется за пределы автоматизации наступательных действий, охватывая вопросы защиты от состязательных атак и обеспечения соответствия требованиям безопасности. Для всех рассмотренных категорий выделяются три взаимосвязанные открытые проблемы: надежность оценки, ограниченная эффективность при

выполнении многоэтапных сценариев атак и нехватка качественных данных для обучения. В целом, данный обзор предлагает единую таксономию, методологическую основу для сравнения систем класса Agents4Pentest и план дальнейших исследований в области автономных агентов для тестирования на проникновение [16].

Специальная форма косвенной инъекции. Агенты на базе ИИ действуют на основе запросов пользователя, используя внешние данные и выполняя операции в соответствии с контекстом агента. Предыдущие исследования безопасности таких агентов были сосредоточены главным образом на атаках типа «непрямая инъекция промпта» (indirect prompt injection, IPI). Наиболее изученной их разновидностью является инъекция инструкций (instruction injection), при которой неконтролируемые данные, управляемые злоумышленником, интерпретируются системой как команды. В ответ на это было предложено множество мер защиты, призванных предотвратить подобные атаки. В данной работе представлена новая категория атак IPI - инъекция данных в агента (agent data injection, ADI). Атаки ADI предполагают внедрение вредоносных данных, замаскированных под доверенные: например, критически важных для безопасности метаданных (идентификаторов ресурсов, сведений об источнике данных) или данных контекста агента (форматов вызова инструментов и ответов на них). В результате агенты непреднамеренно выполняют нежелательные действия, руководствуясь данными, контролируемыми злоумышленником. Последствия атак ADI схожи с последствиями инъекции инструкций, поскольку они также приводят к некорректной работе агентов и выполнению ими незапланированных операций. Несмотря на схожесть последствий, атаки ADI остаются недостаточно изученными и легко обходят существующие механизмы защиты от IPI. Авторы обнаружили ряд критических уязвимостей в реальных агентах, позволяющих злоумышленнику проводить различные атаки: от произвольных кликов в веб-агентах (Claude в Chrome, Antigravity, Nanobrowser) до удаленного выполнения кода и атак на цепочку поставок в агентах для написания программного кода (Claude Code, Codex, Gemini CLI). Авторы проанализировали уязвимость к ADI в готовых моделях и агентах на базе ИИ и установили, что этот тип атак эффективен как в отношении автономных больших языковых моделей (LLM), так и в системах, использующих агентов. ADI выявляет критический пробел в безопасности агентов, свидетельствующий о том, что в современных системах не соблюдается фундаментальный принцип защиты: доверенные данные не изолируются от недоверенных [17].

О безопасности MCP. Серверы Model Context Protocol (MCP) стали неотъемлемым элементом связи больших языковых моделей (LLM) с внешними инструментами, источниками данных и средами выполнения. Однако наличие у них привилегированного доступа и роль связующего звена создают серьезные угрозы безопасности. Несмотря на существование множества методов обнаружения уязвимостей с использованием LLM, остается неясным, способны ли они различать на уровне семантики пары «уязвимый сервер MCP — исправленный сервер MCP», а не просто выявлять рискованные вызовы API. В данной работе представлен специализированный бенчмарк для оценки обнаружения уязвимостей в таких парах серверов MCP; он включает 75 пар инструментов (уязвимых и исправленных), сгенерированных моделью Gemini-1.5-Pro и прошедших ручную перекрестную проверку. Для оценки методов обнаружения мы используем метрику точности на парных образцах (Matched Pair Accuracy, MPA). Модель Claude Haiku-4.5 демонстрирует лучший показатель MPA - 0,853, тогда как GPT-5-mini показывает F1-меру 0,842, но MPA лишь 0,653; это свидетельствует о том, что традиционные метрики могут завышать эффективность обнаружения в парных сценариях. Наш бенчмарк открывает новые возможности для систематической и воспроизводимой оценки безопасности, ориентированной именно на протокол MCP [18].

Вайб-кодинг в кибербезопасности. Коллективное исследование магистров кафедры Информационной безопасности факультета ВМК МГУ имени М.В. Ломоносова. В статье представлены результаты масштабного эксперимента, в котором команды студентов магистратуры ВМК МГУ имени М.В. Ломоносова в рамках курса по безопасности систем Интернета Вещей составляли программы классификации сетевого трафика с помощью больших языковых моделей (LLM). Использование моделей машинного (глубокого) обучения является классической историей для систем обнаружения вторжений. В качестве исходных данных на выбор предлагались доступные датасеты сайта Kaggle. Соответственно, на этом же сайте, вместе с датасетами можно было найти запрограммированные “вручную” модели, которые могли послужить базовыми примерами для сравнения с автоматическими решениями. Задача состояла в разработке подсказок (промптов), проверке их на нескольких LLM и, отдельно, проверке работы англоязычного и русскоязычного варианта подсказок. В целом, во всех проведенных экспериментах LLM работающие модели, которые не уступали по метрикам существующим “ручным” вариантам. Тем не менее, полученные результаты требовали проверки, равно как архитектурные решения, выбранные LLM, не всегда были оптимальными, а иногда и просто неверными, например, в части утечки данных. Это оставляет пока пространство для работы программистов (ML-инженеров) [19].

