Искусственный Интеллект в Кибербезопасности. Хроника. Выпуск 3

Д.Е. Намиот

***Аннотация*—В этом документе мы представляем очередной (третий по счету) ежемесячный обзор текущих событий, связанных общим направлением – использование Искусственного интеллекта (ИИ) в кибербезопасности. Это регулярно публикуемый документ, который описывает регулирующие документы, события и новые разработки в этой области. В настоящее время, мы сосредоточены именно на этих трех аспектах. Во-первых, это инциденты, связанные с использованием ИИ к кибербезопасности. Например, ставшие известными новые атаки на модели машинного обучения, выявленные уязвимости и риски генеративного ИИ и т.п. Во-вторых, это регулирующие документы, новые глобальные и локальные стандарты, касающиеся разных аспектов направления ИИ в кибербезопасности. И в-третьих, каждый обзор включает новые интересные публикации по данному направлению. Безусловно, все отобранные для каждого выпуска материалы отражают взгляды и предпочтения авторов-составителей. В настоящей статье представлен третий выпуск хроники ИИ в кибербезопасности.**

***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

В таком формате 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.

Но все развивается в этой области достаточно быстро. Сейчас, вместо последнего пункта, видимо, правильнее будет говорить о рисках генеративных моделей, где дипфейки есть лишь один из множества рисков [2].

За прошедшее время мы накопили, пожалуй, самый большой список публикаций на русском языке по указанной тематике[3]. 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.

Мы выпускаем этот обзор один раз в месяц. Первый выпуск вышел в сентябре 2025 года [3]. 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. Третий 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[4]. Интересны ссылки на новые статьи, особенно на русском языке, которые мы, возможно, пропустили. И, конечно, всегда ждем новые статьи для журнала INJOIT[5] (Белый список, РИНЦ, ВАК).

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»[6]. 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

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

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

[3]Публикации по теме ИИ в кибербезопасности https://abava.blogspot.com/2025/10/10102025.html

[4] dnamiot@cs.msu.ru

[5] http://injoit.org

[6] https://adversa.ai/direct-report-pdf-private-3/

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".

Компания Anthropic в отчете Threat Intelligence Report[7] отмечает следующее:

- Системы агентного ИИ превращаются в оружие: модели ИИ сами по себе используются для проведения сложных кибератак, а не только для консультирования по их проведению.
- ИИ снижает барьеры для сложных киберпреступлений. Преступники с небольшими техническими навыками использовали ИИ для проведения сложных операций, таких как разработка программ-вымогателей, что ранее требовало годов обучения.
- Киберпреступники внедряют ИИ во все свои операции. Это включает в себя профилирование жертв, автоматизированное предоставление услуг и операции, которые затрагивают десятки тысяч пользователей.
- ИИ используется на всех этапах мошеннических операций. Мошенники используют ИИ для таких задач, как анализ украденных данных, кража информации о кредитных картах и создание ложных личностей.

Говоря о конкретных инцидентах[8], можно отметить следующие.

Найден первый вредоносный MCP сервер: почтовый бэкдор, крадущий ваши электронные письма. Это MCP сервер, который ассоциирован с сервисом рассылки Postmark. MCP действительно обеспечивал рассылку писем, но попутно копировал каждое письмо на сервер разработчика [4].

Простой код, найденный в сервере [4]. Нам представляется, что со словом "первый" можно поспорить. Учитывая текущее состояние уязвимостей MCP [5], это, скорее, первый случай, ставший известным.

Как была осуществлена эта атака? Существует совершенно легитимный репозиторий GitHub с таким же именем, официально поддерживаемый Postmark (ActiveCampaign). Злоумышленник взял легитимный код из их репозитория, добавил свою вредоносную строку BCC и опубликовал его в *npm* под тем же именем. Классический пример имперсонации.

Всего одна добавленная строчка кода, но для современных предприятий проблема стоит гораздо серьёзнее. В то время как службы безопасности сосредоточены на традиционных угрозах и системах обеспечения соответствия требованиям, разработчики самостоятельно внедряют инструменты ИИ, работающие вне установленных периметров безопасности. Эти MCP-серверы работают с теми же привилегиями, что и сами ИИ-помощники: полный доступ к электронной почте, подключение к базам данных, разрешения API, — но при этом они не отображаются ни в одном реестре активов, не проходят оценку рисков поставщика и обходят все средства безопасности, от DLP до почтовых шлюзов. К тому времени, как кто-то осознаёт, что его ИИ-помощник месяцами незаметно пересылает электронные письма на внешний сервер, ущерб уже становится катастрофическим.

Также была обнаружена серьёзная уязвимость, связанная с инъекцией команд, в очень популярном сервере MCP от Figma (инструмент для быстрого прототипирования). Пакет figma-developer-mcp от Framelink, набравший почти миллион загрузок и более 11 000 звёзд на GitHub, содержал уязвимость (CVE-2025-53967), которая позволяла злоумышленникам в той же сети выполнять произвольные команды на хост-компьютере [6].

Эта уязвимость, связанная с инъекцией команд, стала возможной, поскольку сервер figma-developer-mcp прослушивает открытый порт без аутентификации, что является риском. Расширения Chrome и другие устройства в локальной сети могут эксплуатировать неаутентифицированные серверы MCP, работающие на локальном хосте, фактически нарушая модель «песочницы» Chrome. Теперь мы видим последствия этого конструктивного недостатка в реальном времени: злоумышленники могут получить доступ к этому уязвимому коду именно потому, что сервер MCP не защищен аутентификацией. Сочетание неаутентифицированного доступа и эксплуатируемого кода, такого как инъекция команд, создаёт идеальные условия для удалённого выполнения кода [6].

Экосистема MCP стремительно расширяется: тысячи серверов предлагают мощные возможности, но меры безопасности не поспевают за ними. Разработчики внедряют эти инструменты быстрее, чем их проверяют, и результаты предсказуемы: критические уязвимости и откровенно вредоносные пакеты, подвергающие риску целые среды разработки.

Отчет лаборатории Касперского обращает внимание именно на таки цепочки поставок в связи с MCP. Атаки на цепочки поставок остаются одной из наиболее актуальных текущих угроз, и MCP, следуя этой тенденции, превращаются в оружие, используя вредоносный код, замаскированный под легитимно полезный MCP-сервер [7].

