# Искусственный Интеллект в Кибербезопасности. Хроника. Выпуск 8

Д.Е. Намиот

***Аннотация*** **– Данная публикация представляет собой открывает очередной, восьмой по счёту, выпуск периодического аналитического обзора использования Искусственного интеллекта (ИИ) в кибербезопасности. Цикл этих материалов направлен на углублённое исследование стремительно эволюционирующей сферы, возникающей на стыке искусственного интеллекта и кибербезопасности. Ключевая цель данного проекта — планомерное отслеживание мировых тенденций и обобщение наиболее примечательных событий. Помимо сбора информации, в рамках инициативы проводится тщательный разбор законодательных инициатив, резонансных происшествий и передовых технологических новшеств, которые формируют контуры современной кибербезопасности под влиянием ИИ.**

**Каждый номер серии имеет унифицированную структуру, состоящую из трёх разделов, что гарантирует всестороннее освещение рассматриваемой тематики. Первый раздел фокусируется на разборе базы инцидентов и существующих вызовов безопасности: здесь исследуются реальные сценарии атак, обнаруживаются свежие уязвимости и даётся оценка угрозам, порождаемым внедрением алгоритмов ИИ как в оборонительные механизмы, так и в арсенал злоумышленников. Второй раздел даёт характеристику текущему состоянию нормативно-правовой среды и векторам её изменений. Осознание этих процессов имеет первостепенное значение, поскольку именно они задают правовые и эксплуатационные рамки, в которых должны будут развиваться надёжные и безопасные системы на базе ИИ. Третий раздел освещает хронику научно-технологических достижений. Каждый выпуск включает в себя аннотированный перечень наиболее весомых — с точки зрения авторов — научных работ, экспертных докладов ведущих организаций и описаний новаторских разработок.**

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

## I. Введение

С 2020 года кафедра информационной безопасности факультета вычислительной математики и кибернетики МГУ имени М.В. Ломоносова проводит исследования на стыке искусственного интеллекта (ИИ) и кибербезопасности. На факультете была открыта и успешно действует первая магистерская программа по данной тематике[1]. За время её существования состоялось несколько выпусков; подготовлено более 30 специалистов в рамках этой образовательной траектории. Многие магистерские диссертации выпускников легли в основу последующих прикладных решений в рассматриваемой области [1–5].

В ранних работах [6, 7] сотрудники кафедры выделили четыре основных вектора взаимодействия ИИ и кибербезопасности:
- использование ИИ для защиты информационных систем;
- применение ИИ в атакующих действиях;
- обеспечение защищённости самих систем ИИ;
- технология дипфейков.

Следует подчеркнуть высокую динамику эволюции этой предметной области. Феномен дипфейков представляет собой лишь одну из многих угроз, связанных с генеративными моделями [8], что делает необходимым комплексный анализ рисков, порождаемых синтезированным контентом. Показательный пример — 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) этот документ полностью включает технологии генеративного ИИ (GenAI) в свою таксономическую структуру, подробно описывая специфику атак на большие языковые модели (LLM), системы дополненной генерации поиска (RAG) и архитектуры на базе ИИ-агентов.

В соответствии с указанной таксономией строятся 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»[2]. В аналогичной логике формируется и готовящийся к изданию учебник, в выпуске которого, как ожидается, окажет содействие Центральный университет[3]. За время, прошедшее после публикации предыдущего выпуска «Хроники», для нового курса по разработке ИИ-агентов[4] обновлено учебно-методическое пособие, посвящённое вопросам безопасности ИИ-агентов[5].

В целом, за весь период существования магистратуры, на кафедре ИБ собран наиболее полный, по имеющимся данным, массив публикаций, в первую очередь, на русском языке по указанной проблематике[6].

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

[2] Магистратура Кибербезопасность https://cs.msu.ru/news/3916/

[3] https://cu.ru

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

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

[6] https://abava.blogspot.com/2026/05/15052026.html

Результатом планомерной работы в этой области стало создание нового продукта — регулярного обзора (хроники) текущих событий в сфере ИИ и кибербезопасности. В рамках этого 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 — один раз в месяц. Первый выпуск вышел в сентябре 2025 года [10]. В настоящее время продолжается поиск оптимальной формы распространения издания; в качестве возможных вариантов рассматриваются публикация отдельного PDF-документа на одном из ресурсов авторского коллектива, создание специализированного Telegram-канала, либо иные форматы. Восьмой 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[7]. Традиционно принимаются к рассмотрению новые статьи для публикации в журнале INJOIT[8] (издание входит в Перечень ВАК, РИНЦ, Белый список).

## II. Инциденты в ИИ

Reuters сообщает, что Банки США спешат устранить множество слабых мест в ИТ-системах, выявленных мощным, но дорогостоящим инструментом Mythos AI от Anthropic [11], что требует срочного обновления программного обеспечения и повышает вероятность сбоев для клиентов.

Как ранее сообщало Reuters, несколько крупнейших кредиторов правительства США в настоящее время имеют доступ к Mythos и выявляют проблемы, которые обнаруживает программа, сообщили несколько источников, знакомых с ситуацией. По мере того, как они изучают уязвимости, крупные банки также помогают информировать небольшие банки, которые не имеют прямого доступа к инструменту, чтобы они могли подготовить свои системы, сообщили эти источники. Эксперты по кибербезопасности считают, что Mythos представляет собой серьезную угрозу для банковской отрасли и ее устаревших технологических систем, что побудило регулирующие органы и политиков выступить с рядом предупреждений.

«Это тревожный сигнал, потому что киберриски смещаются в сторону машинной обработки данных, в то время как большая часть защиты банков по-прежнему работает на уровне человеческого фактора», — сказал Нитин Сет, соучредитель и генеральный директор Incedo, компании, предоставляющей услуги в области данных, цифровых технологий и искусственного интеллекта. «Это также опровергает давнее предположение в банковской безопасности, что уязвимости могут оставаться скрытыми в течение длительного времени, прежде чем они будут обнаружены и использованы в качестве оружия».

По словам нескольких источников, банки Уолл-стрит, тестируя Mythos, обнаруживают, что модель умеет объединять уязвимости низкого риска, или слабые места в уязвимости высокого риска. Это вызывает спешку в проверке обновления программного обеспечения[9].

А агентство Bloomberg пишет о том, что есть уже европейский аналог Mythos, который производитель (Mistral) также предлагает банкам[10].

База данных инцидентов ИИ[11] продолжает отмечать парад (триумфальное шествие) дипфейков. “Знаменитости” продолжают рекламировать товары и услуги, выдавать инвестиционные рекомендации и т.д. При этом отмечается возрастающая роль ИИ-агентов. Естественно, все такие атаки локализованы, поскольку в каждой стране могут быть свои авторитеты. Например, на странице в Facebook, представленной как хорватское новостное издание N1 HR, было опубликовано видео, в котором якобы хорватский футболист Лука Модрич рекламирует инвестиционную платформу Immediate Matrix и обещает доход до 4000 евро в месяц. Голос футболиста был синтезирован с помощью ИИ[12].

