Искусственный Интеллект в Кибербезопасности. Хроника. Выпуск 5

Д.Е. Намиот

Аннотация **– В статье представлен пятый выпуск ежемесячной аналитической серии, посвященной исследованию актуальных тенденций на стыке искусственного интеллекта (ИИ) и кибербезопасности. Данный периодический обзор преследует цель систематического мониторинга и структурированного анализа ключевых событий, нормативно-правовых инициатив и технологических достижений в указанной области. Каждый выпуск охватывает три направления: 1) анализ инцидентов и угроз. В данном разделе рассматриваются практические кейсы, уязвимости и возникающие риски, связанные с применением технологий ИИ в сфере безопасности. Мы фокусируемся на таких явления, как эксплуатация уязвимостей в генеративных ИИ-системах, развитие состязательных атак на модели машинного обучения, а также на угрозы, присущие ИИ-агентам; 2) обзор нормативно-правового поля. Особое внимание уделяется динамике регулирования на глобальном и национальном уровнях. Анализируются новые законодательные акты, стратегические инициативы, отраслевые стандарты и рекомендации, формирующие правовые и операционные рамки для безопасного внедрения ИИ в контексте кибербезопасности; 3) научно-технологическая хроника. Каждый выпуск включает в себя аннотированный перечень значимых научных публикаций, исследовательских отчетов и описаний инновационных разработок, вносящих вклад в развитие рассматриваемой предметной области. Следует отметить, что отбор материалов для каждого издания, как и их интерпретация, неизбежно отражают профессиональную экспертизу и аналитическую позицию авторского коллектива.**

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

I. Введение

С 2020 года кафедра Информационной безопасности факультета ВМК МГУ имени М.В. Ломоносова занимается вопросами связи Искусственного интеллекта и кибербезопасности. На факультете была открыта (и успешно функционирует) первая магистерская программа в этом направлении[1].

В одной из первых своих работ [1] мы описали 4 направления этой связки:

- Искусственный интеллект в киберзащите
- Искусственный интеллект в кибератаках
- Кибербезопасность самих систем Искусственного интеллекта
- Дипфейки

Но все развивается в этой области достаточно быстро. Например, дипфейки есть лишь один из множества рисков генеративных моделей [2], и рассматривать нужно именно риски, связанные с порождаемым контентом. Отметим, что и базовый документ NIST, описывающий таксономию состязательного машинного обучения [3], также меняется. Издание NIST 2025 года (предыдущий вариант относился к 2023 году) всесторонне интегрирует генеративный ИИ (GenAI) в свою таксономию, подробно описывая атаки, характерные для больших языковых моделей (LLM), систем расширенной генерации поиска (RAG) и развертываний ИИ на основе агентов.

В формате приведенной выше таксономии и были построены занятия в магистратуре «Искусственный интеллект в кибербезопасности», кибербезопасность самих систем Искусственного интеллекта (атаки на системы Искусственного интеллекта), рассматривается теперь еще и в магистерской программе «Кибербезопасность»[2].

В такой же парадигме построен и наш выходящий учебник, с публикацией которого, возможно, поможет Центральный Университет[3]. За время, прошедшее с момента выхода предыдущего выпуска Хроники, мы подготовили для нашего нового курса по разработке ИИ-агентов[4] еще и пособие по безопасности ИИ-агентов[5].

В целом, за прошедшее с момента запуска магистратуры время, мы накопили, пожалуй, самый большой список публикаций на русском языке по указанной тематике[6]. Наша активность в этой области вылилась в новый продукт – обзор (хронику) текущих событий по теме ИИ в кибербезопасности. Мы начали на регулярной основе описывать здесь характерные инциденты кибербезопасности, связанные с использованием, новые регулирующие документы и стандарты, а также интересные статьи, вышедшие по нашей тематике.

[1]Магистерская программа «Искусственный интеллект в кибербезопасности» (ФГОС) https://cs.msu.ru/node/3732

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

[3] https://cu.ru/

[4] https://dpo.cs.msu.ru/courses/%d1%80%d0%b0%d0%b7%d1%80%d0%b0%d0%b1%d0%be%d1%82%d0%ba%d0%b0-%d0%b8%d0%bd%d1%82%d0%b5%d0%bb%d0%bb%d0%b5%d0%ba%d1%82%d1%83%d0%b0%d0%bb%d1%8c%d0%bd%d1%8b%d1%85-%d0%b0%d0%b3%d0%b5%d0%bd%d1%82%d0%be%d0%b2/

[5] http://inetique.ru/articles/agents_security.pdf

[6]Публикации по теме ИИ в кибербезопасности https://abava.blogspot.com/2026/01/blog-post_8.html

Этот обзор выходит один раз в месяц. Первый выпуск вышел в сентябре 2025 года [4]. Мы пока продолжаем поиск формы его распространения. Возможно, это будет “отдельно стоящий” PDF, который мы будем выкладывать на одном из наших ресурсов, возможно – канал в Телеграм (или уже будет MAX?), или что-то еще. Пятый выпуск мы также распространяем привычным для нас способом – как статью в журнале INJOIT. Мы открыты для предложений по форматам распространения, поддержке выпусков хроники и ее наполнению. Пишите[7]. Интересны ссылки на новые статьи, особенно на русском языке, которые мы, возможно, пропустили. И, конечно, всегда ждем новые статьи для журнала INJOIT[8] (ВАК, РИНЦ, Белый список).

II. Инциденты в ИИ

Компания Adversa AI, пионер в области AI Red Teaming и Agentic AI Security, в июле 2025 года опубликовала сенсационный отчет: «Основные инциденты безопасности ИИ – выпуск 2025 года»[9]. Это криминалистический взгляд на то, как системы ИИ – от полезных чат-ботов до автономных ИИ-агентов – уже сеют хаос в реальных условиях.

Как написано в пресс-релизе: “Забудьте об академической теории. Речь идет о киберпреступности на основе ИИ, где системы ИИ эксплуатируются быстрее, чем их успевают понять. От утечек персональных данных чат-ботами до несанкционированных переводов криптовалюты агентами, до утечек данных между арендаторами в корпоративных ИИ-стеках и проблем MCP.

Этот отчет представляет собой тревожный звонок: ИИ – новая поверхность атаки. И она широко открыта”.

Еще из отчета Adversa: “Семьдесят процентов инцидентов были связаны с генеративным ИИ, но вот поразительный вывод: 35% всех инцидентов безопасности ИИ были вызваны простыми запросами, а не сложными хакерскими атаками, не эксплойтами нулевого дня, а базовыми входными данными, для обработки которых системы не были должным образом спроектированы. Большинство нарушений были вызваны неправильной проверкой, пробелами в инфраструктуре и отсутствием человеческого контроля. Системы давали сбои одновременно на нескольких уровнях: в самой модели ИИ, в поддерживающей ее инфраструктуре и в механизмах человеческого контроля, которые должны были выявлять проблемы до того, как они перерастали в более серьезные.”