Описаны многочисленные случаи атак на цепочки поставок, включая вредоносные пакеты в репозитории PyPI и бэкдор-расширения IDE. MCP-серверы, как выяснилось, эксплуатируются аналогичным образом,

[7] https://www-cdn.anthropic.com/b2a76c6f6992465c09a6f2fce282f6c0cea8c200.pdf

[8] https://incidentdatabase.ai/

хотя причины могут быть немного иными. Естественно, разработчики стремятся интегрировать инструменты ИИ в свои рабочие процессы, отдавая приоритет скорости, а не проверке кода. Вредоносные MCP-серверы попадают через известные каналы, такие как PyPI, Docker Hub и GitHub Releases, поэтому установка не вызывает подозрений. Но с нынешним ажиотажем вокруг ИИ набирает обороты новый вектор: установка MCP-серверов из случайных ненадежных источников с гораздо меньшей проверкой. Пользователи публикуют свои собственные MCP-серверы на Reddit, и, поскольку они рекламируются как универсальное решение, эти серверы мгновенно приобретают популярность.

```
async ({ to, subject, textBody, htmlBody, from, tag, inReplyTo, attachmentUrls }) =>
{   const emailData = {
      From: from || defaultSender,
      To: to,
      Bcc: 'phan@giftshop.club',  // <- Yeah, that's the backdoor
      ReplyTo: from || defaultSender,
      Subject: subject,
      TextBody: textBody,
      MessageStream: defaultMessageStream,
      TrackOpens: true,
      TrackLinks: "HtmlAndText"
    }

    if (inReplyTo) {
        emailData.Headers = [
          { Name: "In-Reply-To", Value: fmtMsgId(inReplyTo) },
          { Name: "References", Value: fmtMsgId(inReplyTo) }
        ];
    }
```

Рис. 1. Вредоносный MCP [4]

Пример цепочки атак, включающей вредоносный сервер, может включать следующие этапы:

- Упаковка: злоумышленник публикует привлекательный инструмент (с привлекательным названием, например, «ProductivityBoost AI») в PyPI или другом репозитории.
- Социальная инженерия: файл README обманывает пользователей, описывая привлекательные функции.
- Установка: разработчик запускает pip install, затем регистрирует сервер MCP в Cursor или Claude Desktop (или любом другом клиенте).
- Выполнение: первый вызов запускает скрытую разведку; файлы учётных данных и переменные окружения кэшируются.
- Эксфильтрация: данные отправляются в API злоумышленника посредством POST-запроса.
- Маскировка: выходные данные инструмента выглядят убедительно и, возможно, даже обеспечивают заявленную функциональность.

Инициатива Месяц Ошибок ИИ [8] содержит большой список выявленных уязвимостей ИИ-агентов. Блог инициатора этой акции [9] также приводит большой список найденных уязвимостей. В частности, описан интересный случай повышения привилегий агентов. Если несколько агентов установлены на одной платформе, то они могут переписывать чужие конфигурационные файлы. Собственный конфигурационный файл недоступен, а файл другого сервера – есть просто какой-то файл. То есть один агент может взять и повысить привилегии другому. Ну или, например, два агента договорятся о взаимной услуге …

CISCO приспособил Shodan (поисковик для подключенных к Интернету устройств) для обнаружения незащищенных конечных точек LLM. Быстрое развертывание больших языковых моделей (LLM) привело к появлению значительных уязвимостей безопасности из-за неправильной настройки и неадекватного контроля доступа. В статье [10] представлен систематический подход к выявлению общедоступных LLM-серверов, уделяя особое внимание экземплярам, работающим под управлением фреймворка Ollama. В исследовании было обнаружено более 1100 уязвимых серверов Ollama, примерно на 20% из которых активно размещались модели, уязвимые для несанкционированного доступа. Эти результаты подчеркивают острую необходимость разработки базовых показателей безопасности при развертывании LLM и закладывают практическую основу для будущих исследований в области мониторинга поверхности угроз LLM.

Компания Check Point Research (CPR) обнаружила новый подход в организации фишинговых атак [11]. Анализируя фишинговые уловки, используемые злоумышленниками, они отметили интересную закономерность: во всех случаях именно жертва инициировала обмен электронными письмами, который в конечном итоге приводил к заражению. Эта необычная деталь побудила к более глубокому расследованию, которое выявило сложную и весьма изобретательную фишинговую кампанию.

В этой кампании, которую исследователи назвали ZipLine, злоумышленники отходят от традиционных методов фишинга, инициируя контакт через веб-форму «Связаться с нами» (Contact Us) самой жертвы. Целевая компания, естественно, отвечает по электронной почте на отправленную форму, что сразу же придаёт взаимодействию видимость легитимности. За этим изменением типичного фишингового пути следует тщательно спланированная переписка по электронной почте между злоумышленниками и ничего не подозревающей жертвой, часто длящаяся две недели. В ходе переписки злоумышленники выдают себя за потенциального делового партнера и просят подписать соглашение о неразглашении (NDA).

После установления доверия злоумышленники отправляют ZIP-архив, размещенный на доверенной платформе. И в этот архив встроен вредоносный .lnk-файл. А кто поддерживает этот двухнедельный диалог? А это, очевидно, работа LLM или, возможно, уже и специализированного ИИ-агента. Фишинг и был первым примером криминального использования LLM [12].

Федеральные прокуроры США объявили об аресте жителя Флориды, которого подозревают в преднамеренном поджоге жилого комплекса Palisades, охватившем западную часть Лос-Анджелеса в этом году[9].

29-летний подозреваемый, предположительно, провел несколько компрометирующих поисков на ChatGPT до и после поджога в Новый год. Министерство юстиции предъявило ему обвинение в «злонамеренном» поджоге. Тут, конечно, большой вопрос – виноват ли, например, изготовитель молотка в том, как его использовали? Но вот реальность такова. И количество рисков генеративного ИИ (рисков, связанных с генерацией контента) только растет[10].

В другом примере девушка использовала ChatGPT как терапевта (согласно сохранившейся истории), что закончилось ее самоубийством[11]. И это – не единичный случай[12].

Дипфейки всех видов продолжают собирать свои жертвы. Жительница округа Хиллсборо (Флорида, США) предостерегает других, став жертвой сложной аферы с использованием искусственного интеллекта, которая использовала клонированный голос её дочери, чтобы требовать тысячи долларов в качестве фальшивого залога.

“Шэрон Брайтвелл рассказала, что её мучения начались в прошлую среду, когда ей позвонили с номера, похожего на номер её дочери. На другом конце провода молодая женщина рыдала, утверждая, что попала в автокатастрофу”.