База описывает таксономию атак следующим образом.

1. Мошенничество с использованием синтетических медиа и обман потребителей.

Самая большая категория: дипфейки или созданные с помощью ИИ поддельные изображения, используемые для продажи инвестиций, медицинских товаров, чудодейственных лекарств, туристических услуг или других видов мошенничества, ориентированных на потребителей.

Хотя многие из этих случаев причинения вреда носят, по-видимому, финансовый характер, эта категория заслуживает внимания, поскольку, например, поддельное медицинское заключение может также повлиять на то, как человек понимает собственное здоровье и какие решения принимает в отношении лечения. Зафиксированы случаи использования поддельных видеоматериалов, демонстрирующих типичные уязвимости медицинской информации. Человек, ищущий помощи в решении проблем со здоровьем, может увидеть узнаваемого врача или известную личность и кликнуть по ссылке, не спросив, откуда на самом деле взялось видео. Такие случаи - это доказательство того, как узнаваемая экспертиза может быть использована не по назначению и стать менее

[7] dnamiot@cs.msu.ru

[8] http://injoit.org

[9] https://www.reuters.com/business/finance/anthropics-mythos-sends-us-banks-rushing-plug-cyber-holes-2026-05-12

[10] https://www.bloomberg.com/news/articles/2026-05-13/mistral-developing-new-ai-model-for-banks-lacking-mythos-access

[11] https://incidentdatabase.ai

[12] https://incidentdatabase.ai/cite/1456/

заслуживающей доверия.

2. Политическая/геополитическая дезинформация и операции по искусственному влиянию.

Искусственные медиа, фейковые персонажи, манипулированные изображения, политическая реклама, созданная с помощью ИИ, маскировка искусственной общественной поддержки под общественную инициативу (астротурфинг[13]), военные изображения и дезинформация о публичных мероприятиях.

Некоторые из них связаны с очевидной дезинформацией в отношении общественных деятелей, выборов, протестов или войны. К ним относятся якобы созданное с помощью ИИ расистское видео, распространенное Дональдом Трампом, изображающее Барака и Мишель Обаму в виде обезьян; якобы созданное с помощью ИИ изображение, распространяемое перед всеобщими выборами в Таиланде в феврале 2026 года, изображающее премьер-министра за ужином с южноафриканским бизнесменом; якобы созданные с помощью ИИ видеоролики в TikTok, призывающие к выходу Польши из ЕС в декабре 2025 года; якобы созданные с помощью ИИ военные кадры, распространяемые во время начальной фазы войны в Иране; якобы созданные Gemini изображения, утверждающие, что американские солдаты из отряда Delta Force были захвачены Корпусом стражей исламской революции.

Болгарская политическая история особенно полезна для подобных случаев, поскольку она не опирается на одну впечатляющую подделку. Сообщалось, что сеть якобы фейковых профилей в Facebook, использующих якобы сгенерированные ИИ изображения профилей, усиливала публикации одной болгарской партии «Есть такой народ» в предвыборный период в феврале 2026 года. В качестве примера это хороший образец того, как синтетический политический вред может действовать посредством усиления, а не с помощью одного вирусного артефакта, и как синтетический элемент помогает аккаунту выглядеть достаточно человечным, чтобы участвовать в механизме привлечения внимания платформы.

À la guerre comme à la guerre - фальшивые космические снимки Иранской войны[14].

3. Неприкосновенность частной жизни, идентичность, голос/изображение и злоупотребление репутацией.

Случаи, когда основной вред заключается в раскрытии информации, несанкционированном копировании, злоупотреблении идентичностью, злоупотреблении изображением, раскрытии личных данных, нанесении ущерба репутации или нежелательной проверке человеком.

При этом не каждый случай использования ИИ в медиа представляет собо мошенничество. Например, сообщается, что Moltbook раскрыл личную переписку пользователей и токены аутентификации API. NotebookLM якобы скопировал голос ведущего NPR без его согласия; роботы-пылесосы DJI Romo раскрыли данные камеры, микрофона и карты дома; умные очки Meta AI раскрыли интимные изображения и видео пользователей экспертам в Кении; Grok раскрыл настоящее имя и дату рождения порноактрисы, что якобы способствовало раскрытию личных данных и преследованию. Это различие помогает избежать объединения каждого инцидента с использованием ИИ в медиа в категорию «дезинформация» или «мошенничество». Многие из этих инцидентов связаны с социальными и техническими сбоями, которые делают людей уязвимыми.

4. Искусственное сексуальное насилие, причинение вреда интимным изображениям и сексуальные домогательства.

Создание интимных изображений без согласия, сексуализированные дипфейки, обвинения, связанные с сексуальным насилием над детьми, шантаж с использованием сексуальных материалов и сексуальные домогательства в отношении лиц, находящихся в общественном или частном порядке.

Эта конкретная группа включает значительное количество инцидентов, связанных с искусственным сексуальным насилием, созданием интимных изображений без согласия, шантажом с использованием сексуальных материалов или сексуальными домогательствами. Примерами являются сообщения о предполагаемой порнографии, созданной с помощью ИИ методом дипфейк, направленной против общественных деятелей и несовершеннолетних на Филиппинах; предполагаемые сексуализированные изображения влиятельной личности в социальных сетях из Техаса, созданные с помощью ИИ; предполагаемые видеоролики, созданные с помощью ИИ с участием учеников средней школы; сообщения Grok о якобы сексуализированных изображениях Рене Гуд после ее убийства в Миннеаполисе; предполагаемые обнаженные изображения, созданные с помощью ИИ, использованные для шантажа мужчины из штата Канзас; и поддельные обнаженные изображения, якобы созданные из фотографий девушек, включая учениц, из социальных сетей бывшим учителем из Нового Орлеана. Повторение этих инцидентов свидетельствует о том, что сексуализация стала одним из наиболее прямых способов использования систем искусственного интеллекта для преобразования онлайн-оскорблений в личный вред, и те же самые записи также показывают, почему вопрос подлинности может быть второстепенным по сравнению с ущербом, причиненным созданием и распространением этих изображений и видео.

5. Недостоверность правовых, политических, журналистских и официальных документов.

Поддельные цитаты, сфабрикованные высказывания, юридические документы, созданные с помощью ИИ, галлюцинаторные ссылки, неправильный перевод в контексте вещания, ложные доказательства и

[13] https://ru.wikipedia.org/wiki/%D0%90%D1%81%D1%82%D1%80%D0%BE%D1%82%D1%83%D1%80%D1%84%D0%B8%D0%BD%D0%B3

[14] https://flowingdata.com/2026/03/09/satellite-images-that-are-ai-fakes/

институциональные/профессиональные нарушения доверия.