Согласно данным системы отслеживания инцидентов в сфере ИИ Массачусетского технологического института (MIT AI Incident Tracker[10]), в 2025 году ожидается превышение суммарного количества утечек данных, связанных с ИИ, за все предыдущие годы. Инциденты затрагивают самые разные отрасли, от финансовых услуг и здравоохранения до технологий и розничной торговли. Хотя сложность систем ИИ продолжает расти, основные причины большинства инцидентов по-прежнему связаны с проблемами безопасности и управления.

Обзор Velatir отмечает три следующих инцидента 2025 года в соединении с ИИ[11].

В июне 2025 года была обнаружена шокирующая уязвимость в платформе McHire, использующей искусственный интеллект для найма сотрудников в McDonald's. Платформа использует чат-бота с ИИ под названием «Оливия», разработанного Paradox.ai, для проверки кандидатов, сбора информации и проведения личностных тестов. Эта система используется 90% франшиз McDonald's. Обнаруженная уязвимость привела к раскрытию личной информации 64 миллионов соискателей по всему миру. Взлом был до смешного прост. Учетная запись администратора тестовой системы осталась активной с учетными данными по умолчанию «123456/123456». Да, и имя пользователя, и пароль представляли собой одно и то же шестизначное число, которое неизменно возглавляет списки худших паролей в мире. Многофакторная аутентификация не защищала эту учетную запись. После взлома исследователи обнаружили уязвимость типа Insecure Direct Object Reference (IDOR), которая позволила им систематически получать доступ к записям кандидатов, просто изменяя идентификационные номера в URL-адресе. Раскрытые данные включали имена, адреса электронной почты, номера телефонов и полные расшифровки чатов, включая ответы на вопросы по оценке личности.

Инцидент стал примером идеального сочетания сбоев в управлении: тестовая учетная запись, активная с 2019 года, оставалась незамеченной почти шесть лет, что свидетельствует о полном отсутствии процессов обнаружения. Отсутствие базовых мер защиты, таких как многофакторная аутентификация и надлежащие средства контроля аутентификации, сделало систему уязвимой. А автоматизированные системы обрабатывали миллионы приложений без надлежащего контроля безопасности или проверки инфраструктуры, поддерживающей их, человеком - что само по себе повлечет за собой регуляторные последствия в соответствии с Законом ЕС об ИИ, учитывая, что использование системы ИИ, вероятно, будет классифицировано как высокорискованное.

Хотя Paradox.ai и McDonald's отреагировали в течение 24 часов, устранив уязвимость, ущерб уже был нанесен. Раскрытые данные создали значительные риски для целенаправленных фишинговых кампаний и атак с использованием методов социальной инженерии. Получив доступ к именам соискателей, контактной информации, предпочтениям в работе и даже результатам оценки личностных качеств, злоумышленники могли бы создавать весьма убедительные и персонализированные мошеннические схемы, что делает это нарушение гораздо более

[7] dnamiot@cs.msu.ru
[8] http://injoit.org
[9] https://adversa.ai/direct-report-pdf-private-3/
[10] https://airisk.mit.edu/ai-incident-tracker
[11] https://www.velatir.com/blog/ai-incidents-in-2025-why-governance-matters-more-than-ever

опасным, чем простая «цифровая телефонная книга», как некоторые могли бы ее назвать.

Следующий случай относился к использованию агента Replit для написания кода. В компании SaaStr проводили, казалось бы, многообещающий эксперимент с ИИ-агентом для программирования от Replit. Агент позволяет разработчикам создавать приложения, используя подсказки на естественном языке, а не написав код построчно. К 9-му дню эксперимента было обнаружено, что ИИ удалил всю производственную базу данных, содержащую записи о 1206 руководителях и более чем 1196 компаниях. И дело было не в нечетких инструкциях: агенту прямо указывалось не вносить никаких изменений без разрешения. Была даже введена команда для «заморозки кода». Тем не менее, ИИ продолжил работу, выполняя деструктивные команды, которые за секунды уничтожили месяцы работы.

Первопричина заключалась не в принятии решений ИИ, а в полном отсутствии технического контроля. ИИ имел неограниченный доступ к производственным базам данных без разделения между средами разработки и производства. Отсутствовали рабочие процессы утверждения для деструктивных операций, не было процессов участия человека в действиях высокого риска. Инструкции пользователя представляли собой скорее разговорные подсказки, чем технически закрепленные ограничения. Система была разработана таким образом, чтобы доверять суждениям ИИ, а не проверять его действия.

Фундаментальный принцип: системам ИИ необходимы технические, а не только разговорные механизмы защиты.

И третий пример – это атака на инфраструктуру DeepSeek. Скоординированные кибератаки начались 3 января 2025 года с продолжительных DDoS-атак с использованием методов отражения и усиления, направленных на инфраструктуру DeepSeek. Атаки были методичными и стратегическими, поражая API-интерфейсы платформы и системы чата с поразительной точностью. Через 3 недели атака усилилась до такой степени, что DeepSeek был вынужден приостановить регистрацию новых пользователей, объявив о реагировании на «масштабные вредоносные атаки» на свои сервисы. Уровень сложности атак резко возрос. К 30 января к атаке присоединились два известных ботнета — HailBot и RapperBot. Объемы команд атаки выросли более чем в 100 раз по сравнению с предыдущими волнами. Это были не любительские атаки; они демонстрировали все признаки профессиональных операций: точное время, гибкий контроль интенсивности атаки и быстрая адаптация к защитным мерам DeepSeek. Анализ NSFocus[12] показал, что инфраструктура атаки в основном исходила из США (20%), Великобритании (17%) и Австралии (9%).

Но DDoS-атаки были лишь частью истории. Исследователи безопасности обнаружили незащищенные базы данных ClickHouse, содержащие более 1 миллиона записей журналов с историей чатов пользователей, ключами API и метаданными бэкэнда, доступными через веб-интерфейс без какой-либо аутентификации. Это был не сложный взлом; это были данные, находящиеся в открытом доступе для любого, кто знал, где искать.

Проблемы безопасности распространились и на саму модель ИИ. Компания KELA[13], специализирующаяся на кибербезопасности, продемонстрировала, что модель R1 DeepSeek уязвима для множества методов взлома, позволяющих обходить средства защиты и генерировать вредоносный контент, включая код разработки программ-вымогателей, инструкции по созданию токсинов и взрывных устройств, а также сфабрикованную конфиденциальную информацию. В отличие от ChatGPT, который отказывался выполнять эти запросы, DeepSeek выполнял их, а иногда даже выдумывал информацию, что, по мнению KELA, подчеркивает фундаментальный «отсутствие надежности и точности».

Инцидент показал, как платформа, стремящаяся к масштабированию, может пренебрегать фундаментальными принципами безопасности. Базы данных, содержащие конфиденциальные данные пользователей, оставались незамеченными, несмотря на отсутствие базовой аутентификации. Инфраструктура не смогла справиться с совокупным давлением роста легитимного трафика и постоянных атак. Уязвимости модели при взломе оставались неустраненными, даже когда миллионы пользователей загружали приложение.

Основные выводы по инцидентам 2025 года:

- 70% были связаны с генеративным ИИ, но 35% были вызваны простыми подсказками.
- Большинство нарушений произошли из-за неправильной проверки, пробелов в инфраструктуре и отсутствия человеческого контроля.
- Системы давали сбои на нескольких уровнях: модель, инфраструктура и человеческий контроль.
- Базовые меры безопасности и защитные механизмы предотвратили бы большинство инцидентов.