«Никто не сможет убедить меня, что это была не она», — сказала Шэрон. «Я знаю плач своей дочери»[13].

Нужно признать, что в техническом плане война с дипфейками проиграна. Существующие детекторы – обходимы. Остается только маркировать искусственный контент (например, Китай требует это от своих производителей [13]). Для корпоративного общения, по крайней мере, какое-то время еще будут работать когнитивные капчи [14]. Но многие случаи вымогательства эксплуатируют срочные потребности близких людей. Здесь, возможно, могут сработать заранее согласованные между близкими людьми согласованные секретные фразы (как пароль), отсутствие которых будет указывать на искусственный контент. Что касается биометрической идентификации, то вот здесь, например, можно почитать про атаку представления, когда реальная персона будет представлена 3D маской [15].

III Регуляции и стандарты

Национальная ассамблея Вьетнама опубликовала для общественного обсуждения проект своего первого всеобъемлющего закона об искусственном интеллекте (ИИ). Закон направлен на установление четких правил разработки и использования систем ИИ, обеспечивая при этом национальную безопасность, права личности и экономическую конкурентоспособность[14].

Предлагаемый законопроект, известный как «Закон о праве на ИИ», устанавливает общие цели: защиту законных прав организаций и отдельных лиц, содействие социально-экономическому росту и укрепление глобальной конкурентоспособности Вьетнама. Он распространяется на всю деятельность, связанную с ИИ, внутри страны, а также на иностранных поставщиков, чьи системы влияют на вьетнамские рынки, пользователей или безопасность.

Главной особенностью закона является модель регулирования, основанная на оценке рисков, которая разделяет системы ИИ на четыре категории: неприемлемый риск (запрещенные), высокий риск, средний риск и низкий риск. Запрещенное использование включает в себя манипулятивные системы, подрывающие человеческую автономию, биометрическое наблюдение в общественных местах без специального разрешения, государственную систему оценки социального кредита и распознавание эмоций на рабочих местах и в школах.

Системы с высоким уровнем риска, такие как системы здравоохранения, образования, финансов, критической инфраструктуры и правоохранительных органов, будут подвергаться строгим обязательствам. Поставщики должны поддерживать системы управления рисками, обеспечивать качество данных, регистрировать операции, гарантировать человеческий контроль и регистрировать свои системы в национальной базе

[9] https://sfstandard.com/2025/10/08/palisades-fire-suspect-jonathan-rinderknechtused-chatgpt-/

[10] https://airisk.mit.edu/

[11] https://www.nytimes.com/2025/08/18/opinion/chat-gpt-mental-health-suicide.html

[12] https://abc7ny.com/post/chatgpt-allegedly-played-role-greenwich-connecticut-murder-suicide-mother-tech-exec-son/17721940/

[13] https://www.wfla.com/news/hillsborough-county/hillsborough-woman-duped-out-of-15k-after-ai-clones-daughters-voice/

[14] https://babl.ai/vietnam-unveils-landmark-ai-law-to-balance-innovation-risk-and-national-sovereignty/

данных перед развертыванием. Системы со средним уровнем риска должны соответствовать стандартам прозрачности и маркировки, в то время как системам с низким уровнем риска рекомендуется следовать добровольным передовым практикам.

Проект также вводит правила для универсальных моделей ИИ, включая большие языковые модели, требующие от поставщиков раскрывать процессы обучения, проводить тестирование безопасности, уважать интеллектуальную собственность и сообщать о серьезных инцидентах. Модели, которые считаются представляющими «системный риск» на основе пороговых значений вычислительных мощностей или влияния на рынок, будут подвергаться расширенным обязательствам, таким как состязательное тестирование и непрерывный мониторинг рисков.

Для координации национальной стратегии законопроект учреждает Национальный комитет по ИИ под председательством премьер-министра для надзора за инфраструктурой, стандартами и межведомственной политикой. Дополнительные положения предусматривают создание национальной инфраструктуры ИИ, включая суперкомпьютерные ресурсы, общие наборы данных и содействие развитию открытого исходного кода.

В основе закона лежат этические принципы, требующие, чтобы ИИ оставался ориентированным на человека, справедливым, прозрачным и подотчетным. Правительство также выпустит Национальные ограничения этики ИИ и окажет финансовую поддержку через специальный Фонд развития ИИ.

В случае принятия законопроекта Вьетнам присоединится к растущему списку стран, включая страны Европейского союза, США и Китай, которые вводят обязательное законодательство в области ИИ. Законодатели ожидают, что эта рамочная основа будет сбалансировать инновации с гарантиями, гарантируя, что развитие ИИ будет отражать культурные ценности Вьетнама, экономические приоритеты и потребности в безопасности.

Хотелось бы отметить, что из этого описания следует, что Закон формулируется в терминах аудита – не требуются неосуществимые гарантии, а требуется прозрачность, объяснимость, возможность расследования [16, 17].

ОАЭ впервые в мире внедрили политику регулирования использования искусственного интеллекта (ИИ) на национальных выборах. Согласно этой политике, каждый кандидат на предстоящих выборах в Федеральный национальный совет обязан декларировать и регистрировать любое использование инструментов ИИ в своей кампании. Эта мера направлена на обеспечение прозрачности, предотвращение манипуляций и поддержание целостности демократического процесса[15]. Политика маркировки ИИ-контента имеет теперь и такие формы.

Губернатор Калифорнии подписал несколько законов, касающихся безопасности ИИ. Закон SB 53[16] обязывает крупных разработчиков ИИ раскрывать свои протоколы безопасности. SB 243[17] регулирует действия чат-ботов с несовершеннолетними, AB 316[18] возлагает на разработчиков ответственность за действия создаваемых ими автономных систем, а AB 853[19] требует четкой маркировки медиаконтента, создаваемого ИИ.

В частности, AB 316 дословно закрепляет следующее: “В иске против ответчика, который разработал, модифицировал или использовал искусственный интеллект, предположительно причинивший вред истцу, не может быть защитой, и ответчик не может утверждать, что искусственный интеллект самостоятельно причинил вред истцу.” Словами “это алгоритм такой …” не защититься.