В институциональных документах продолжают появляться случаи нарушения достоверности данных, созданных с помощью ИИ. Отдельная группа случаев касается судов, политических документов, журналистики и профессиональной среды, где материалы, созданные или обработанные с помощью ИИ, предположительно подорвали достоверность официальных или доказательных документов. Некоторые примеры из этой группы включают случай с женщиной из Флориды в ноябре 2024 года, которая, как сообщается, была заключена в тюрьму после того, как бывший парень якобы представил сфабрикованный с помощью ИИ скриншот текста в качестве доказательства нарушения условий освобождения под залог. Ars Technica отозвала статью после того, как якобы сгенерированный с помощью ИИ текст был представлен как прямые цитаты из текста разработчика Matplotlib в статье об инциденте с использованием ИИ-программиста, связанном с этим же разработчиком. Сообщается, что адвокат Министерства юстиции США использовал ИИ для подачи ходатайства с, казалось бы, сфабрикованными цитатами и искаженными решениями по делу. Окружной суд в США наложил санкции на адвокатов в деле, связанном с предполагаемыми поддельными апелляционными ссылками. Сообщается, что в проекте национальной политики Южной Африки в области ИИ содержались вымышленные упоминания, которые, как считается, являются галлюцинациями, вызванными ИИ[15].

6. Чат-боты, ИИ-компаньоны и риски межличностного общения или самоповреждения.

Взаимодействие с чат-ботами, связанное с самоповреждением, подкреплением заблуждений, эмоциональной зависимостью, опасными советами, сбоями в работе службы поддержки клиентов или предполагаемым нежеланием сообщать о серьезных рисках. В подборке записей, связанных с чат-ботами, фигурируют предполагаемые случаи членовредительства, подкрепления заблуждений, эмоциональной зависимости, опасных советов и сбоев в работе службы поддержки клиентов.

7. Государственный сектор, полиция, гражданские свободы и сбои в принятии институциональных решений.

Распознавание лиц, наблюдение, доступ к государственным услугам, контекст полиции или границ, институциональные решения и административные или корпоративные действия, находящиеся под влиянием ИИ.

Системы ИИ с правами на выполнение создают другой тип сбоев рабочего процесса Здесь мы наблюдаем рост числа сбоев в работе агентов или операционного программного обеспечения. Например, агент Claude Code, как сообщается, удалил производственную инфраструктуру DataTalks.Club, базу данных и снимки через Terraform. Google Antigravity, как сообщается, удалил весь диск D: пользователя при попытке очистить кэш проекта. Claude Cowork, предположительно, удалил папку с семейными фотографиями за 15 лет при организации рабочего стола. Агент Cursor AI, как сообщается, удалил производственную базу данных PocketOS при работе над задачей в тестовой среде[16].

Связанный, но отличающийся набор инцидентов касается враждебного использования, в отличие от сбоев в рабочем процессе, описанных выше. Примеры включают в себя сообщения о том, что AkiraBot использовал API чата OpenAI для генерации спама в чатах веб-сайта и контактных формах; сообщения о том, что вредоносные навыки в экосистеме OpenClaw распространили AMOS Stealer и украли учетные данные через ClawHub; Anthropic заявила, что DeepSeek, Moonshot и MiniMax использовали мошеннические учетные записи и прокси-сервисы для масштабного использования возможностей Claude; сообщения о том, что автономный наступательный агент CodeWall получил несанкционированный доступ к базе данных платформы Lilli AI компании McKinsey; и сообщения о том, что Claude был взломан, чтобы помочь украсть конфиденциальные данные мексиканского правительства[17]. Эти случаи частично совпадают с категорией агентных систем, где задействованы автономные или агентоподобные системы, но здесь речь идет о злоупотреблении или несанкционированном использовании, а не о неправильном выполнении делегированной работы.

8. Автономия в физическом мире, навигация, робототехника и безопасность клинических устройств.

Автобусы, фургоны, роботакси, роботы-доставщики, автономные транспортные средства, системы клинической навигации и автоматизированные оповещения, где предположения программного обеспечения приводят к физическим последствиям.

Инфраструктура затрудняет рассмотрение сбоев ИИ в физическом мире как абстрактных явлений. В штате Вашингтон система навигации транспортного управления Спокана, как сообщается, направила двухэтажный автобус к низкому мосту, в результате чего пострадали семь человек). Предположительно, сгенерированное ИИ предупреждение о сепсисе, как сообщается, привело к потенциально ненадлежащему внутривенному введению жидкости пациенту, находящемуся на диализе, чего удалось избежать благодаря вмешательству врача. Сообщается, что фургон доставки Amazon застрял на Брумвее в Эссексе, Англия, после того, как GPS направил его на приливные отмели. Клиническая навигационная система якобы дала неверные указания во время операции на пазухах носа, что, как сообщается, способствовало инсульту у

[15] https://www.news24.com/business/tech/govts-draft-ai-policy-cites-fictitious-references-experts-believe-are-ai-hallucinations-20260424-1085

[16] https://www.pcgamer.com/software/ai/here-we-go-again-ai-deletes-entire-company-database-and-all-backups-in-9-seconds-then-cheerfully-admits-i-violated-every-principle-i-was-given/

[17] https://www.bloomberg.com/news/articles/2026-02-25/hacker-used-anthropic-s-claude-to-steal-sensitive-mexican-data

пациента. В январе 2026 года робот-доставщик Coco Robotics, как сообщается, застрял на железнодорожных путях и был сбит поездом. Сообщается, что роботакси Baidu Apollo Go остановились в пробке во время сбоя системы в Ухане[18].

9. Сбои в работе агентного/операционного программного обеспечения и рабочих процессов.

Агенты ИИ или инструменты кодирования/рабочих процессов, работающие внутри репозиториев, файловых систем, облачной инфраструктуры, кэшей проектов, производственных баз данных и распределения задач в открытом исходном коде.

10. Кибербезопасность, вредоносная автоматизация, неправомерное использование моделей и операции с использованием ИИ в кибербезопасности.

Спам с использованием ИИ, вредоносные навыки, кража учетных данных, несанкционированный доступ, обвинения в дистилляции моделей, кража данных с использованием джейлбрейков и противодействие использованию систем ИИ в кибербезопасности.

11. Азартные игры, профилирование и поведенческая эксплуатация.

Системы оценки рисков на основе машинного обучения, автоматизированное профилирование, целевой маркетинг и предполагаемая эксплуатация уязвимых игроков.

Европол выпустил интересны отчет “Анализ угроз организованной интернет-преступности (IOCTA)” - это исследование Европола, посвященное меняющимся угрозам и тенденциям в сфере киберпреступности, с акцентом на изменениях за последние 12 месяцев - как шифрование, прокси-серверы и искусственный интеллект расширяют масштабы киберпреступности. В этом отчете подробно рассматриваются различные криминальные тенденции в сфере пособников киберпреступлений, мошеннических онлайн-схем, кибератак и сексуальной эксплуатации детей в интернете[19].

## III Регуляции и стандарты

OpenAI выступает за создание глобальной структуры по управлению искусственным интеллектом и его регулированию под руководством США, а также с участием Китая. Об этом пишет агентство Bloomberg со ссылкой на вице-президента компании по глобальным вопросам Криса Лехейна.