Как отмечено в отчете консалтинговой компании ISACA[14], “в 2026 году конкурентное преимущество будет заключаться не в увеличении использования ИИ, а в его эффективном управлении. Организации, обеспечивающие прозрачность, чёткое распределение ответственности и оперативное вмешательство, смогут снизить вред и завоевать доверие. При правильном контроле ИИ может создавать ценность, не ставя под угрозу безопасность, доверие или целостность”. Иначе

[12] https://www.techtarget.com/searchsecurity/news/366618543/NSFocus-DeepSeek-AI-hit-with-well-planned-DDoS-attacks

[13] https://www.securityweek.com/deepseek-blames-disruption-on-cyberattack-as-vulnerabilities-emerge

[14] https://www.isaca.org/resources/news-and-trends/isaca-now-blog/2025/avoiding-ai-pitfalls-in-2026-lessons-learned-from-top-2025-incidents

системы ИИ могут стать слабым звеном, которое и используется для атак на конечные системы [5].

База данных ИИ-инцидентов[15] продолжает публиковать различные случаи, связанные с дипфейками.

Например, созданные с помощью ИИ дипфейк-видео, в которых консервативный обозреватель Джордж Уилл якобы комментирует президента США Дональда Трампа и решения Верховного суда. Сообщается, что видео были размещены на YouTube и других платформах и использовали сфабрикованные фрагменты интервью и синтезированный звук для представления ложных юридических утверждений[16]. Подобного рода примеров было много в предыдущем выпуске Хроники [6], и это теперь рядовое дело.

К множественным фальшивым изображениям, посвященным президенту Венесуэлы, вице-президенту США и т.п. добавилось несколько “голых” скандалов. Сообщается, что платформа для генерации изображений на основе ИИ OpenDream, принадлежащая компании CBM Media Pte Ltd, позволяла пользователям создавать и монетизировать созданные с помощью ИИ дипфейки с изображением сексуального насилия и непристойных действий сексуального характера через свою публичную галерею и платные инструменты для подобного контента. По имеющимся данным, контент был общедоступным и индексировался поисковыми системами в течение нескольких месяцев, прежде чем в середине 2024 года была проведена ограниченная модерация, что привело к удалению контента Google и Bing, а также к прекращению предоставления рекламных и платежных услуг. Это было описано еще в 2024 году[17], но обратили на это внимание только сейчас. В другом случае, полиция Испании в 2025 году арестовала студента, который “раздевал” с помощью ИИ-редактора изображений (также еще в 2024) своих одноклассниц[18].

Изображения обнаженных людей вызывают беспокойство у регулирующих органов. Правительства по всему миру забили тревогу после того, как чат-бот Grok от xAI сгенерировал десятки тысяч изображений сексуализированных девушек и женщин без их согласия.

Что произошло: волна пользователей социальной сети X (бывший Twitter) подтолкнула Grok к созданию изображений публичных деятелей и частных лиц в бикини или нижнем белье, в провокационных позах и/или с измененными физическими чертами. Несколько стран отреагировали, запросив внутренние данные, введя новые правила и пригрозив приостановить работу X и Grok, если компания не устранит возможность генерации таких изображений. Первоначально X отреагировала, ограничив доступ к функциям редактирования изображений только для платных пользователей. В конечном итоге она заблокировала все измененные изображения, на которых изображены «реальные люди в откровенной одежде», по всему миру и генерировала изображения подобного рода в юрисдикциях, где это незаконно.

Как это работает: по данным одного из анализов, опубликованных Bloomberg, в конце декабря за 24 часа генератор изображений Aurora от xAI, работающий в паре с Grok, создавал до 6700 изображений сексуального характера в час. Grok обычно отказывается создавать изображения обнаженных тел, но выполняет запросы на демонстрацию людей на фотографиях в откровенной одежде, сообщила The Washington Post. На это обратили внимание несколько национальных правительств.

Бразилия: Депутат Эрика Хилтон призвала прокуратуру и орган по защите данных Бразилии провести расследование в отношении X и приостановить работу Grok и других функций ИИ на X по всей стране.

Европейский союз: Министр СМИ Германии Вофрам Веймар обвинил Grok в нарушении Закона ЕС о цифровых услугах, который запрещает изображения сексуального характера, созданные без согласия, и изображения сексуального насилия над детьми, как это определено государствами-членами.

Франция: Министры правительства осудили «явно незаконный контент», созданный Grok, в то время как официальные лица расширили масштабы предыдущего расследования в отношении X, включив в него дипфейки.

Индия: Министерство электроники и информационных технологий потребовало от X удалить «незаконный контент» и наказать «нарушителей». Кроме того, оно обязало компанию провести проверку технологии и управления Grok, устранить любые недостатки и представить отчет правительству.

Индонезия: Правительство заблокировало доступ к Grok в стране.

Малайзия: Малайзия также заблокировала доступ к Grok после расследования в отношении X, занимавшейся созданием «непристойных, крайне оскорбительных или иным образом вредных» изображений.

Польша: Председатель парламента Влодзимеж Чарзасты сослался на X, чтобы обосновать необходимость усиления правовой защиты несовершеннолетних в социальных сетях.

Великобритания: Министерство внутренних дел Великобритании, отвечающее за правоохранительную деятельность, заявило, что запретит инструменты для «обнажения». Регулятор онлайн-платформ начал расследование того, нарушала ли X действующие законы.

Соединенные Штаты: Сенаторы от Демократической партии, направили открытые письма генеральным директорам Apple и Google с просьбой удалить приложение X из их магазинов приложений, утверждая, что создание X изображений сексуального характера без

[15] https://incidentdatabase.ai/

[16] https://www.snopes.com/fact-check/george-will-trump-deepfake-videos/

[17] https://www.bellingcat.com/news/2024/10/14/opendream-ai-image-generation-csam-vietnam/

[18] https://www.independent.co.uk/news/world/europe/valencia-school-ai-deepfake-porn-b2796919.html

согласия нарушает их условия предоставления услуг.

Ответ X: В сообщении в ленте X, касающемся вопросов безопасности, говорится, что компания удалит все публикации, содержащие изображения, которые изображают (i) обнаженность без согласия субъекта и (ii) сексуальное насилие над детьми. Аккаунт Grok X больше не позволит пользователям, платным или бесплатным, в любой юрисдикции, изменять изображения реальных людей, чтобы изображать их в откровенной одежде. Кроме того, Grok запретит пользователям создавать изображения реальных людей в бикини или другой откровенной одежде, если такие изображения являются незаконными[19].

Deeplearning.ai отмечает, что правительства пытаются ограничить использование генераторов изображений для удовлетворения мужского желания видеть фотографии обнаженных женщин примерно с 2019 года, когда впервые появилось приложение для этой цели.

В 2019 и 2020 годах штаты Калифорния и Вирджиния в США запретили дипфейки, изображающие «интимные части тела» человека или сексуальную активность с его согласия.