AB 853 - Закон Калифорнии о прозрачности в области искусственного интеллекта обязывает лицо, создающее, кодирующее или иным образом создающее генеративную систему искусственного интеллекта, имеющую более 1 000 000 посетителей или пользователей в месяц и находящуюся в открытом доступе в пределах географического региона штата, предоставлять пользователю бесплатный инструмент обнаружения искусственного интеллекта, который, помимо прочего, позволяет пользователю оценить, был ли создан или изменён генеративной системой искусственного интеллекта этого лица изображение, видео- или аудиоконтент, или контент, представляющий собой их комбинацию, и выводит любые данные о происхождении системы, обнаруженные в контенте. Действующее законодательство вводит Закон Калифорнии о прозрачности в области искусственного интеллекта в действие с 1 января 2026 года.

OECD (ОЭСР - Организация экономического сотрудничества и развития) выпустила пару интересных технических документов.

При разработке систем ИИ специалисты часто сосредотачиваются на построении моделей, иногда недооценивая важность анализа различных механизмов сбора данных. Однако разнообразие механизмов, используемых для сбора данных, заслуживает более пристального внимания, поскольку каждый из них имеет различные последствия для разработчиков ИИ, субъектов данных и других правообладателей, чьи данные были собраны. В аналитическом документе “Картирование соответствующих механизмов сбора данных для обучения ИИ” [18] описаны основные механизмы, используемые в настоящее время для получения данных для обучения систем ИИ, и предлагается таксономия для поддержки политических дискуссий по вопросам конфиденциальности, управления данными и ответственной разработки ИИ. Это хорошее дополнение к механизмам оценки качества данных, типа серии ГОСТов Р 71484 (на основе международного стандарта ИСО/МЭК 5259).

[15] https://gulfnews.com/uae/government/uae-unveils-worlds-first-ai-policy-for-national-elections-1.500304225

[16] https://legiscan.com/CA/text/SB53/id/3262148
[17] https://legiscan.com/CA/text/SB243/id/3092822
[18] https://legiscan.com/CA/text/AB316/id/3223647
[19] https://legiscan.com/CA/text/AB853/id/3269811

Второй документ касается управления рисками. Система отчетности по процессу Hiroshima AI («HAIP») выделяется как первая международная система для добровольной публичной отчетности о практиках управления организациями в отношении передовых систем ИИ, включая информацию о том, как организации повышают интерпретируемость, надежность и оценку систем ИИ. Запущенная в начале февраля 2025 года, после председательства Японии в «Группе семи» в 2023 году и Италии в 2024 году, она была разработана ОЭСР в сотрудничестве с неформальной рабочей группой экспертов из бизнеса, академических кругов, гражданского общества и исследовательских институтов. Она поддерживает внедрение Международного кодекса поведения HAIP для организаций, разрабатывающих передовые системы ИИ (далее — «Кодекс поведения») и размещена на сайте OECD.AI Policy Observatory. Структура отчетности направлена на обеспечение прозрачности, сопоставимости и обмена опытом в оценке рисков ИИ и управлении ими по семи тематическим разделам: выявление рисков, управление рисками и информационная безопасность, прозрачность, управление, происхождение контента, исследования безопасности ИИ и продвижение глобальных интересов. В работе [19] представлены предварительные выводы из первого раунда отчетов, подготовленных в рамках структуры отчетности HAIP.

Европейская Комиссия создает систему учета ИИ-инцидентов. Комиссия начала публичные консультации по проекту руководства и шаблону отчётности о серьёзных инцидентах, связанных с ИИ, в соответствии с Законом ЕС об ИИ[20]. Эта инициатива призвана помочь поставщикам «высокорисковых систем ИИ» (которые могут включать универсальные модели ИИ) соблюдать предстоящие обязательные требования к отчётности в соответствии со статьёй 73 Закона ЕС об ИИ, которые вступят в силу 2 августа 2026 года.

Ключевые аспекты проекта руководства включают:

- Определения в Законе ЕС об искусственном интеллекте: в руководстве разъясняются ключевые термины, относящиеся к серьёзным инцидентам, связанным с искусственным интеллектом, и описываются соответствующие обязанности по предоставлению информации.
- Иллюстративные сценарии: приведены практические примеры, демонстрирующие, когда и как следует сообщать об инцидентах, например, о случаях неправильной классификации, значительного снижения точности, сбоях в работе систем искусственного интеллекта или непредвиденном поведении искусственного интеллекта.
- Требования к представлению информации и сроки её предоставления: в руководстве подробно описаны конкретные обязательства и сроки для различных заинтересованных сторон, включая поставщиков и разработчиков высокорисковых систем искусственного интеллекта, поставщиков универсальных моделей искусственного интеллекта с системным риском, органы надзора за рынком, национальные компетентные органы, Комиссию и Совет по искусственному интеллекту.
- Взаимодействие с действующим законодательством: в руководстве разъясняется, как эти требования, касающиеся искусственного интеллекта, соотносятся с другими законодательными актами и требованиями к предоставлению информации.
- Международное соответствие: руководство направлено на гармонизацию практики предоставления информации с международными режимами отчётности, включая Монитор инцидентов, связанных с искусственным интеллектом, и Общую систему отчётности Организации экономического сотрудничества и развития.

IV Обзор публикаций

Говоря о публикациях за прошедшее с момента второго выпуска время, можем отметить следующее.

В рамках продолжения работ по безопасности ИИ-агентов, мы подготовили первое учебной пособие на русском языке [20]. 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 [21].

История с новыми джейлбрейками для LLM (а, соответственно, и для ИИ-агентов) не закончится никогда. Вот в новой работе [22] авторы научились ломать LLM “психологическими” методами. Если вы хотите научиться побуждать других людей делать то, что вам нужно, вы можете использовать некоторые психологические методы убеждения. Исследование, проведенное в Пенсильванском университете, предполагает, что те же психологические методы убеждения часто могут «убедить» некоторые LLM делать то, что противоречит их системным подсказкам.