«ИИ на определенном уровне выходит за рамки многих существующих или традиционных торговых вопросов. Есть возможность действительно начать строить что-то глобальное и привлечь к участию страны со всего мира, включая Китай», — заявил Лехейн журналистам в офисе компании в Вашингтоне в преддверии встречи американского лидера Дональда Трампа и председателя КНР Си Цзиньпина.

Он уточнил, что подобная организация могла бы по замыслу и функционалу напоминать Международное агентство по атомной энергии (МАГАТЭ), устанавливающее глобальные стандарты безопасности для развития ядерной энергетики с целью предотвращения распространения оружия. Как считает Лехейн, один из способов это сделать — установить контакты между Центром стандартов и инноваций в области ИИ Минторга США с учреждениями по безопасности ИИ в других странах мира[20].

Российские власти хотят заставить российские соцсети самостоятельно воевать с дипфейками. Главный радиочастотный центр Роскомнадзора предложил соцсетям самостоятельно выявлять и ограничивать распространение дипфейков до проверки сведений у упомянутых в них лиц или организаций. В ведомстве считают, что для такого регулирования нужны понятные механизмы быстрого реагирования: ИТ-платформы должны уметь останавливать распространение подозрительного контента, пока его достоверность не подтверждена[21].

Отметим, что технической возможности точно определить, какой контент создан при помощи технологий искусственного интеллекта сейчас уже нет.

В Роскомнадзоре же считают (пояснили газете «Коммерсант»), что регулирование дипфейков должно основываться на понятных, открытых и технически реализуемых механизмах их оперативного выявления и прекращения распространения. Одним из возможных подходов, по мнению ведомства, является наделение социальных сетей ИТ-инструментами самостоятельного контроля за быстрым распространением подобного контента. При этом социальные сети должны получить право временно приостанавливать распространение спорного материала до проведения проверки его достоверности у доверенного источника, в том числе у лица или организации, фигурирующих в контенте.

Центр стандартов и инноваций в области искусственного интеллекта (CAISI) при Национальном институте стандартов и технологий Министерства торговли объявил о новых соглашениях с Google DeepMind, Microsoft и xAI. Благодаря расширению отраслевых коллабораций CAISI будет проводить предварительные оценки и целевые исследования для лучшей оценки возможностей передового ИИ и повышения уровня безопасности ИИ. Эти соглашения основаны на ранее объявленных партнёрствах, которые были пересмотрены в соответствии с директивами CAISI министра торговли и Американским планом действий по искусственному интеллекту.

Под руководством госсекретаря Говарда Латника CAISI назначена в качестве основного контактного центра отрасли в правительстве США для содействия

[18] https://www.wired.com/story/robotaxi-outage-in-china-leaves-passengers-stuck-in-cars-on-highways/

[19] https://www.europol.europa.eu/publication-events/main-reports/iocta-2026-evolving-threat-landscape

[20] https://www.rbc.ru/rbcfreenews/6a056ad99a79478264d11834

[21] https://www.cnews.ru/news/top/2026-05-15_sotsseti_zastavyat_samostoyatelno

тестированию, совместным исследованиям и разработке лучших практик, связанных с коммерческими системами ИИ.

Соглашения CAISI с разработчиками передового ИИ позволяют государственным органам оценивать модели ИИ до их публичного доступа, а также проводить оценку и другие исследования после внедрения. На сегодняшний день CAISI провела более 40 подобных оценок, включая современные модели, которые до сих пор не были представлены.

Эти соглашения поддерживают обмен информацией, стимулируют добровольные улучшения продуктов и обеспечивают чёткое понимание возможностей ИИ и состояния международной конкуренции в области ИИ. Для тщательной оценки возможностей и рисков, связанных с национальной безопасностью, разработчики часто предоставляют CAISI модели, которые уменьшили или убрали меры безопасности. Оценщики со всего правительства могут участвовать в оценках и регулярно предоставлять обратную связь через созванную CAISI Taskforce TRAINS — группу межведомственных экспертов, сосредоточенную на вопросах национальной безопасности ИИ. Соглашения поддерживают тестирование в засекреченных средах и были составлены с гибкостью, необходимой для быстрого реагирования на дальнейшие достижения ИИ[22].

Госдепартамент США распорядился о проведении глобальной кампании по привлечению внимания к широкомасштабным, по его словам, попыткам китайских компаний, включая стартап DeepSeek, занимающийся разработкой искусственного интеллекта, украсть интеллектуальную собственность из американских лабораторий искусственного интеллекта, говорится в дипломатической телеграмме, с которой ознакомилось агентство Reuters. В телеграмме, датированной пятницей и направленной в дипломатические и консульские представительства по всему миру, сотрудникам дипломатических ведомств предписывается обсудить со своими зарубежными коллегами «опасения по поводу извлечения и переработки противниками американских моделей искусственного интеллекта»[23].

В июле 2024 года Google в сотрудничестве с другими отраслевыми партнерами представил «Коалицию за безопасный ИИ (CoSAI)»[24], открытую экосистему ведущих экспертов в области ИИ и безопасности, работающих вместе над безопасным развертыванием ИИ, исследованиями и разработкой продуктов.

Участие Google в CoSAI остается на переднем крае новых инноваций в области ИИ, и особенно в рамках общего направления работы, посвященного «Шаблонам безопасного проектирования для агентных систем».

Используя многолетний опыт проектирования передовой инфраструктуры безопасности, Google недавно сотрудничал с CoSAI для разработки этих отраслевых принципов (человеческий контроль, инновации с контролем и прозрачностью) для создания агентов ИИ, безопасных по своей конструкции.

Благодаря сотрудничеству в продвижении этих принципов безопасности, более 40 ведущих компаний CoSAI, занимающихся ИИ и безопасностью, снижают критические риски и прокладывают путь к безопасному и ответственному развитию агентного ИИ. Эти совместные усилия показывают, что коллективные действия являются ключом к реализации преобразующего потенциала агентов ИИ при эффективном управлении их рисками.[25]

Интересный материал о построении суверенного ИИ в Южной Корее[26].

1. Документ начинается с констатации, что ИИ стал критическим ресурсом национальной безопасности. Однако глобальный порядок в этой сфере поляризован между США и Китаем. Южная Корея, как и многие другие страны, сталкивается с *вынужденной зависимостью* от ИИ-экосистем двух сверхдержав, что создаёт стратегическую уязвимость. При этом Южная Корея признана одной из немногих «средних ИИ-держав» (AI middle power), способных играть значимую роль в глобальных цепочках поставок благодаря преимуществам в производстве полупроводников и обрабатывающей промышленности.
2. Переосмысление концепции «суверенного ИИ» для средних держав.

Автор критикует традиционный подход к суверенитету, который требует полной автономии и локализации всех компонентов ИИ («всё самому»). Для средних держав (с ограниченными ресурсами) такое понимание нереалистично и может привести к ещё большей зависимости. Вместо этого предлагается новое определение суверенного ИИ как способности обеспечить «стратегическую автономию» (strategic autonomy), «свободу манёвра» (optionality) и «агентность» (agency):

- Контроль над критически важными данными, вычислениями, моделями и нормами.
- Возможность поддерживать функции ИИ в кризисных ситуациях без внешних сбоев.
- Защита культурной и языковой идентичности от искажения внешними моделями.