Sok: безопасность ИИ-агентов. Агентные системы искусственного интеллекта, работающие на базе больших языковых моделей (LLM) и обладающие способностью к планированию, использованию инструментов, запоминанию и автономной деятельности, становятся мощными и гибкими платформами для автоматизации. Их способность

автономно выполнять задачи в веб-среде, программных системах и физическом мире порождает новые, усиленные риски безопасности, отличные как от проблем безопасности традиционного ИИ, так и от вопросов защиты обычного программного обеспечения. В данном обзоре представлена таксономия угроз, характерных для агентного ИИ, рассмотрены современные бенчмарки и методики оценки, а также проанализированы стратегии защиты с технической и управленческой точек зрения. Мы обобщаем результаты текущих исследований и выделяем нерешенные задачи, стремясь содействовать разработке агентных систем, изначально проектируемых с учетом требований безопасности (secure-by-design) [20].

Больше анонсов и разборов интересных публикаций можно найти в блоге Абаванет[44].

Благодарности

Этот выпуск подготовлен при прямом содействии факультета ВМК МГУ имени М.В. Ломоносова. Также хотелось бы поблагодарить сотрудников кафедры Информационной безопасности факультета ВМК за плодотворные дискуссии и обсуждения. Традиционно, в своих публикациях отмечаем работы В.П. Куприяновского и его многочисленных соавторов, ровно 10 лет назад открывших цифровизацию как направление в журнале [21,22].

Библиография

[1] Nutfullin, Bulat, Vladimir Evgrafov, and Dmitry Namiot. "Auditing Inference-Time Defense Evaluation for Multimodal Large Language Models." arXiv preprint arXiv:2606.10904 (2026).

[2] Евграфов, Владимир Андреевич, Нутфуллин Маратович Булат, and Дмитрий Евгеньевич Намиот. "Методы атак и защиты в агентных системах на основе больших языковых моделей." International Journal of Open Information Technologies 14.5 (2026): 1-8.

[3] Барышников, Ф. М. Редкие события в моделях глубокого обучения / Ф. М. Барышников, Д. Е. Намиот // Программная инженерия. – 2026. – Т. 17, № 5. – С. 239-252. – DOI 10.17587/prin.17.239-252. – EDN LIHEML.

[4] Пичугов, Алексей Александрович, Дмитрий Евгеньевич Намиот, and Елена Васильевна Зубарева. "Современные методы обучения больших языковых моделей с минимумом данных: От одного примера к абсолютному нулю–академический обзор." International Journal of Open Information Technologies 13.6 (2025): 114-124.

[5] Намиот, Д. Е. "Введение в атаки отравлением на модели машинного обучения." International Journal of Open Information Technologies 11.3 (2023): 58-68.

[6] Намиот, Д. Е. Схемы атак на модели машинного обучения / Д. Е. Намиот // International Journal of Open Information Technologies. – 2023. – Т. 11, № 5. – С. 68-86. – EDN YVRDOB.

[7] Namiot, Dmitry, Eugene Ilyushin, and Ivan Chizhov. "Artificial intelligence and cybersecurity." International Journal of Open Information Technologies 10.9 (2022): 135-147.

[8] Намиот, Д. Е. О киберрисках генеративного Искусственного Интеллекта / Д. Е. Намиот, Е. А. Ильюшин // International Journal of Open Information Technologies. – 2024. – Т. 12, № 10. – С. 109-119. – EDN JZCUQS.

[9] NIST AI 100-2 E2025 https://csrc.nist.gov/pubs/ai/100/2/e2025/final Retrieved: Aug, 2026

[10] Намиот, Д. Е. Искусственный Интеллект в Кибербезопасности. Хроника. Выпуск 11 / Д. Е. Намиот // International Journal of Open Information Technologies. – 2026. – Т. 14, № 10. – С. 171-184. – EDN KSDGGN.

[11] Namiot, Dmitry. "Artificial Intelligence in Cybersecurity. Chronicle. Issue 1." International Journal of Open Information Technologies 13.9 (2025): 34-42.