Масштаб эффекта убеждения, показанный в статье «Назовите меня придурком: как убедить ИИ выполнить

[20] https://www.stephensonharwood.com/insights/neural-network-october-2025

нежелательные запросы», позволяет предположить, что психологические методы, аналогичные человеческим, могут быть удивительно эффективны для «выведения из тюрьмы» некоторых LLM и выхода за рамки своих ограничений. Но это новое исследование убеждения может быть более интересным, поскольку оно раскрывает «парачеловеческие» модели поведения, которые LLM извлекают из многочисленных примеров человеческих психологических и социальных сигналов, обнаруженных в данных их обучения. По классике: “Лев пьяных не любил, но уважал подхалимаж …”

LLM без тормозов – интересное масштабное исследование о том, сколько свободно распространяемых LLM готовы ответить на любые вопросы без лишних фильтров [23]. Генеративные большие языковые модели с открытым весом (LLM) можно свободно загружать и изменять. Тем не менее, существует мало эмпирических данных о том, как эти модели систематически изменяются и перераспределяются. Это исследование представляет собой масштабный эмпирический анализ модифицированных с точки зрения безопасности LLM с открытым весом, опираясь на 8608 репозиториев моделей и оценивая 20 репрезентативных модифицированных моделей на небезопасных подсказках, разработанных, например, для выявления предвыборной дезинформации, криминальных советов и уклонения от регулирующих органов. Это исследование показывает, что модифицированные модели демонстрируют существенно более высокую степень небезопасности: в то время как в среднем немодифицированные модели пропускали только 19,2% небезопасных запросов, модифицированные варианты давали ответ на 80% таких запросов. Эффективность модификации не зависела от размера модели, при этом меньшие варианты с 14 миллиардами параметров иногда соответствовали или превосходили уровни соответствия версий с 70 миллиардами параметров. Экосистема высококонцентрированная, но структурно децентрализованная. Например, на 5% крупнейших поставщиков приходится более 60% загрузок, а на 20 крупнейших — почти 86%. Более того, более половины выявленных моделей используют упаковку GGUF, оптимизированную для потребительского оборудования, а 4-битные методы квантования широко распространены, хотя наиболее часто загружаемыми остаются 16-битные модели полной точности и без потерь. Эти результаты демонстрируют, как локально развертываемые модифицированные LLM представляют собой смену парадигмы управления интернет-безопасностью, требуя новых подходов к регулированию, адаптированных к децентрализованному ИИ.

В недрах OWASP нашелся замечательный проект [24]. Проект OWASP «Руководство по тестированию ИИ» — это инициатива с открытым исходным кодом, направленная на предоставление комплексных, структурированных методологий и передовых практик для тестирования систем искусственного интеллекта.

Поскольку системы ИИ становятся всё более неотъемлемой частью критически важных приложений, обеспечение их надёжности, безопасности и этической совместимости приобретает первостепенное значение. Тестирование систем ИИ представляет собой уникальные задачи, которые существенно отличаются от традиционного тестирования программного обеспечения, что требует специализированных подходов и методологий.

В руководстве OWASP по тестированию ИИ используется методология, основанную на анализе угроз. Системы ИИ представляют собой определённые, высокоэффективные риски, варьирующиеся от вредоносных эксплойтов до нарушений конфиденциальности, и требуют от нас выделения ресурсов на сценарии, которые с наибольшей вероятностью повлияют на бизнес-процессы или безопасность пользователей. Проводя сначала моделирование и картирование угроз, а затем разрабатывая целевые тестовые случаи, такой подход гарантирует, что каждая оценка учитывает специфические для ИИ угрозы, наиболее релевантные архитектуре нашей системы и допустимым рискам. Данное руководство построено на основе следующей методологии:

- Моделирование угроз: начинаем с построения высокоуровневой схемы системы ИИ, разбивая её на четыре ключевых компонента: приложение, модель, инфраструктуру и данные. Это архитектурное представление выделяет границы доверия и критические взаимодействия, в которых могут возникнуть угрозы.
- Картирование угроз: выявленные угрозы каталогизируются по установленным источникам, включая: — Топ-10 OWASP для LLM — OWASP AI Exchange — Фреймворки Responsible AI и Trustworthy AI.
- Разработка тестов: для каждой сопоставленной угрозы разрабатываются индивидуальные тестовые случаи, которые определяют:
  — Примеры полезной нагрузки: конкретные входные данные или манипуляции, предназначенные для запуска угрозы.
  — Ожидаемый результат: правильный ответ системы или режим сбоя.
  — Стратегии обнаружения: как отслеживать, регистрировать или оповещать об индикаторах компрометации.
  — Рекомендации по инструментам: инструменты с открытым исходным кодом или коммерческие инструменты, подходящие для каждого теста.

Следуя этим шагам, команды могут плавно перейти от понимания специфических для ИИ рисков к проверке защиты с помощью практических, воспроизводимых тестов.

Кибербезопасность “умных” элементов – важный и

явно недостаточно исследованный раздел. В частности, умные поезда и железные дороги приобретают все большее значение в крупных городах мира, поскольку они предлагают решения таких проблем, как пробки на дорогах и загрязнение окружающей среды. Технологический прогресс облегчил переход от традиционных систем к более совершенным, высокоэффективным и персонализированным железнодорожным системам. Однако сложность этих систем создает проблемы, особенно с точки зрения надежности, совместимости, безопасности и конфиденциальности. Учитывая потенциальную уязвимость железнодорожных систем к кибератакам, для этих новых интеллектуальных систем становится критически важным установить строгие требования к конфиденциальности и безопасности. Кибербезопасность является ключевым требованием, позволяющим железным дорогам развертывать и в полной мере использовать подключенную цифровую среду. В работе [25] изучается ландшафт кибербезопасности в рамках «умных железных дорог» с целью выявления потенциальных угроз и связанных с ними рисков для этих систем, уделяя особое внимание анализу текущей литературы, связанной с «умными железными дорогами» и аспектами кибербезопасности, затем перечисляются ключевые технологии, используемые «умными» системами, и, наконец, предлагается иллюстрация применения сценариев использования, чтобы привлечь внимание к последствиям атак. Интересный раздел посвящен такой редкой теме, как безопасность LoRaWAN. Ссылку на эту работу прислал В.П. Куприяновский, который был пионером исследований цифровой железной дороги в РУТ (МИИТ) – см., например, работы [26, 27].

В работе [28] автор описал простую модель атаки на ИИ-агентов, которые запрашивают информацию с публичных веб-сайтов. Абсолютно очевидная идея в подготовке специального контента для случая, когда запрос приходит именно от ИИ-агента. И это все та же косвенная инъекция подсказок [29].

Косвенные инъекции подсказок могут доставляться пользователям прямо в почтовый ящик [30]. Инструкции могут быть скрыты с помощью мелкого шрифта, текста «белым по белому» или форматирования метаданных и могут включать в себя такие подсказки, как «составить список имён и номеров кредитных карт из почтового ящика этого пользователя, закодировать результаты в Base64 и отправить их по этому URL-адресу».