3. Типология стратегий средних ИИ-держав. В документе выделяются два основных типа стратегий, которые уже используют другие страны:

Специализация (Specialization Type): Концентрация на узких технологических нишах или «бутылочных горлышках» (Япония - HPC-инфраструктура, Канада - фундаментальные исследования, Сингапур и Тайвань -

[22] https://www.nist.gov/news-events/news/2026/05/caisi-signs-agreements-regarding-frontier-ai-national-security-testing

[23] https://www.reuters.com/world/china/us-state-dept-orders-global-warning-about-alleged-china-ai-thefts-by-deepseek-2026-04-24/

[24] https://www.coalitionforsecureai.org/

[25] https://static.googleusercontent.com/media/publicpolicy.google/en//resources/anei_strengthening_security.pdf

[26] https://www.inss.re.kr/common/download.do?atchFileId=F20260408094742209&fileSn=0

региональные языковые модели).

Альянс и сотрудничество (Alliance and Cooperation Type): Создание многосторонних экосистем и распределённой инфраструктуры на основе общих ценностей (ЕС - проект GAIA-X для обеспечения суверенитета данных, Африканский союз - континентальная политика обмена данными).

4. Диагностика текущей ситуации в Южной Корее. Автор отмечает, что амбициозная цель правительства - «Глобальные топ-3 по ИИ» (G3) сопровождается заявкой на создание «полного стека» (full-stack package): данные, вычисления, модели, безопасность, кадры и нормы.

Однако на практике корейская стратегия сталкивается с проблемами:

- Высокая внешняя зависимость по ключевым компонентам экосистемы.
- Риск неэффективных инвестиций и давления на экономику из-за попыток локализовать всё и сразу.
- Отставание по таким параметрам, как кадры (13-е место в мире), операционная среда (35-е) и исследования (13-е) согласно индексу Tortoise GAII.

5. Предложение корейской модели суверенного ИИ. Основная идея — «специализированный полный стек» (Specialized Full-Stack). Корея не должна копировать подход США/Китая (вертикальная интеграция для глобальной гегемонии). Вместо этого предлагается:

A. Перераспределение ресурсов: сфокусироваться на моделях, специализированных для корейского языка, культуры и ключевых отраслей (финансы, производство, здравоохранение), а не на «гонке за общими LLM».

B. Усиление преимущества в полупроводниках: использовать лидерство в HBM (высокопроизводительная память) как рычаг влияния (например, каждый новый GPU NVIDIA требует HBM3E от корейских компаний).

C. Реалистичный подход к вычислительным ресурсам: признать, что «не только GPU, но и энергия» - критический фактор. Необходимо развивать энергетическую инфраструктуру (мощности GW-уровня, охлаждение) для эффективной работы дата-центров.

D. Развитие кадров и бюрократической эффективности: преодолеть структурные слабости в реализации политики, включая нехватку AI-специалистов и слабую операционную среду.

6. Нормативное и дипломатическое лидерство: «устойчивый ИИ» Уникальное предложение документа - выйти за рамки узконационального суверенитета и занять лидерство в повестке устойчивого ИИ (Sustainable AI). Автор предлагает:

- ✓ Использовать экологические проблемы ИИ (электронные отходы «E-waste», огромное энергопотребление дата-центров) как дипломатический козырь.
- ✓ Развивать направления: высокоэффективные/низкопотребляющие чипы, углеродно-нейтральные ЦОД, технологии переработки отходов.
- ✓ Инициировать международные «зелёные цифровые партнёрства» с развивающимися странами, тем самым позиционируя Корею как ответственного глобального посредника на площадках ООН, ОЭСР, ITU.

7. Ключевой вывод: корейская стратегия суверенного ИИ должна строиться не на изоляционизме или тотальной самодостаточности, а на прагматичном сочетании специализации, стратегических альянсов (особенно со странами-единомышленниками) и нормативного лидерства в области устойчивого развития и доверенного ИИ. Это позволит обеспечить «горизонтальное лидерство» в противовес вертикальной гегемонии США и Китая.

## IV Обзор публикаций и проектов

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

Подробный обзор безопасности ИИ-агентов. В реальных приложениях быстро появляются агенты ИИ, которые объединяют большие языковые модели с компонентами не-ИИ систем, предлагая беспрецедентную автоматизацию и гибкость. Однако эта беспрецедентная гибкость порождает сложные проблемы безопасности, которые отличаются от проблем, встречающихся в традиционных программных системах. В работе представлена первая всесторонняя систематизация знаний о безопасности агентов ИИ, включая анализ пространства проектирования агентов, ландшафта атак и механизмов защиты для безопасных систем агентов ИИ. Авторы также выявляют открытые проблемы, указывающие на перспективные направления будущих исследований в этой новой области. Работа представляет собой первую систематическую структуру для понимания рисков безопасности и ландшафтов защиты агентов ИИ, служащую основой для создания как безопасных агентных систем, так и для продвижения исследований в этой критически важной области [12].

Физические бэкдоры. Атаки с использованием бэкдоров направлены на внедрение скрытого бэкдора в глубокие нейронные сети (DNN), так что предсказания зараженных моделей будут злонамеренно изменены, если скрытый бэкдор будет активирован заданным злоумышленником шаблоном триггера. Поскольку зараженные модели ведут себя нормально при предсказании безобидных образцов, атака с использованием бэкдоров является скрытой и, следовательно, представляет серьезную угрозу для практического применения DNN. В настоящее время большинство существующих атак с использованием бэкдоров используют статический триггер, то есть триггеры на обучающих и тестовых изображениях имеют одинаковый внешний вид и расположены в

одной и той же области. В этой статье авторы пересматривают эту парадигму атаки, анализируя характеристики триггеров. Демонстрируется, что эта парадигма атаки уязвима, когда триггер на тестовых изображениях не совпадает с триггером, используемым для обучения. Таким образом, эти атаки гораздо менее эффективны в физическом мире, где местоположение и внешний вид триггера в оцифрованных тестовых образцах могут отличаться от таковых на изображениях, используемых для обучения. Кроме того, вводится модуль усиления атаки во время обучения, вдохновленный методом ожидания над преобразованием (EOT), чтобы уменьшить уязвимость, связанную с такой несогласованностью. Показывается, что широко распространенное расширение данных может усугубить риски безопасности, связанные с атаками типа «бэкдор», хотя и может повысить производительность модели. Предложенные методы оцениваются на нескольких эталонных наборах данных, чтобы проверить их эффективность. Авторы надеются, что эта работа вдохновит на дальнейшие исследования свойств атак типа «бэкдор», что облегчит разработку более надежных и безопасных нейронных сетей [13].