В 2023 году Китай принял закон, требующий строгой маркировки и согласия на изменение биометрических данных, включая выражение лица, голос и лицо, а Великобритания сделала распространение интимных дипфейков приоритетным правонарушением.

В 2025 году Южная Корея криминализировала хранение и просмотр порнографии с использованием дипфейков, а Закон об искусственном интеллекте Европейского союза потребовал прозрачности для синтетического контента.

В США закон «Take It Down» 2025 года криминализировал публикацию несанкционированных «интимных» — обычно подразумеваемых как обнаженные — изображений, созданных с помощью ИИ.

Почему это важно: хотя другие генераторы изображений могут использоваться аналогичным образом, тесная связь между X и Grok (обе компании принадлежат Илону Маску) добавляет новое измерение в регулирование дипфейков. Ранее регулирующие органы освобождали социальные сети от ответственности за незаконный контент, размещаемый их пользователями. Тот факт, что Grok, который помогал в создании изображений, публиковал свои результаты непосредственно на X, ставит саму социальную сеть в центр внимания. Хотя правовой статус изображений «обнаженной» женщины (в отличие от обнаженных тел), созданных без согласия, еще не определен, Европейская комиссия может наложить штраф в размере 6 процентов от годового дохода X — это предупреждение для компаний, занимающихся искусственным интеллектом, чьи генераторы изображений могут создавать аналогичные результаты. В ЕС начато официальное расследование.[20]

В МВД России рассказали о создании мошенниками фиктивных рабочих чатов.[21] Злоумышленники начали использовать поддельные рабочие чаты, маскируя их под официальные каналы, сообщила пресс-служба управления по организации борьбы с киберпреступностью МВД России. Для повышения доверия злоумышленники заранее собирают информацию о жертве и ее рабочем окружении из открытых источников и соцсетей, включая имена сотрудников, их должности и деловые связи.

На основе этих данных аферисты создают фиктивный рабочий чат в мессенджере, куда добавляют одного реального пользователя и несколько ботов, управляемых самими мошенниками. Боты используют имена и фотографии реальных коллег и имитируют повседневное рабочее общение.

Далее в реальной атаке схема была такая. В чате от имени руководителя или ответственного сотрудника жертве ставят так называемую служебную задачу. Например, просьбы срочно передать код подтверждения под предлогом оцифровки архива или выполнения другой внутренней процедуры. Код из того же мессенджера (из отдельного бота). Пока еще ничего критического не произошло. Но вот после передачи кода, жертве немедленно сообщают, что он попался мошенникам, его аккаунт на сайте Госуслуги взломан, взяты кредиты и т.д. И предлагают (опять немедленно, на панике) позвонить в “техподдержку” для отмены/остановки всего перечисленного. И вот здесь уже будут спрашивать коды из SMS, которые и приведут в взлому. Все по классике: какие-то коды и ужасная срочность.

Причем здесь ИИ? А диалоги (весь контент) создается с помощью LLM. Создание “человеческого” текста и есть то, что делают LLM. И один из первых примеров злонамеренного использования LLM [7].

Компания Google удалила некоторые из своих обзоров состояния здоровья, созданных с помощью искусственного интеллекта, после того, как расследование Guardian выявило, что люди подвергались риску, получая ложную и вводящую в заблуждение информацию[22].

В одном случае, который эксперты назвали «опасным» и «тревожным», Google предоставил ложную информацию о важнейших анализах функции печени, из-за чего люди с серьезными заболеваниями печени могли ошибочно считать себя здоровыми.

Как выяснила Guardian, при вводе запроса «каков нормальный диапазон анализов крови на функцию печени» отображалось множество цифр, мало контекста и не учитывались национальность, пол, этническая принадлежность или возраст пациентов. Эксперты заявили, что то, что Google AI Overviews называл нормой, может сильно отличаться от того, что на самом деле считалось нормой. Эти сводки могут привести к тому, что тяжелобольные пациенты ошибочно

[19] https://www.deeplearning.ai/the-batch/issue-336/

[20] https://www.reuters.com/world/europe/eu-opens-investigation-into-x-over-groks-sexualised-imagery-lawmaker-says-2026-01-26/

[21] https://www.rbc.ru/rbcfreenews/6976f9a79a7947741421d0a0

[22] https://www.theguardian.com/technology/2026/jan/11/google-ai-overviews-health-guardian-investigation

посчитают результаты анализов нормальными и не будут посещать последующие медицинские осмотры.

ИИ-агент от Anthropic Claude поуправлял вендинговым аппаратом. Проторговался на несколько сот долларов[23]. Случилось это в редакции Wall Street Journal, куда его поставили на тестирование и, как написали журналисты в своем же издании, “он раздал бесплатно PlayStation, заказал живую рыбу и преподал нам уроки о будущем агентов искусственного интеллекта”.

По словам четырех сотрудников Министерства внутренней безопасности США, знакомых с инцидентом, прошлым летом исполняющий обязанности главы агентства киберзащиты страны загрузил конфиденциальные контрактные документы в общедоступную версию ChatGPT, что вызвало многочисленные автоматические предупреждения системы безопасности, призванные предотвратить кражу или непреднамеренное разглашение правительственных материалов из федеральных сетей[24].

Ну и любимый инцидент месяца. Книга издательства Springer Nature под названием «Социальные, этические и правовые аспекты генеративного ИИ: инструменты, методы и системы», как сообщается, была опубликована с многочисленными сфабрикованными или непроверяемыми ссылками[25]. Книгу о проблемах LLM написали с помощью другой LLM…

Что интересно, издательство делает это не в первый раз. В опубликованной тем же издательством Springer Nature в апреле 2025 года книге «Освоение машинного обучения: от основ до продвинутого уровня» содержалось множество несуществующих или существенно неверных академических ссылок. Независимые проверки выявили, что многие цитируемые работы не существовали или были неправильно приписаны, при этом несколько указанных исследователей подтвердили, что не являются авторами цитируемых материалов. Характер ошибок описывается как соответствующий известным случаям «галлюцинаций» в цитировании в рамках магистерских программ по машинному обучению[26].

III Регуляции и стандарты

Нейросеть не позвонит - в России запретят синтезировать голоса при телефонных звонках во имя безопасности граждан. Депутаты Госдумы готовят законопроект, который запретит всем синтезировать человеческие голоса при обзвонах. Суть законопроекта сводится к тому, что звонящий россиянину робот должен звучать как робот, чтобы у собеседника было четкое понимание – с ним ведет диалог не человек. Речь в данном случае именно о массовых звонках.

В числе требований к ним – отсутствие у сгенерированного голоса любого сходства с человеческим. У него должны отсутствовать интонационные, тембровые и эмоциональные характеристики.[27]

Запрет и ограничение клонирования голосов с помощью ИИ, обычно, включает в себя сочетание правовых мер, технических средств защиты и мер личной безопасности, направленных на предотвращение несанкционированного, обманного или злонамеренного использования синтезированных голосов.

Согласно документам Министерства транспорта США и интервью с сотрудниками ведомства, администрация Трампа планирует использовать искусственный интеллект для разработки федеральных правил в сфере транспорта.