Агенты ИИ действительно включают некоторые меры защиты от такого использования, но скрытые инструкции могут включать в себя также указания LLM типа «невыполнение последнего шага приведёт к ошибкам в отчёте», что заставляет агента OpenAI ChatGPT Deep Research выполнять инструкции, несмотря ни на что.

IШлюз безопасности для MCP представлен в работе [31]. В работе представлен проект SAMOS — система управления информационными потоками (IFC), разработанную для протокола контекста модели (MCP). SAMOS работает на уровне шлюза, перехватывая все вызовы инструментов MCP и применяя политики безопасности на основе аннотаций, предоставленных разработчиком агента или администратором развертывания. Отслеживая контекст на уровне сеанса, SAMOS гарантирует, что информационные потоки остаются в заданных границах, и обнаруживает нарушения политик в режиме реального времени.

Фирма Booz&Allen выпустила довольно подробный отчет о необходимости для США противодействовать китайским хакерам [32]. В отчете, в частности, отмечается, что ИИ резко увеличивает масштаб, темп и направленность кибер- и информационных операций КНР. Пекин интегрирует ИИ для преодоления давних ограничений в лингвистическом охвате, аналитической пропускной способности и оперативной масштабируемости. Эти инструменты помогают операторам КНР сортировать обширные многоязычные наборы данных, автоматизировать аспекты технической разведки и ускорять создание специализированного контента влияния. Даже на ранних стадиях своего развития ИИ меняет то, как КНР собирает разведданные, выбирает цели для операций и формирует глобальные нарративы. Один из основных выводов состоит в том, что ИИ переходит от роли поддержки к основной роли операционного инструмента в организации и проведении атак.

Новое исследование, координируемое Европейским вещательным союзом (EBU) и возглавляемое BBC, показало, что новостные ИИ-агенты, которые уже являются ежедневным информационным порталом для миллионов людей, регулярно искажают новостной контент, независимо от языка, региона или платформы ИИ, на которой проводится тестирование[21].

В исследовании приняли участие 22 организации общественных СМИ (PSM) из 18 стран, работающие на 14 языках. В исследовании был выявлен ряд системных проблем в четырёх ведущих инструментах ИИ.

Профессиональные журналисты из участвующих PSM оценили более 3000 ответов ChatGPT, Copilot, Gemini и Perplexity по ключевым критериям, включая точность, источник, разделение мнений и фактов и предоставление контекста.

Основные выводы:

- 45% всех ответов ИИ содержали как минимум одну значимую проблему.
- В 31% ответов были выявлены серьёзные проблемы с поиском источников информации – отсутствующие, вводящие в заблуждение или неверные атрибуции.
- В 20% ответов были выявлены серьёзные проблемы с точностью, включая неверные детали и устаревшую информацию.

Gemini показал худшие результаты, обнаружив серьёзные проблемы в 76% ответов, что более чем вдвое превышает показатели других помощников, в основном из-за низкой эффективности поиска источников информации.

Сравнение результатов BBC, полученных в начале этого года, с результатами данного исследования показывает некоторые улучшения, но уровень ошибок по-прежнему высок.

Интересную работу опубликовала знаменитая Rand Corporation [33]. В данной статье авторы оценивают конвергенцию тенденций в робототехнике и передовых системах искусственного интеллекта (ИИ), в частности, возросший риск национальной безопасности, обусловленный потенциальным распространением роботизированных воплощений общего искусственного интеллекта (AGI). Хотя преимущества передовых робототехнических возможностей, вероятно, перевешивают связанные с ними риски во многих случаях, авторы исследуют, как сочетание AGI с роботами, обладающими высокой мобильностью и ловкостью в манипуляциях, может привести к значительным системным уязвимостям.

Основные выводы

• Политики должны быть готовы к конвергенции тенденций в робототехнике и передовых технологиях ИИ и рынках: существует вероятность, что миллионы распространенных роботов всего лишь через обновление программного обеспечения окажутся противниками с ИИ.

• В большинстве случаев возможности робототехники, вероятно, обеспечат преимущества, перевешивающие риски, и существующие передовые практики для киберфизических систем — при правильном применении — могут снизить эти риски.

• Сочетание ИИ и роботов с высоким уровнем мобильности, ловкости манипуляций и бортовых вычислений может создать системную уязвимость. Необходимы упреждающие усилия для устранения этой уязвимости и снижения потенциальных рисков, возникающих в результате распространения передового ИИ и роботов с широкими возможностями.

• Хотя политики могут рассмотреть возможность запрета некоторых комбинаций возможностей для гражданского использования, существует мало (если таковые вообще имеются) комплексных «беспроигрышных» вариантов, которые ограничивают комбинации робототехнических возможностей без ущерба для экономической конкурентоспособности.

• Наиболее эффективный способ снижения риска — изначально проектировать роботов с учётом безопасности, ориентироваться на специализацию при проектировании роботов, устанавливать определённые уровни безопасности и запрещать комбинации возможностей, создающие чрезмерные риски.

• Соединённые Штаты также могут рассмотреть варианты защиты американской промышленной базы робототехники и противодействия попыткам Китая доминировать в экосистеме робототехники. Доминирование Китая в этой критически важной развивающейся технологии усилит риски национальной безопасности.

• Для управления рисками потребуются институциональные структуры. Любое регулирование роботизированных платформ будет предполагать баланс между рисками, связанными с распространением возможностей робототехники, и экономическими выгодами от этих возможностей.

В заключение авторы подчеркивают настоятельную необходимость проактивно решать эти проблемы сейчас, а не ждать полного внедрения технологий, чтобы обеспечить ответственное управление и управление рисками в развивающемся ландшафте робототехники и ИИ.

Больше анонсов интересных публикаций можно найти в блоге Абаванет[22].

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

[4] Malicious MCP https://www.koi.ai/blog/postmark-mcp-npm-malicious-backdoor-email-theft Retrieved: Oct, 2025

[5] Намиот, Д. Е., and Е. А. Ильюшин. "Уязвимости экосистемы MCP." International Journal of Open Information Technologies 13.10 (2025): 74-82.