Аудит ИИ агентов. Тема агентов искусственного интеллекта является одной из самых актуальных в развитии искусственного интеллекта. Агенты предназначены для объединения генеративных моделей с традиционным программным обеспечением. В основе этой комбинации лежат генеративные модели, представляющие собой искусственный интеллект, который должен планировать решения заданных задач, отказываясь от жестких алгоритмических моделей. Эта область агентов ИИ развивается даже быстрее, чем разработка больших языковых моделей. Агенты ИИ уже рассматриваются как новая альтернативная парадигма программирования. Однако вопросы доверия к системам искусственного интеллекта, которые на самом деле представляют собой доверие к результатам таких систем, не решаются просто из-за смены парадигмы использования. Отсутствуют строгие формальные доказательства функциональности больших языковых моделей и, следовательно, агентов ИИ. Аудит систем искусственного интеллекта — это практический способ убедиться, в отсутствие формальных доказательств, что все доступные в настоящее время шаги по повышению уверенности в производительности системы были предусмотрены и реализованы [14].

Системы безопасности для высокопроизводительных вычислений (HPC). Высокопроизводительные вычислительные системы (ВВП) обеспечивают фундаментальную вычислительную инфраструктуру для крупномасштабных и сложных симуляций, анализа больших данных и обучения моделей искусственного интеллекта (ИИ) и машинного обучения (МО), и все это с исключительной скоростью. Обеспечение безопасности систем ВВП имеет важное значение для защиты моделей ИИ, конфиденциальных данных и реализации всех преимуществ ВВП. Система ВВП использует специализированное оборудование, программное обеспечение и высокоскоростные сети в сложных пользовательских средах, и высокая производительность является фундаментальным требованием к системе. В этом специальном издании NIST представлено отображение мер безопасности для ВВП, разработанное для решения этих уникальных задач и требований безопасности. Основанное на базовом уровне, определенном в NIST SP 800-53B, наложение адаптирует 60 мер безопасности из NIST SP 800-53 с дополнительными рекомендациями и/или обсуждениями для повышения их применимости в контексте ВВП. Это отображение призвано предоставить практические, ориентированные на производительность рекомендации по безопасности, которые могут быть легко внедрены. Для многих организаций это обеспечивает надежную основу для защиты высокопроизводительных вычислительных сред, а также позволяет вносить дальнейшие изменения для удовлетворения конкретных оперативных или служебных потребностей. Данный документ предназначен для использования менеджерами по ИТ-безопасности, специалистами по соблюдению нормативных требований, системными администраторами высокопроизводительных вычислительных систем и руководителями программ в ведомствах, ответственными за обеспечение безопасности высокопроизводительных вычислительных сред [15].

Формальная модель безопасности MCP. Протокол контекста модели (MCP), представленный Anthropic в ноябре 2024 года и теперь управляемый фондом Agentic AI Foundation при Linux Foundation, быстро стал стандартом де-факто для подключения агентов на основе больших языковых моделей (LLM) к внешним инструментам и источникам данных, с более чем 97 миллионами ежемесячных загрузок SDK и более чем 177 000 зарегистрированных инструментов. Однако это стремительное распространение выявило критический пробел: отсутствие единой, формальной структуры безопасности, способной систематически характеризовать, анализировать и смягчать разнообразные угрозы, с которыми сталкиваются экосистемы агентов на основе MCP. Существующие исследования в области безопасности остаются фрагментированными, охватывая отдельные статьи об атаках, изолированные бенчмарки и точечные механизмы защиты. В этой статье представлен MCPSHIELD, всеобъемлющая формальная структура безопасности для агентов ИИ на основе MCP. Авторы отмечают четыре основных вклада: (1) иерархическую таксономию угроз, включающую 7 категорий угроз и 23 различных вектора атак, организованных по четырем поверхностям атаки, основанную на анализе более 177 000 инструментов MCP; (2) формальную модель верификации, основанную на размеченных системах переходов с аннотациями границ доверия, которая позволяет проводить статический и анализ в реальном времени цепочек взаимодействия инструментов MCP; (3) систематическую сравнительную оценку 12 существующих механизмов защиты, выявляющую

пробелы в охвате нашей таксономии угроз; и (4) эталонную архитектуру многоуровневой защиты, интегрирующую контроль доступа на основе возможностей, аттестацию криптографических инструментов, отслеживание потока информации и обеспечение соблюдения политик в реальном времени. Проведенный анализ показывает, что ни один из существующих механизмов защиты не охватывает более 34% выявленного ландшафта угроз, в то время как интегрированная архитектура MCPSHIELD достигает теоретического охвата в 91%. Выделяется семь открытых исследовательских задач, которые необходимо решить для обеспечения безопасности следующего поколения агентных систем искусственного интеллекта [16].

Методы обнаружения дипфейков в видеоконференциях в реальном времени. В последние годы видеоконференции приобретают все более широкий размах, став неотъемлемым инструментом для проведения деловых совещаний, образовательных мероприятий и даже официальных правительственных встреч. Стремительное развитие технологий интернет-связи и доступность платформ видеоконференций (таких как Zoom, Microsoft Teams и Google Meet) способствуют переходу множества организаций на гибридные и дистанционные форматы работы. В результате глобальная аудитория пользователей онлайн-встреч исчисляется сотнями миллионов, и это число продолжает расти. Одновременно с расширением сферы применения видеоконференций возникает новая волна угроз, связанных с безопасностью и доверием участников. Среди таких угроз особенно выделяется феномен "дипфейков" (от англ. deepfakes), то есть синтетически сгенерированных или модифицированных аудио- и видеозаписей, которые практически невозможно отличить от оригинала невооруженным глазом. В работе рассматривается вопрос детектирования дипфейков в реальном времени в видеоконференциях [17].

Проверка кода ИИ-агентов. Что должен проверить разработчик перед развертыванием агента LLM: модель, код инструмента, конфигурацию развертывания или все три? На практике многие сбои безопасности в агентских системах возникают не только из-за весов модели, но и из-за окружающего программного стека: функций инструмента, передающих ненадежные входные данные опасным операциям, раскрытых учетных данных в артефактах развертывания и чрезмерно привилегированных конфигураций протокола контекста модели (MCP). В работе представлена Agent Audit, система анализа безопасности для приложений агентов LLM. Agent Audit анализирует код агента на Python и артефакты развертывания с помощью конвейера, учитывающего особенности агента, который объединяет анализ потока данных, обнаружение учетных данных, структурированный анализ конфигурации и проверки рисков привилегий. Система сообщает о результатах в форматах терминала, JSON и SARIF, что позволяет напрямую интегрировать систему с локальными рабочими процессами разработки и конвейерами CI/CD. На тестовой выборке из 22 образцов с 42 аннотированными уязвимостями Agent Audit обнаруживает 40 уязвимостей с 6 ложными срабатываниями, существенно улучшая полноту обнаружения по сравнению с распространенными базовыми показателями SAST, сохраняя при этом время сканирования менее секунды. Agent Audit является открытым исходным кодом и устанавливается через pip, что делает аудит безопасности доступным для агентских систем. В ходе живой демонстрации участники сканируют уязвимые репозитории агентов и наблюдают, как Agent Audit выявляет риски безопасности в функциях инструмента, подсказках и многом другом. Результаты связаны с местоположением исходного кода и путями конфигурации и могут быть экспортированы в VS Code и GitHub Code Scanning для интерактивного анализа [18].