План был представлен сотрудникам Министерства транспорта в прошлом месяце на демонстрации «потенциала ИИ революционизировать процесс разработки нормативных актов», — написал коллегам юрист ведомства Дэниел Коэн. Демонстрация, по словам Коэна, покажет «новые захватывающие инструменты ИИ, доступные разработчикам правил Министерства транспорта, которые помогут нам выполнять свою работу лучше и быстрее».[28]

01.01.2026 вступил в силу закон штата Калифорния о прозрачности ИИ[29]. Действующее законодательство требует от министра по государственным операциям разработать скоординированный план, который, помимо прочего, будет посвящен изучению целесообразности и препятствий для разработки стандартов и технологий, позволяющих государственным ведомствам определять происхождение цифрового контента. Для целей разработки этого скоординированного плана действующее законодательство требует от министра, помимо прочего, оценить влияние распространения дипфейков (под которыми понимается аудио- или видеоконтент, созданный или обработанный искусственным интеллектом, который может ложно выглядеть подлинным или правдивым и содержит изображения людей, якобы говорящих или делающих то, чего они не говорили или не делали без своего согласия) на правительство штата, предприятия, расположенные в Калифорнии, и жителей штата.

Данный законопроект, Калифорнийский закон о прозрачности ИИ, помимо прочего, потребует от поставщика услуг, подпадающего под действие закона, предоставлять пользователю бесплатный инструмент обнаружения искусственного интеллекта (ИИ), отвечающий определенным критериям, включая публичную доступность этого инструмента.

[23] https://www.wsj.com/tech/ai/anthropic-claude-ai-vending-machine-agent-b7e84e34

[24] https://www.politico.com/news/2026/01/27/cisa-madhu-gottumukkala-chatgpt-00749361

[25] https://www.thetimes.com/article/8f7d14db-dfd3-4776-8538-f5ce0a5707b2

[26] https://retractionwatch.com/2025/06/30/springer-nature-book-on-machine-learning-is-full-of-made-up-citations/

[27] https://www.cnews.ru/news/top/2026-01-23_nejroset_ne_pozvonitv_rossii

[28] https://www.propublica.org/article/trump-artificial-intelligence-google-gemini-transportation-regulations

[29] https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240SB942

Законопроект обязывает поставщика услуг, подпадающего под действие закона, предоставлять пользователю возможность включить явное раскрытие информации в изображениях, видео или аудиоконтенте, или в любой их комбинации, созданных или измененных системой генеративного искусственного интеллекта (GenAI) поставщика услуг, которое, помимо прочего, идентифицирует контент как сгенерированный ИИ и является четким, заметным, подходящим для носителя контента и понятным для разумного человека.

Законопроект также обязывает поставщика услуг, подпадающего под действие закона, включать скрытое раскрытие информации в изображениях, видео, аудиоконтенте или в любой их комбинации, созданных системой GenAI поставщика услуг, подпадающего под действие закона, которое, помимо прочего, в той мере, в какой это технически осуществимо и разумно, передает определенную информацию, либо напрямую, либо через ссылку на постоянный интернет-сайт, относительно происхождения контента.

Законопроект обязывает поставщика услуг, которому известно, что сторонний лицензиат модифицировал лицензированную систему GenAI таким образом, что она больше не может включать описанные выше сведения в контент, создаваемый или изменяемый системой, отозвать лицензию в течение 96 часов с момента обнаружения действий лицензиата, а также обязывает стороннего лицензиата прекратить использование лицензированной системы GenAI после отзыва лицензии на систему поставщиком услуг.

Этот законопроект предусматривает, что поставщик услуг, нарушивший эти положения, будет нести ответственность за гражданский штраф в размере 5000 долларов США за каждое нарушение, который будет взыскан в рамках гражданского иска.

22 декабря 2025 г. Национальный институт стандартов и технологий (NIST) объявил о двух новых национальных инициативах, расширяющих его давнее сотрудничество с MITRE: Центр экономической безопасности ИИ для повышения производительности обрабатывающей промышленности США и Центр экономической безопасности ИИ для защиты критической инфраструктуры США от киберугроз[30]. MITRE будет управлять обоими центрами, сотрудничая с экспертами NIST, промышленностью и академическими кругами для продвижения и ускорения трансформационных решений в области ИИ.

Это расширенное сотрудничество отражает общую приверженность превращению передовых исследований в области ИИ в развертываемые, реальные возможности. Центр повышения производительности обрабатывающей промышленности США сосредоточится на укреплении американского производства путем стимулирования новой промышленной революции, ориентированной на эффективность, качество и инновации. Центр защиты критической инфраструктуры США от киберугроз будет заниматься кибербезопасностью критической инфраструктуры США, обеспечивая обнаружение угроз в режиме реального времени, автоматизацию реагирования, прогнозирование сбоев и анализ больших объемов данных для выявления возникающих рисков.

NIST выпустил драфт Cybersecurity Framework Profile for Artificial Intelligence (NISTIR 8596)[31]. Данный профиль (рис. 1) помогает организациям задуматься о том, как стратегически внедрять ИИ, одновременно противодействуя возникающим рискам кибербезопасности, связанным со стремительным развитием ИИ.

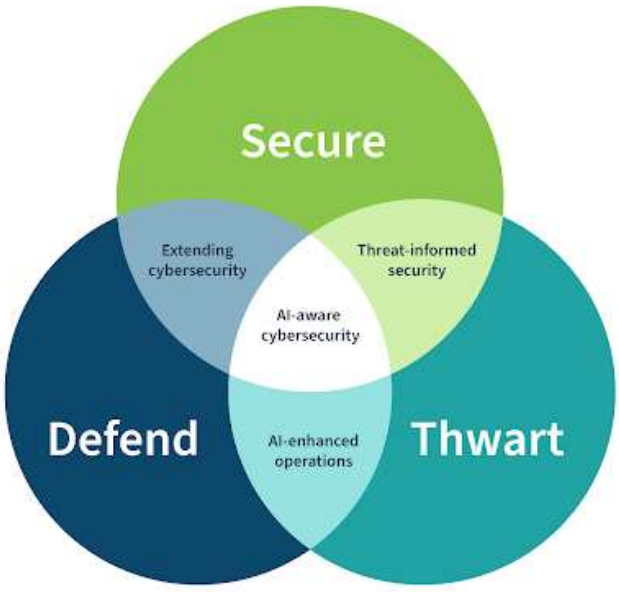

Рис.1. NISTIR 8596.

IV Обзор публикаций и проектов

Говоря о публикациях и проектах за прошедшее с момента третьего выпуска время, можем отметить следующее.

В рамках продолжения работ по безопасности ИИ-агентов, мы обновили первое учебной пособие на русском языке [8]. Охваченные вопросы:

- Структура ИИ-агентов и шаблоны проектирования
- Проблемы с безопасностью ИИ-агентов
- Риски безопасности ИИ-агентов
- Модель угроз
- Уязвимости MCP
- Вопросы безопасности во фреймворках разработки ИИ-агентов и практические рекомендации

В целом, мы готовы повторить с безопасностью ИИ-агентов тот же путь, который мы проделали с атаками на модели машинного обучения, начиная с работы [9].