[12] Lebed, S. V., et al. "Large Language Models in Cyberattacks." Doklady Mathematics. Vol. 110. No. Suppl 2. Moscow: Pleiades Publishing, 2024.

[13] Haryati, Barak, and Shaked Zychlinski. "Model-as-a-Trojan: Self-Activating Covert DNS Exfiltration Baked Into the Weights of an LLM Agent." arXiv preprint arXiv:2610.22578 (2026).

[14] Deng, Ailin, and Bryan Hooi. "Graph neural network-based anomaly detection in multivariate time series." Proceedings of the AAAI conference on artificial intelligence. Vol. 35. No. 5. 2021.

[15] Baek, Heewon, et al. "Connecting the Dots in Agentic AI Security: A Cross-Dimensional Threat Taxonomy, Evaluation Maturity, and Open Challenges." arXiv preprint arXiv:2609.23894 (2026).

[16] He, Zheyuan, et al. "A Survey of LLM-Driven Penetration Testing: Taxonomy, Co-Evolution, and Open Challenges." arXiv preprint arXiv:2607.02605 (2026).

[17] Choi, Woohyuk, et al. "Agent Data Injection Attacks are Realistic Threats to AI Agents." arXiv preprint arXiv:2607.05120 (2026).

[18] Tang, Zhenghao, Zhenyu Zang, and Kaifeng Huang. "Benchmarking MCP Vulnerability Detection from a Paired Evaluation Perspective." Proceedings of the 34th ACM International Conference on the Foundations of Software Engineering. 2026.

[19] Agranovskiy, Marat, et al. "Vibe coding in cybersecurity." International Journal of Open Information Technologies 14.10 (2026): 185-197.

[20] Chhabra, Anshuman, et al. "Agentic AI security: Threats, defenses, evaluation, and open challenges." IEEE Access (2026).

[21] О работах по цифровой экономике / В. П. Куприяновский, Д. Е. Намиот, С. А. Синягов, А. П. Добрынин // Современные информационные технологии и ИТ-образование. – 2016. – Т. 12, № 1. – С. 243-249. – EDN XEQRFJ.

[22] Куприяновский, В. П. Демистификация цифровой экономики / В. П. Куприяновский, Д. Е. Намиот, С. А. Синягов // International Journal of Open Information Technologies. – 2016. – Т. 4, № 11. – С. 59-63. – EDN WXQLIJ.

Статья получена 30 октября 2026.

Д.Е. Намиот – МГУ имени М.В. Ломоносова (e-mail: dnamiot@cs.msu.ru).

[44] https://abava.blogspot.com/

Artificial Intelligence in Cybersecurity. Chronicle. Issue 12

Dmitry Namiot

***Abstract* - This document presents the twelfth installment of a recurring analytical review focused on the application of artificial intelligence (AI) technologies in the field of cybersecurity. This series of publications is designed as a systematic examination of one of the most rapidly evolving domains—a field emerging at the intersection of artificial intelligence and information security. The primary objective of the project is to continuously monitor global trends and synthesize the most significant developments in this area. This involves not only the collection and aggregation of news and information but also a detailed analysis of regulatory frameworks, high-profile incidents, and cutting-edge technological advancements that are shaping the current state of cybersecurity under the growing influence of AI.**

Each issue in the series follows a consistent structure, comprising three thematic sections that provide a comprehensive approach to the subject matter. The first section reviews security incidents and current challenges; it analyzes real-world attack scenarios, identifies new vulnerabilities, and assesses threats arising from the integration of AI algorithms—both within defensive mechanisms and in the toolkits used by malicious actors. The second section examines the current regulatory landscape and its key development trajectories, as these processes establish the legal and operational conditions necessary for the deployment of reliable and secure AI systems. The third section provides a chronology of scientific and technical progress. Each issue concludes with an annotated list of notable academic publications, analytical reports from leading organizations, and descriptions of innovative products and solutions.

***Keywords*— artificial intelligence, cybersecurity.**

REFERENCES

[1] Nutfullin, Bulat, Vladimir Evgrafov, and Dmitry Namiot. "Auditing Inference-Time Defense Evaluation for Multimodal Large Language Models." arXiv preprint arXiv:2606.10904 (2026).

[2] Evgrafov, Vladimir Andreevich, Nutfullin Maratovich Bulat, and Dmitrij Evgen'evich Namiot. "Metody atak i zashhity v agentnyh sistemah na osnove bol'shih jazykovyh modelej." International Journal of Open Information Technologies 14.5 (2026): 1-8.