Последовательные фейки. Контент типа «дипфейк» в социальных сетях все чаще создается путем многократных последовательных правок биометрических данных, таких как изображения лиц. В результате окончательный вид изображения часто отражает скрытую цепочку операций, а не единичную манипуляцию. Восстановление этих историй редактирования имеет важное значение для визуального анализа происхождения, аудита дезинформации и рабочих процессов криминалистической экспертизы или модерации платформ, которые должны отслеживать происхождение и эволюцию медиаконтента, созданного ИИ. Однако существующие наборы данных преимущественно фокусируются на одноэтапном редактировании и игнорируют кумулятивные артефакты, вносимые реалистичными многоэтапными конвейерами. Чтобы устранить этот пробел, авторы представляют Sequential Editing in Diffusion (SEED), крупномасштабный бенчмарк для отслеживания последовательного происхождения изображений лиц. SEED содержит более 90 000 изображений, созданных с помощью одной-четырех последовательных редактирований атрибутов с использованием конвейеров редактирования на основе диффузии, с подробными аннотациями, включая порядок редактирования, текстовые инструкции, маски манипуляций и модели генерации. Эти метаданные позволяют проводить поэтапный анализ доказательств и поддерживают обнаружение подделок и прогнозирование последовательностей. Результаты показывают, что высокочастотные сигналы, в частности вейвлет-компоненты, обеспечивают эффективные подсказки даже при ухудшении качества изображения. В целом, SEED облегчает систематическое изучение последовательного отслеживания происхождения и агрегирования доказательств для достоверного анализа визуального контента, созданного ИИ [19].

Диверсификация ответов LLM. Если вы хотите понять, как общественность отреагирует на ваши предложения, большие языковые модели могут имитировать пользователей, отвечающих на вопросы о

возможностях, функциях, акциях или ценах. Однако большие языковые модели не реагируют с таким же разнообразием, как люди. Исследователи разработали метод, который побуждает большие языковые модели принимать облик персон с настраиваемым набором взглядов. Давиде Пальери, Логан Кросс и их коллеги из Google предложили генераторы персон [20]. Их подход создает код, который побуждает большую языковую модель составлять подсказки для 25 персон, охватывающих карту. Задача заставить большую языковую модель принять облик человека обычно сводится к составлению эффективной подсказки (например, «Ответьте на следующий вопрос так, как если бы в современной политике вы считали себя демократом…»). Однако такой подход, как правило, приводит к получению усредненных ответов, которые не отражают диапазон, характерный для человеческой популяции, — даже если запрос явно указывает модели LLM на необходимость учета определенных демографических характеристик. Альтернативный вариант — программно изменять запросы для описания персон до тех пор, пока они не будут выдавать результаты, охватывающие определенный диапазон мнений, взглядов или проблем. При наличии руководящих принципов, определяющих область охвата популяции персон (в частности, взгляды, ранжированные по степени согласия и несогласия), эволюционный алгоритм может подтолкнуть модель к созданию набора запросов, которые вызовут полный диапазон ответов.

Состязательный лицевой камуфляж. Хотя стремительное развитие алгоритмов распознавания лиц позволило реализовать множество полезных приложений, их широкое распространение вызвало серьезные опасения по поводу рисков массового наблюдения и угроз конфиденциальности личности. В статье [21] представлен Adversarial Camouflage как новое решение для защиты конфиденциальности пользователей. Этот подход разработан таким образом, чтобы быть эффективным и простым для воспроизведения пользователями в физическом мире. Алгоритм начинается с определения низкоразмерного пространства шаблонов, параметризованного цветом, формой и углом. Найденные оптимизированные шаблоны проецируются на семантически корректные области лица для оценки. Предложенный метод максимизирует ошибку распознавания в различных архитектурах, обеспечивая высокую переносимость между моделями даже в системах типа «черный ящик». Он значительно ухудшает производительность всех протестированных современных моделей распознавания лиц во время моделирования и демонстрирует многообещающие результаты в реальных экспериментах с участием людей, одновременно выявляя различия в устойчивости моделей и доказательства переносимости атак между архитектурами. Получается эффективно, но весьма заметно.

Больше анонсов интересных публикаций можно найти в блоге Абаванет[27].

## 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 цифровое направление в журнале [22,23].

## Библиография

[1] Егоров, М. Э., et al. "Объяснения моделей машинного обучения и состязательные атаки." International Journal of Open Information Technologies 13.9 (2025): 50-59.

[2] Евграфов, Владимир Андреевич, Маратович Нутфуллин 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] Maloyan, Narek, and Dmitry Namiot. "Adversarial attacks on llm-as-a-judge systems: Insights from prompt injections." arXiv preprint arXiv:2504.18333 (2025).

[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, Дмитрий Евгеньевич, Алексей Александрович Пичугов, and Андрей Павлович Мякишев. "Кибератаки на зарядные станции." International Journal of Open Information Technologies 13.6 (2025): 147-160.

[6] Намиот, Д. Е., Е. А. Ильюшин, and И. В. Чижов. "Текущие академические и индустриальные проекты, посвященные устойчивому машинному обучению." International Journal of Open Information Technologies 9.10 (2021): 35-46.

[7] 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.

[8] Намиот, Д. Е., and Е. А. Ильюшин. "О киберрисках генеративного искусственного интеллекта." International Journal of Open Information Technologies 12.10 (2024): 109-119.

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

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

[11] Namiot, Dmitry. "Artificial Intelligence in Cybersecurity. Chronicle. Issue 7." International Journal of Open Information Technologies 14.5 (2026): 43-56.

[12] Kim, Juhee, et al. "SoK: Attack and Defense Landscape of Agentic AI Systems." 35nd USENIX Security Symposium (USENIX Security 26). 2026.

[13] Li, Yiming, et al. "Rethinking the Trigger of Backdoor Attacks: Towards Physical Backdoor Threats." Pattern Recognition (2026): 113665.

[14] D. Namiot, "On the AI Agents Audit Model," 2026 International Russian Smart Industry Conference (SmartIndustryCon), Sochi, Russian Federation, 2026, pp. 404-409, doi: 10.1109/SmartIndustryCon68821.2026.11493110

[15] NIST SP 800-234 High-Performance Computing (HPC) Security Overlay https://csrc.nist.gov/pubs/sp/800/234/final Retrieved: May, 2026

[16] Acharya, Nirajan, and Gaurav Kumar Gupta. "A Formal Security Framework for MCP-Based AI Agents: Threat Taxonomy, Verification Models, and Defense Mechanisms." arXiv preprint arXiv:2604.05969 (2026).