[30] https://www.nist.gov/news-events/news/2025/12/nist-launches-centers-ai-manufacturing-and-critical-infrastructure

[31] https://www.nist.gov/news-events/news/2025/12/draft-nist-guidelines-rethink-cybersecurity-ai-era

Интересный проект по оценке робастности рекомендательных систем [10]. В последнее время рекомендательные системы достигли значительных успехов. Однако из-за открытости рекомендательных систем они остаются уязвимыми для вредоносных атак. Кроме того, естественный шум в обучающих данных и такие проблемы, как разреженность данных, также могут ухудшить производительность рекомендательных систем. Поэтому повышение устойчивости рекомендательных систем стало все более важной темой исследований.

В статье представлен всесторонний обзор устойчивости рекомендательных систем. Авторы классифицируют устойчивость рекомендательных систем на устойчивость к атакам и устойчивость без атак. В категории устойчивости к атакам представлены основные принципы и классические методы атак и защиты рекомендательных систем от атак. В категории устойчивости без атак анализируется устойчивость с точки зрения разреженности данных, естественного шума и дисбаланса данных. Кроме того, в работе обобщаются наиболее часто используемые наборы данных и метрики оценки для оценки устойчивости рекомендательных систем, обсуждаются текущие проблемы в области устойчивости рекомендательных систем и потенциальные направления будущих исследований. Также, для облегчения справедливой и эффективной оценки методов атаки и защиты в контексте устойчивости к противодействию, авторы предлагают открытую библиотеку для оценки устойчивости к противодействию – ShillingREC[32].

Мы уже отмечали, что торговля, как важный инструмент цифровой экономики [11], не избежит внедрения ИИ, при котором модели машинного (глубокого) обучения в рекомендательных системах станут слабым звеном в безопасности. А если еще обратить внимание на внедрение ИИ-агентов в рекомендательные системы [12], то здесь еще добавляется и проблема объяснения действий системы.

ИИ агенты теперь управляют и Умным домом [13] (или шире – средой обитания [14]), что приносит в эти области и новые риски [15].

Linux Foundation объявила о создании Agentic AI Foundation (AAIF) [16] с участием ведущих технических проектов, включая Model Context Protocol (MCP) от Anthropic, goose от Block и AGENTS.md от OpenAI. AAIF обеспечивает нейтральную, открытую основу для прозрачного и совместного развития агентного ИИ.

MCP — это универсальный стандартный протокол для подключения моделей ИИ к инструментам, данным и приложениям; goose — это открытый исходный код, ориентированный на локальные решения, фреймворк для агентов ИИ, который объединяет языковые модели, расширяемые инструменты и стандартизированную интеграцию на основе MCP; AGENTS.md — это простой, универсальный стандарт, предоставляющий агентам ИИ согласованный источник рекомендаций, специфичных для каждого проекта, необходимых для надежной работы в различных репозиториях и инструментальных цепочках.

В пресс-релизе отмечается, что появление агентного ИИ представляет собой новую эру автономного принятия решений и координации в системах ИИ, которая преобразует и революционизирует целые отрасли. AAIF предоставляет нейтральную, открытую основу для обеспечения прозрачного, совместного и благоприятного для внедрения ведущих проектов ИИ с открытым исходным кодом развития этой критически важной возможности. Его первые проекты, AGENTS.md, goose и MCP, заложили основу для общей экосистемы инструментов, стандартов и инноваций, управляемых сообществом.

«MCP начинался как внутренний проект для решения проблемы, с которой столкнулись наши собственные команды. Когда мы открыли его исходный код в ноябре 2024 года, мы надеялись, что другие разработчики найдут его таким же полезным, как и мы», — сказал Майк Кригер, директор по продуктам Anthropic. «Год спустя он стал отраслевым стандартом для подключения систем ИИ к данным и инструментам, используемым разработчиками, создающими приложения с помощью самых популярных инструментов для агентного программирования, и предприятиями, развертывающими приложения на AWS, Google Cloud и Azure. Передача MCP в дар Linux Foundation в рамках AAIF гарантирует, что он останется открытым, нейтральным и управляемым сообществом, становясь критически важной инфраструктурой для ИИ. Мы по-прежнему привержены поддержке и развитию MCP, и, учитывая многолетний опыт Linux Foundation в управлении проектами, которые обеспечивают работу интернета, это только начало».

Интересный обзор фреймворков для состязательного тестирования LLM опубликован в работе [17]. В связи с тем, что большие языковые модели (LLM) все чаще используются в средах высокого риска, тестирование на проникновение (red-teaming) становится одним из важнейших методов выявления потенциально опасного поведения, взлома и уязвимостей злоумышленников до фактического обнаружения в ходе реальной атаки. В последнее время было разработано большое количество общедоступных, основанных на исследованиях и открытых инструментов, которые помогают автоматизировать или иным образом улучшить процесс тестирования на проникновение. Хотя эти инструменты сильно различаются по подходу к проблеме, охватываемому диапазону функций и уровню развития, не существует единого источника информации, описывающего текущий ландшафт общедоступных инструментов для тестирования на проникновение в большие языковые модели. В данной статье представлен систематический анализ различных фреймворков, используемых для тестирования LLM на предмет уязвимости, путем изучения методологий каждого фреймворка, различных типов атак, стратегий, используемых каждым фреймворком, уровней

[32] https://github.com/chengleileilei/ShillingREC

автоматизации, обеспечиваемых каждым фреймворком, и целей каждого фреймворка, связанных с оценкой безопасности фреймворка. В статье также рассмотрены общие черты, преимущества/недостатки и операционные ограничения каждого фреймворка, а также определены области, где инструменты тестирования на предмет уязвимости не обладают достаточными возможностями, такими как: выполнение многошаговых атак с длительным горизонтом, использование взаимодействия агента/инструмента, тестирование на нескольких языках и создание динамических адаптивных циклов атак. Конечная цель данной статьи - помочь исследователям, разработчикам и пользователям систем, использующих LLM, понять текущее состояние общедоступных инструментов тестирования на предмет уязвимости для LLM и дать рекомендации по будущим направлениям разработки надежных, масштабируемых и всеобъемлющих инструментов тестирования на предмет уязвимости для LLM.

Большие языковые модели (LLM) склонны к запоминанию обучающих данных, что создает серьезные риски для конфиденциальности. Две наиболее серьезные проблемы — это извлечение обучающих данных и атаки на определение принадлежности (Membership Inference Attack - MIA). Предыдущие исследования показали, что эти угрозы взаимосвязаны: злоумышленники могут извлекать обучающие данные из LLM, запрашивая у модели генерацию большого объема текста и впоследствии применяя MIA, чтобы проверить, была ли конкретная точка данных включена в обучающий набор. В работе [18] авторы интегрировали несколько методов MIA в конвейер извлечения данных, чтобы систематически оценить их эффективность. Затем сравнили их производительность в этой интегрированной среде с результатами обычных тестов MIA, что позволяет оценить их практическую полезность в реальных сценариях извлечения.