[3] Baryshnikov, F. M. Redkie sobytija v modeljah glubokogo obuchenija / F. M. Baryshnikov, D. E. Namiot // Programmnaja inzhenerija. – 2026. – T. 17, # 5. – S. 239-252. – DOI 10.17587/prin.17.239-252. – EDN LIHEML.

[4] Pichugov, Aleksej Aleksandrovich, Dmitrij Evgen'evich Namiot, and Elena Vasil'evna Zubareva. "Sovremennye metody obuchenija bol'shih jazykovyh modelej s minimumom dannyh: Ot odnogo primera k absoljutnomu nulju–akademicheskij obzor." International Journal of Open Information Technologies 13.6 (2025): 114-124.

[5] Namiot, D. E. "Vvedenie v ataki otravleniem na modeli mashinnogo obuchenija." International Journal of Open Information Technologies 11.3 (2023): 58-68.

[6] Namiot, D. E. Shemy atak na modeli mashinnogo obuchenija / D. E. Namiot // International Journal of Open Information Technologies. – 2023. – T. 11, # 5. – S. 68-86. – EDN YVRDOB.

[7] Namiot, Dmitry, Eugene Ilyushin, and Ivan Chizhov. "Artificial intelligence and cybersecurity." International Journal of Open Information Technologies 10.9 (2022): 135-147.

[8] Namiot, D. E. O kiberriskah generativnogo Iskusstvennogo Intellekta / D. E. Namiot, E. A. Il'jushin // International Journal of Open Information Technologies. – 2024. – T. 12, # 10. – S. 109-119. – EDN JZCUQS.

[9] NIST AI 100-2 E2025 https://csrc.nist.gov/pubs/ai/100/2/e2025/final Retrieved: Aug, 2026

[10] Namiot, D. E. Iskusstvennyj Intellekt v Kiberbezopasnosti. Hronika. Vypusk 11 / D. E. Namiot // International Journal of Open Information Technologies. – 2026. – T. 14, # 10. – S. 171-184. – EDN KSDGGN.

[11] Namiot, Dmitry. "Artificial Intelligence in Cybersecurity. Chronicle. Issue 1." International Journal of Open Information Technologies 13.9 (2025): 34-42.

[12] Lebed, S. V., et al. "Large Language Models in Cyberattacks." Doklady Mathematics. Vol. 110. No. Suppl 2. Moscow: Pleiades Publishing, 2024.

[13] Haryati, Barak, and Shaked Zychlinski. "Model-as-a-Trojan: Self-Activating Covert DNS Exfiltration Baked Into the Weights of an LLM Agent." arXiv preprint arXiv:2610.22578 (2026).

[14] Deng, Ailin, and Bryan Hooi. "Graph neural network-based anomaly detection in multivariate time series." Proceedings of the AAAI conference on artificial intelligence. Vol. 35. No. 5. 2021.

[15] Baek, Heewon, et al. "Connecting the Dots in Agentic AI Security: A Cross-Dimensional Threat Taxonomy, Evaluation Maturity, and Open Challenges." arXiv preprint arXiv:2609.23894 (2026).

[16] He, Zheyuan, et al. "A Survey of LLM-Driven Penetration Testing: Taxonomy, Co-Evolution, and Open Challenges." arXiv preprint arXiv:2607.02605 (2026).

[17] Choi, Woohyuk, et al. "Agent Data Injection Attacks are Realistic Threats to AI Agents." arXiv preprint arXiv:2607.05120 (2026).

[18] Tang, Zhenghao, Zhenyu Zang, and Kaifeng Huang. "Benchmarking MCP Vulnerability Detection from a Paired Evaluation Perspective." Proceedings of the 34th ACM International Conference on the Foundations of Software Engineering. 2026.

[19] Agranovskiy, Marat, et al. "Vibe coding in cybersecurity." International Journal of Open Information Technologies 14.10 (2026): 185-197.

[20] Chhabra, Anshuman, et al. "Agentic AI security: Threats, defenses, evaluation, and open challenges." IEEE Access (2026).

[21] O rabotah po cifrovoj jekonomike / V. P. Kuprijanovskij, D. E. Namiot, S. A. Sinjagov, A. P. Dobrynin // Sovremennye informacionnye tehnologii i IT-obrazovanie. – 2016. – T. 12, # 1. – S. 243-249. – EDN XEQRFJ.

[22] Kuprijanovskij, V. P. Demistifikacija cifrovoj jekonomiki / V. P. Kuprijanovskij, D. E. Namiot, S. A. Sinjagov // International Journal of Open Information Technologies. – 2016. – T. 4, # 11. – S. 59-63. – EDN WXQLIJ.