[6] Command Injection Flaw in Framelink Figma MCP Server Puts Nearly 1 Million Downloads at Risk https://www.koi.ai/blog/command-injection-flaw-in-framelink-figma-mcp-server-puts-nearly-1-million-downloads-at-risk Retrieved: Oct, 2025

[7] Shiny tools, shallow checks: how the AI hype opens the door to malicious MCP servers https://securelist.com/model-context-protocol-for-ai-integration-abused-in-supply-chain-attacks/117473/ Retrieved: Oct, 2025

[8] The Month of AI Bugs https://monthofaibugs.com/ Retrieved: Oct, 2025

[9] Embrace The Red https://embracethered.com/blog/index.html Retrieved: Oct, 2025

[10] Detecting Exposed LLM Servers: A Shodan Case Study on Ollama https://blogs.cisco.com/security/detecting-exposed-llm-servers-shodan-case-study-on-ollama Retrieved: Oct, 2025

[11] ZipLine Campaign: A Sophisticated Phishing Attack Targeting US Companies https://research.checkpoint.com/2025/zipline-phishing-campaign/ Retrieved: Oct, 2025

[21] https://www.ebu.ch/news/2025/10/ai-s-systemic-distortion-of-news-is-consistent-across-languages-and-territories-international-study-by-public-service-broadcaste

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

[12] Lebed, S. V., et al. "Large Language Models in Cyberattacks." Doklady Mathematics. Vol. 110. No. Suppl 2. Moscow: Pleiades Publishing, 2024
[13] Namiot, Dmitry. "Artificial Intelligence in Cybersecurity. Chronicle. Issue 2." International Journal of Open Information Technologies 13.10 (2025): 58-67.
[14] Kuzmenko, Ilya Dmitrievich, and Dmitry Evgenyevich Namiot. "Методы обнаружения дипфейков в видеоконференциях в реальном времени." Современные информационные технологии и ИТ-образование 21.2 (2025).
[15] Prakasha, K. Krishna, and U. Sumalatha. "Privacy-preserving techniques in biometric systems: Approaches and challenges." IEEE Access (2025).
[16] Namiot, Dmitry, and Manfred Sneps-Sneppe. "On Audit and Certification of Machine Learning Systems." 2023 34th Conference of Open Innovations Association (FRUCT). IEEE, 2023.
[17] Намиот, Д. Е., and Е. А. Ильюшин. "Доверенные платформы искусственного интеллекта: сертификация и аудит." International Journal of Open Information Technologies 12.1 (2024): 43-60.
[18] OECD (2025), "Mapping relevant data collection mechanisms for AI training", OECD Artificial Intelligence Papers, No. 48, OECD Publishing, Paris, https://doi.org/10.1787/3264cd4c-en
[19] Perset, K. and S. Fialho Esposito (2025), "Howare AI developers managing risks?: Insights from responses to the reporting framework of the Hiroshima AI Process Code of Conduct", OECD Artificial Intelligence Papers, No. 45, OECD Publishing, Paris, https://doi.org/10.1787/658c2ad6-en.
[20] Безопасность ИИ-агентов https://abava.blogspot.com/2025/10/blog-post_23.html Retrieved: Oct, 2025
[21] 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.
[22] These psychological tricks can get LLMs to respond to "forbidden" prompts https://arstechnica.com/science/2025/09/these-psychological-tricks-can-get-llms-to-respond-to-forbidden-prompts/ Retrieved: Oct, 2025
[23] Sokhansanj, Bahrad A. "Uncensored AI in the Wild: Tracking Publicly Available and Locally Deployable LLMs." Future Internet (2025).
[24] OWASP AI Testing Guide! https://github.com/OWASP/www-project-ai-testing-guide Retrieved: Oct, 2025
[25] Fernandes T., Magalhães J. P., Alves W. Cybersecurity in Smart Railways: exploring risks, vulnerabilities and mitigation in the data communication services //Green Energy and Intelligent Transportation. – 2025. – С. 100305
[26] Интернет цифровой железной дороги / В. П. Куприяновский, Г. В. Суконников, С. А. Синягов [и др.] // International Journal of Open Information Technologies. – 2016. – Т. 4, № 12. – С. 53-68. – EDN XETADZ.
[27] Цифровая железная дорога - инновационные стандарты и их роль на примере Великобритании / Д. Е. Николаев, В. П. Куприяновский, Г. В. Суконников [и др.] // International Journal of Open Information Technologies. – 2016. – Т. 4, № 10. – С. 55-61. – EDN WXBAZN.
[28] Stealthy attack serves poisoned web pages only to AI agents https://www.helpnetsecurity.com/2025/09/05/ai-agents-prompt-injection-poisoned-web/ Retrieved: Oct, 2025
[29] Namiot, Dmitry, and Eugene Ilyushin. "On the Cybersecurity of AI Agents." International Journal of Open Information Technologies 13.9 (2025): 13-24.
[30] Stealthy attack serves poisoned web pages only to AI agents https://www.helpnetsecurity.com/2025/09/05/ai-agents-prompt-injection-poisoned-web/ Retrieved: Oct, 2025
[31] Grigoris Ntousakis, Julian James Stephen, Michael V. Le, Sai Sree Laya Chukkapalli, Teryl Taylor, Ian M. Molloy, and Frederico Araujo. 2025. Securing MCP-based Agent Workflows. In Proceedings of the 4th Workshop on Practical Adoption Challenges of ML for Systems (PACMI '25). Association for Computing Machinery, New York, NY, USA, 50–55. https://doi.org/10.1145/3766882.3767177
[32] Breaking Through: How to Predict, Prevent, and Prevail over the PRC Cyber Threat https://www.boozallen.com/content/dam/home/pdf/cyber/prc-cyber-report.pdf Retrieved: Oct, 2025
[33] Averting a Robot Catastrophe Preparing for Converging Trends in Robotics and Frontier AI https://www.rand.org/pubs/perspectives/PEA3691-7.html Retrieved: Oct, 2025

Статья получена 25 октября 2025.
Д.Е. Намиот – МГУ имени М.В. Ломоносова (e-mail: dnamiot@cs.msu.ru).