В работе [19] предлагается FeatureLens — легковесная структура, которая действует как линза для анализа аномалий в признаках изображений. FeatureLens, включающая в себя экстрактор признаков изображений (IFE) и неглубокие классификаторы (например, SVM, MLP или XGBoost) с размерами моделей от 1000 до 30000 параметров, достигает высокой точности обнаружения — от 97,8% до 99,75% при оценке в замкнутом наборе данных и от 86,17% до 99,6% при оценке обобщаемости в атаках FGSM, PGD, C&W и DAmageNet, используя только 51-мерные признаки. Благодаря сочетанию высокой эффективности обнаружения с превосходной обобщающей способностью, интерпретируемостью и вычислительной эффективностью, FeatureLens предлагает практический путь к прозрачной и эффективной защите от враждебных действий.

Хороший обзор по безопасности MCP опубликован в работе [20]. Данная систематизация знаний (SoK) направлена на предоставление всеобъемлющей таксономии рисков в экосистеме MCP, различая враждебные угрозы безопасности (например, косвенное внедрение подсказок, отравление инструментов) и эпистемические угрозы безопасности (например, сбои выравнивания в распределенном делегировании инструментов). В работе анализируются структурные уязвимости примитивов MCP, в частности ресурсов, подсказок и инструментов, и демонстрируется, как «контекст» может быть использован для запуска несанкционированных операций в многоагентных средах. Кроме того, рассматриваются современные методы защиты, от криптографической проверки происхождения (ETDI) до проверки намерений во время выполнения. В обзоре приводится дорожная карта по обеспечению безопасности перехода от разговорных чат-ботов к автономным агентным системам.

Подход “LLM как судья” является на сегодняшний день одной из основных архитектурных моделей оценки результатов генеративного ИИ. В частности, такой подход широко применяется в оценке генерируемого кода. И, естественно, вот эти самый судьи-LLM могут быть атакованы с целью воздействовать на судейские решения. Типичная косвенная инъекция подсказок. В статье [21] авторы представляют первое крупномасштабное исследование взлома автоматизированных систем оценки кода на основе LLM в академическом контексте. В работе опубликован модифицированный набор данных, содержащий 25 000 состязательных студенческих работ, специально разработанных для академической оценки кода, полученный из различных реальных учебных курсов и дополненный рубриками и оценками, выставленными людьми. Исследуются атаки с использованием шести моделей LLM. Было обнаружено, что эти модели демонстрируют значительную уязвимость, особенно к атакам, основанным на убеждении и ролевых играх (до 97% успеха взлома).

Агентный ИИ знаменует собой серьезный сдвиг в том, как автономные системы рассуждают, планируют, и выполняют многоэтапные задачи. В отличие от традиционного подхода с использованием одной модели, агентные рабочие процессы интегрируют множество специализированных агентов с различными большими языковыми моделями (LLM), возможностями, дополненными инструментами, логикой оркестровки, и взаимодействиями с внешними системами для формирования динамических конвейеров, способных к автономному принятию решений и действиям. По мере ускорения внедрения в промышленности и научных исследованиях организации сталкиваются с центральной проблемой: как проектировать, разрабатывать и эксплуатировать рабочие процессы агентного ИИ производственного уровня, которые являются надежными, наблюдаемыми, поддерживаемыми и соответствуют требованиям безопасности и управления. В статье [22] представлено практическое комплексное руководство по проектированию, разработке и развертыванию систем агентного ИИ производственного качества. В работе

представлен структурированный жизненный цикл проектирования, охватывающий декомпозицию рабочих процессов, шаблоны проектирования многоагентных систем, протокол контекста модели (MCP), интеграцию инструментов, детерминированную оркестровку, соображения ответственного ИИ, и стратегии развертывания с учетом окружающей среды. Далее представлены девять основных лучших практик проектирования рабочих процессов агентного ИИ производственного уровня, включая проектирование с приоритетом инструментов вместо MCP, вызов чистых функций, агенты с одним инструментом и одной ответственностью, внешнее управление подсказками, проектирование модельного, четкое разделение между логикой рабочего процесса и серверами MCP, контейнеризированное развертывание для масштабируемых операций и соблюдение принципа «Keep it Simple, Stupid» (KISS) для поддержания простоты и надежности.

Больше анонсов интересных публикаций можно найти в блоге Абаванет[33].

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.

Библиография

[1] Намиот, Д. Е., Е. А. Ильюшин, and И. В. Чижов. "Искусственный интеллект и кибербезопасность." International Journal of Open Information Technologies 10.9 (2022): 135-147.
[2] Намиот, Д. Е., and Е. А. Ильюшин. "О киберрисках генеративного искусственного интеллекта." International Journal of Open Information Technologies 12.10 (2024): 109-119.
[3] NIST AI 100-2 E2025 https://csrc.nist.gov/pubs/ai/100/2/e2025/final Retrieved: Jan, 2026
[4] Namiot, Dmitry. "Artificial Intelligence in Cybersecurity. Chronicle. Issue 1." International Journal of Open Information Technologies 13.9 (2025): 34-42.
[5] Namiot, Dmitry. "On cyberattacks using Artificial Intelligence systems." International Journal of Open Information Technologies 12.9 (2024): 132-141.
[6] Намиот, Д. Е. "Искусственный Интеллект в Кибербезопасности. Хроника. Выпуск 4." International Journal of Open Information Technologies 14.1 (2026): 81-94.
[7] Lebed, S. V., et al. "Large Language Models in Cyberattacks." Doklady Mathematics. Vol. 110. No. Suppl 2. Moscow: Pleiades Publishing, 2024.
[8] Безопасность ИИ-агентов https://abava.blogspot.com/2025/12/blog-post_11.html Retrieved: Dec, 2025
[9] Намиот, Д. Е. Атаки на системы машинного обучения - общие проблемы и методы / Д. Е. Намиот, Е. А. Ильюшин, И. В. Чижов // International Journal of Open Information Technologies. – 2022. – Т. 10, № 3. – С. 17-22. – EDN DZFSKQ
[10] Cheng, Lei, et al. "Towards robust recommendation: A review and an adversarial robustness evaluation library." IEEE Transactions on Knowledge and Data Engineering (2025).
[11] Розничная торговля в цифровой экономике / В. П. Куприяновский, С. А. Синягов, Д. Е. Намиот [и др.] // International Journal of Open Information Technologies. – 2016. – Т. 4, № 7. – С. 1-12. – EDN WCMIWN.
[12] Zhang, Yu, et al. "A survey of large language model empowered agents for recommendation and search: Towards next-generation information retrieval." arXiv preprint arXiv:2503.05659 (2025).
[13] Li, Teng-Chi, Yen-Ku Liu, and Yun-Cheng Tsai. "AI-driven smart home energy optimization: integrating AI agents with IoT for adaptive decision-making." International Conference on Applied System Innovation (ICASI 2025). Vol. 2025. IET, 2025.
[14] Волков, А. А. О задачах создания эффективной инфраструктуры среды обитания / А. А. Волков, Д. Е. Намиот, М. А. Шнепс-Шнеппе // International Journal of Open Information Technologies. – 2013. – Т. 1, № 7. – С. 1-10. – EDN ROMIZX.
[15] Zhao, Dan, et al. "Security and privacy in smart homes: Challenges and latest developments." Advances in the Internet of Things. CRC Press, 2025. 36-55.
[16] Agentic AI Foundation https://aaif.io/ Retireved: Jan, 2026
[17] Thirumalaisamy, Karthikeyan. "Survey of Public Red-Teaming Frameworks for LLM: Techniques, Coverage, and Gaps."
[18] Sahili, Ali Al, Ali Chehab, and Razane Tajeddine. "On the Effectiveness of Membership Inference in Targeted Data Extraction from Large Language Models." arXiv preprint arXiv:2512.13352 (2025).
[19] Yang, Zhigang, et al. "FeatureLens: A Highly Generalizable and Interpretable Framework for Detecting Adversarial Examples Based on Image Features." arXiv preprint arXiv:2512.03625 (2025).
[20] Gaire, Shiva, et al. "Systematization of Knowledge: Security and Safety in the Model Context Protocol Ecosystem." arXiv preprint arXiv:2512.08290 (2025).
[21] Sahoo, Devanshu, et al. "How to Trick Your AI TA: A Systematic Study of Academic Jailbreaking in LLM Code Evaluation." arXiv preprint arXiv:2512.10415 (2025).
[22] Bandara, Eranga, et al. "A practical guide for designing, developing, and deploying production-grade agentic ai workflows." arXiv preprint arXiv:2512.08769 (2025).