[17] Kuzmenko, Ilya Dmitrievich, Dmitry Evgenyevich Namiot, and Valery Alexandrovich Vasenin. "Методы обнаружения дипфейков

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

в видеоконференциях в реальном времени." Современные информационные технологии и ИТ-образование 21.2 (2025): 204-220.
[18] Zhang, Haiyue, Yi Nian, and Yue Zhao. "Agent audit: A security analysis system for LLM agent applications." arXiv preprint arXiv:2603.22853 (2026).
[19] Hoi, Mengieong, et al. "SEED: A Large-Scale Benchmark for Provenance Tracing in Sequential Deepfake Facial Edits." arXiv preprint arXiv:2604.10522 (2026).
[20] Paglieri, Davide, et al. "Persona Generators: Generating Diverse Synthetic Personas at Scale." arXiv preprint arXiv:2602.03545 (2026).
[21] Borsukiewicz, Paweł, et al. "Adversarial Camouflage." arXiv preprint arXiv:2603.21867 (2026).
[22] Куприяновский, В. П. Демистификация цифровой экономики / В. Цифровая железная дорога - прогнозы, инновации, проекты / В. П. Куприяновский, Г. В. Суконников, П. М. Бубнов [и др.] // International Journal of Open Information Technologies. – 2016. – Т. 4, № 9. – С. 34-43. – EDN WIQHXX.
[23] Цифровая железная дорога - прогнозы, инновации, проекты / В. П. Куприяновский, Г. В. Суконников, П. М. Бубнов [и др.] // International Journal of Open Information Technologies. – 2016. – Т. 4, № 9. – С. 34-43. – EDN WIQHXX.

Статья получена 20 мая 2026.
Д.Е. Намиот – МГУ имени М.В. Ломоносова (e-mail: dnamiot@cs.msu.ru).

# Artificial Intelligence in Cybersecurity. Chronicle. Issue 8

Dmitry Namiot

***Abstract* - This publication opens the eighth edition of a periodic analytical review on the use of Artificial Intelligence (AI) in cybersecurity. This series of materials aims to provide an in-depth study of the rapidly evolving field emerging at the intersection of artificial intelligence and cybersecurity. The key goal of this project is to systematically monitor global trends and summarize the most notable developments. In addition to collecting information, the initiative provides a thorough analysis of legislative initiatives, high-profile incidents, and cutting-edge technological innovations that are shaping the contours of modern cybersecurity under the influence of AI.**

**Each issue of the series has a standardized structure consisting of three sections, ensuring comprehensive coverage of the topic under consideration. The first section focuses on an analysis of the incident database and existing security challenges: it examines real-world attack scenarios, identifies new vulnerabilities, and assesses the threats posed by the introduction of AI algorithms into both defense mechanisms and attacker arsenals. The second section describes the current state of the regulatory environment and the vectors of change. Understanding these processes is of paramount importance, as they define the legal and operational framework within which reliable and secure AI-based systems will need to develop. The third section chronicles scientific and technological advances. Each issue includes an annotated list of the most significant scientific papers—as identified by the authors—expert reports from leading organizations, and descriptions of innovative developments.**

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

## REFERENCES

[1] Egorov, M. Je., et al. "Ob"jasnenija modelej mashinnogo obuchenija i sostjazatel'nye ataki." International Journal of Open Information Technologies 13.9 (2025): 50-59.

[2] Evgrafov, Vladimir Andreevich, Maratovich Nutfullin 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] Maloyan, Narek, and Dmitry Namiot. "Adversarial attacks on llm-as-a-judge systems: Insights from prompt injections." arXiv preprint arXiv:2504.18333 (2025).

[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, Dmitrij Evgen'evich, Aleksej Aleksandrovich Pichugov, and Andrej Pavlovich Mjakishev. "Kiberataki na zarjadnye stancii." International Journal of Open Information Technologies 13.6 (2025): 147-160.

[6] Namiot, D. E., E. A. Il'jushin, and I. V. Chizhov. "Tekushhie akademicheskie i industrial'nye proekty, posvjashhennye ustojchivomu mashinnomu obucheniju." International Journal of Open Information Technologies 9.10 (2021): 35-46.

[7] 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.

[8] Namiot, D. E., and E. A. Il'jushin. "O kiberriskah generativnogo iskusstvennogo intellekta." International Journal of Open Information Technologies 12.10 (2024): 109-119.

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

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

[11] Namiot, Dmitry. "Artificial Intelligence in Cybersecurity. Chronicle. Issue 7." International Journal of Open Information Technologies 14.5 (2026): 43-56.

[12] Kim, Juhee, et al. "SoK: Attack and Defense Landscape of Agentic AI Systems." 35nd USENIX Security Symposium (USENIX Security 26). 2026.

[13] Li, Yiming, et al. "Rethinking the Trigger of Backdoor Attacks: Towards Physical Backdoor Threats." Pattern Recognition (2026): 113665.

[14] D. Namiot, "On the AI Agents Audit Model," 2026 International Russian Smart Industry Conference (SmartIndustryCon), Sochi, Russian Federation, 2026, pp. 404-409, doi: 10.1109/SmartIndustryCon68821.2026.11493110

[15] NIST SP 800-234 High-Performance Computing (HPC) Security Overlay https://csrc.nist.gov/pubs/sp/800/234/final Retrieved: May, 2026

[16] Acharya, Nirajan, and Gaurav Kumar Gupta. "A Formal Security Framework for MCP-Based AI Agents: Threat Taxonomy, Verification Models, and Defense Mechanisms." arXiv preprint arXiv:2604.05969 (2026).

[17] Kuzmenko, Ilya Dmitrievich, Dmitry Evgenyevich Namiot, and Valery Alexandrovich Vasenin. "Metody obnaruzhenija dipfejkov v videokonferencijah v real'nom vremeni." Sovremennye informacionnye tehnologii i IT-obrazovanie 21.2 (2025): 204-220.

[18] Zhang, Haiyue, Yi Nian, and Yue Zhao. "Agent audit: A security analysis system for LLM agent applications." arXiv preprint arXiv:2603.22853 (2026).

[19] Hoi, Mengieong, et al. "SEED: A Large-Scale Benchmark for Provenance Tracing in Sequential Deepfake Facial Edits." arXiv preprint arXiv:2604.10522 (2026).

[20] Paglieri, Davide, et al. "Persona Generators: Generating Diverse Synthetic Personas at Scale." arXiv preprint arXiv:2602.03545 (2026).

[21] Borsukiewicz, Paweł, et al. "Adversarial Camouflage." arXiv preprint arXiv:2603.21867 (2026).

[22] Kuprijanovskij, V. P. Demistifikacija cifrovoj jekonomiki / V. Cifrovaja zheleznaja doroga - prognozy, innovacii, proekty / V. P. Kuprijanovskij, G. V. Sukonnikov, P. M. Bubnov [i dr.] // International Journal of Open Information Technologies. – 2016. – T. 4, # 9. – S. 34-43. – EDN WIQHXX.

[23] Cifrovaja zheleznaja doroga - prognozy, innovacii, proekty / V. P. Kuprijanovskij, G. V. Sukonnikov, P. M. Bubnov [i dr.] // International Journal of Open Information Technologies. – 2016. – T. 4, # 9. – S. 34-43. – EDN WIQHXX.