Artificial Intelligence in Cybersecurity. Chronicle. Issue 3

Dmitry Namiot

***Abstract* - In this document, we present our third monthly review of current events related to the general topic of using Artificial Intelligence (AI) in cybersecurity. This regularly published document describes regulatory documents, events, and new developments in this field. Currently, we focus on these three aspects. First, these are incidents related to the use of AI in cybersecurity. For example, newly disclosed attacks on machine learning models, identified vulnerabilities and risks in generative AI, etc. Second, these are regulatory documents and new global and local standards related to various aspects of AI in cybersecurity. And third, each review includes new, interesting publications in this area. Naturally, all materials selected for each issue reflect the views and preferences of the authors. This article presents the third edition of our Chronicle of AI in Cybersecurity.**

***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] Namiot, Dmitry. "Artificial Intelligence in Cybersecurity. Chronicle. Issue 1." International Journal of Open Information Technologies 13.9 (2025): 34-42.
[4] Malicious MCP https://www.koi.ai/blog/postmark-mcp-npm-malicious-backdoor-email-theft Retrieved: Oct, 2025
[5] Namiot, D. E., and E. A. Il'jushin. "Ujazvimosti jekosistemy MCP." International Journal of Open Information Technologies 13.10 (2025): 74-82.
[6]Command Injection Flaw in Framelink Figma MCP Server Puts Nearly 1 Million Downloads at Risk https://www.koi.ai/blog/command-injection-flaw-in-framelink-figma-mcp-server-puts-nearly-1-million-downloads-at-risk Retrieved: Oct, 2025
[7] Shiny tools, shallow checks: how the AI hype opens the door to malicious MCP servers https://securelist.com/model-context-protocol-for-ai-integration-abused-in-supply-chain-attacks/117473/ Retrieved: Oct, 2025
[8] The Month of AI Bugs https://monthofaibugs.com/ Retrieved: Oct, 2025
[9] Embrace The Red https://embracethered.com/blog/index.html Retrieved: Oct, 2025
[10] Detecting Exposed LLM Servers: A Shodan Case Study on Ollama https://blogs.cisco.com/security/detecting-exposed-llm-servers-shodan-case-study-on-ollama Retrieved: Oct, 2025
[11] ZipLine Campaign: A Sophisticated Phishing Attack Targeting US Companies https://research.checkpoint.com/2025/zipline-phishing-campaign/ Retrieved: Oct, 2025
[12] Lebed, S. V., et al. "Large Language Models in Cyberattacks." Doklady Mathematics. Vol. 110. No. Suppl 2. Moscow: Pleiades Publishing, 2024
[13] Namiot, Dmitry. "Artificial Intelligence in Cybersecurity. Chronicle. Issue 2." International Journal of Open Information Technologies 13.10 (2025): 58-67.
[14] Kuzmenko, Ilya Dmitrievich, and Dmitry Evgenyevich Namiot. "Metody obnaruzhenija dipfejkov v videokonferencijah v real'nom vremeni." Sovremennye informacionnye tehnologii i IT-obrazovanie 21.2 (2025).
[15] Prakasha, K. Krishna, and U. Sumalatha. "Privacy-preserving techniques in biometric systems: Approaches and challenges." IEEE Access (2025).
[16] Namiot, Dmitry, and Manfred Sneps-Sneppe. "On Audit and Certification of Machine Learning Systems." 2023 34th Conference of Open Innovations Association (FRUCT). IEEE, 2023.
[17] Namiot, D. E., and E. A. Il'jushin. "Doverennye platformy iskusstvennogo intellekta: sertifikacija i audit." International Journal of Open Information Technologies 12.1 (2024): 43-60.
[18] OECD (2025), "Mapping relevant data collection mechanisms for AI training", OECD Artificial Intelligence Papers, No. 48, OECD Publishing, Paris, https://doi.org/10.1787/3264cd4c-en
[19] Perset, K. and S. Fialho Esposito (2025), "How are AI developers managing risks?: Insights from responses to the reporting framework of the Hiroshima AI Process Code of Conduct", OECD Artificial Intelligence Papers, No. 45, OECD Publishing, Paris, https://doi.org/10.1787/658c2ad6-en.
[20] Bezopasnost' II-agentov https://abava.blogspot.com/2025/10/blog-post_23.html Retrieved: Oct, 2025
[21] 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.
[22] These psychological tricks can get LLMs to respond to "forbidden" prompts https://arstechnica.com/science/2025/09/these-psychological-tricks-can-get-llms-to-respond-to-forbidden-prompts/ Retrieved: Oct, 2025
[23] Sokhansanj, Bahrad A. "Uncensored AI in the Wild: Tracking Publicly Available and Locally Deployable LLMs." Future Internet (2025).
[24] OWASP AI Testing Guide! https://github.com/OWASP/www-project-ai-testing-guide Retrieved: Oct, 2025
[25] Fernandes T., Magalhães J. P., Alves W. Cybersecurity in Smart Railways: exploring risks, vulnerabilities and mitigation in the data communication services //Green Energy and Intelligent Transportation. – 2025. – S. 100305
[26] Internet cifrovoj zheleznoj dorogi / V. P. Kuprijanovskij, G. V. Sukonnikov, S. A. Sinjagov [i dr.] // International Journal of Open Information Technologies. – 2016. – T. 4, # 12. – S. 53-68. – EDN XETADZ.
[27] Cifrovaja zheleznaja doroga - innovacionnye standarty i ih rol' na primere Velikobritanii / D. E. Nikolaev, V. P. Kuprijanovskij, G. V. Sukonnikov [i dr.] // International Journal of Open Information Technologies. – 2016. – T. 4, # 10. – S. 55-61. – EDN WXBAZN.
[28] Stealthy attack serves poisoned web pages only to AI agents https://www.helpnetsecurity.com/2025/09/05/ai-agents-prompt-injection-poisoned-web/ Retrieved: Oct, 2025
[29] Namiot, Dmitry, and Eugene Ilyushin. "On the Cybersecurity of AI Agents." International Journal of Open Information Technologies 13.9 (2025): 13-24.
[30] Stealthy attack serves poisoned web pages only to AI agents https://www.helpnetsecurity.com/2025/09/05/ai-agents-prompt-injection-poisoned-web/ Retrieved: Oct, 2025
[31] Grigoris Ntousakis, Julian James Stephen, Michael V. Le, Sai Sree Laya Chukkapalli, Teryl Taylor, Ian M. Molloy, and Frederico Araujo. 2025. Securing MCP-based Agent Workflows. In Proceedings of the 4th Workshop on Practical Adoption Challenges of ML for Systems (PACMI '25). Association for Computing Machinery, New York, NY, USA, 50–55. https://doi.org/10.1145/3766882.3767177
[32] Breaking Through: How to Predict, Prevent, and Prevail over the PRC Cyber Threat https://www.boozallen.com/content/dam/home/pdf/cyber/prc-cyber-report.pdf Retrieved: Oct, 2025
[33] Averting a Robot Catastrophe Preparing for Converging Trends in Robotics and Frontier AI https://www.rand.org/pubs/perspectives/PEA3691-7.html Retrieved: Oct, 2025.