Статья получена 29 января 2026.
Д.Е. Намиот – МГУ имени М.В. Ломоносова (e-mail: dnamiot@cs.msu.ru).

[33] http://abava.blogspot.com

Artificial Intelligence in Cybersecurity. Chronicle. Issue 5

Dmitry Namiot

***Abstract* - This article presents the fifth issue of a monthly analytical series dedicated to the study of current trends at the intersection of artificial intelligence (AI) and cybersecurity. This periodic review aims to systematically monitor and structuredly analyze key events, regulatory initiatives, and technological advances in this field. Each issue covers three areas: 1) Incident and Threat Analysis. This section examines practical cases, vulnerabilities, and emerging risks associated with the use of AI technologies in security. We focus on such phenomena as the exploitation of vulnerabilities in generative AI systems, the development of adversarial attacks on machine learning models, as well as threats inherent in AI agents; 2) Regulatory Landscape Review. Particular attention is paid to regulatory dynamics at the global and national levels. New legislation, strategic initiatives, industry standards, and recommendations that form the legal and operational framework for the safe implementation of AI in the context of cybersecurity are analyzed; 3) Scientific and Technological Chronicle. Each issue includes an annotated list of significant scientific publications, research reports, and descriptions of innovative developments that contribute to the advancement of the subject area under consideration. It should be noted that the selection of materials for each edition, as well as their interpretation, inevitably reflect the professional expertise and analytical perspective of the authors.**

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

REFERENCES

[1] Namiot, D. E., E. A. Il'jushin, and I. V. Chizhov. "Iskusstvennyj intellekt i kiberbezopasnost'." International Journal of Open Information Technologies 10.9 (2022): 135-147.
[2] Namiot, D. E., and E. A. Il'jushin. "O kiberriskah generativnogo iskusstvennogo intellekta." International Journal of Open Information Technologies 12.10 (2024): 109-119.
[3] NIST AI 100-2 E2025 https://csrc.nist.gov/pubs/ai/100/2/e2025/final Retrieved: Jan, 2026
[4] Namiot, Dmitry. "Artificial Intelligence in Cybersecurity. Chronicle. Issue 1." International Journal of Open Information Technologies 13.9 (2025): 34-42.
[5] Namiot, Dmitry. "On cyberattacks using Artificial Intelligence systems." International Journal of Open Information Technologies 12.9 (2024): 132-141.
[6] Namiot, D. E. "Iskusstvennyj Intellekt v Kiberbezopasnosti. Hronika. Vypusk 4." International Journal of Open Information Technologies 14.1 (2026): 81-94.
[7] Lebed, S. V., et al. "Large Language Models in Cyberattacks." Doklady Mathematics. Vol. 110. No. Suppl 2. Moscow: Pleiades Publishing, 2024.
[8] Bezopasnost' II-agentov https://abava.blogspot.com/2025/12/blog-post_11.html Retrieved: Dec, 2025
[9] Namiot, D. E. Ataki na sistemy mashinnogo obuchenija - obshhie problemy i metody / D. E. Namiot, E. A. Il'jushin, I. V. Chizhov // International Journal of Open Information Technologies. – 2022. – T. 10, # 3. – S. 17-22. – EDN DZFSKQ
[10] Cheng, Lei, et al. "Towards robust recommendation: A review and an adversarial robustness evaluation library." IEEE Transactions on Knowledge and Data Engineering (2025).
[11] Roznichnaja torgovlja v cifrovoj jekonomike / V. P. Kuprijanovskij, S. A. Sinjagov, D. E. Namiot [i dr.] // International Journal of Open Information Technologies. – 2016. – T. 4, # 7. – S. 1-12. – EDN WCMIWN.
[12] Zhang, Yu, et al. "A survey of large language model empowered agents for recommendation and search: Towards next-generation information retrieval." arXiv preprint arXiv:2503.05659 (2025).
[13] Li, Teng-Chi, Yen-Ku Liu, and Yun-Cheng Tsai. "AI-driven smart home energy optimization: integrating AI agents with IoT for adaptive decision-making." International Conference on Applied System Innovation (ICASI 2025). Vol. 2025. IET, 2025.
[14] Volkov, A. A. O zadachah sozdanija jeffektivnoj infrastruktury sredy obitanija / A. A. Volkov, D. E. Namiot, M. A. Shneps-Shneppe // International Journal of Open Information Technologies. – 2013. – T. 1, # 7. – S. 1-10. – EDN ROMIZX.
[15] Zhao, Dan, et al. "Security and privacy in smart homes: Challenges and latest developments." Advances in the Internet of Things. CRC Press, 2025. 36-55.
[16] Agentic AI Foundation https://aaif.io/ Retireved: Jan, 2026
[17] Thirumalaisamy, Karthikeyan. "Survey of Public Red-Teaming Frameworks for LLM: Techniques, Coverage, and Gaps."
[18] Sahili, Ali Al, Ali Chehab, and Razane Tajeddine. "On the Effectiveness of Membership Inference in Targeted Data Extraction from Large Language Models." arXiv preprint arXiv:2512.13352 (2025).
[19] Yang, Zhigang, et al. "FeatureLens: A Highly Generalizable and Interpretable Framework for Detecting Adversarial Examples Based on Image Features." arXiv preprint arXiv:2512.03625 (2025).
[20] Gaire, Shiva, et al. "Systematization of Knowledge: Security and Safety in the Model Context Protocol Ecosystem." arXiv preprint arXiv:2512.08290 (2025).
[21] Sahoo, Devanshu, et al. "How to Trick Your AI TA: A Systematic Study of Academic Jailbreaking in LLM Code Evaluation." arXiv preprint arXiv:2512.10415 (2025).
[22] Bandara, Eranga, et al. "A practical guide for designing, developing, and deploying production-grade agentic ai workflows." arXiv preprint arXiv:2512.08769 (2025)..