Искусственный Интеллект в Кибербезопасности. Хроника. Выпуск 9

Д.Е. Намиот

***Аннотация* – Данная публикация представляет девятый выпуск периодического аналитического обзора, посвящённого применению искусственного интеллекта (ИИ) в сфере кибербезопасности. Представленный цикл материалов ориентирован на углублённое изучение динамично развивающейся области, формирующейся на пересечении технологий искусственного интеллекта и информационной безопасности. Основной задачей проекта является систематический мониторинг глобальных тенденций и обобщение наиболее значимых событий в указанной предметной области. Помимо агрегации информации, в рамках инициативы осуществляется детальный анализ нормативно-правовых актов, резонансных инцидентов и передовых технологических решений, определяющих современную конфигурацию кибербезопасности под воздействием ИИ.**

Каждый выпуск серии имеет унифицированную структуру, включающую три раздела, что обеспечивает комплексный охват рассматриваемой проблематики. Первый раздел посвящён анализу инцидентной базы и актуальных вызовов безопасности: в нём исследуются реальные сценарии атак, выявляются новые уязвимости и даётся оценка угрозам, возникающим вследствие интеграции алгоритмов ИИ как в оборонительные механизмы, так и в инструментарий злоумышленников. Второй раздел содержит характеристику текущего состояния нормативно-правовой среды и основных направлений ее трансформации. Осмысление данных процессов имеет первостепенное значение, поскольку именно они задают правовые и эксплуатационные параметры, в которых предстоит развиваться надёжным и защищённым системам на базе ИИ. Третий раздел освещает хронику научно-технических достижений. Каждый выпуск включает аннотированный перечень наиболее значимых, с точки зрения авторов, научных работ, экспертных докладов ведущих организаций и описаний инновационных разработок.

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

I. Введение

С 2020 года кафедра информационной безопасности факультета вычислительной математики и кибернетики МГУ имени М.В. Ломоносова ведёт научно-исследовательскую деятельность на пересечении искусственного интеллекта (ИИ) и кибербезопасности. На факультете была открыта и успешно функционирует первая магистерская программа по указанной тематике[1]. За период её реализации состоялось несколько выпусков; в рамках данной образовательной траектории подготовлено более 40 специалистов. Значительная часть магистерских диссертаций выпускников послужила основой для последующих прикладных разработок в рассматриваемой области [1–4]. В целом, за прошедшее с момента запуска первой программы время, мы накопили самый большой массив публикаций на русском языке по данной тематике[2].

В настоящее время вопросы кибербезопасности систем искусственного интеллекта, в частности – атак на модели машинного (глубокого) обучения, рассматриваются и в других магистерских программах кафедры – Кибербезопасность[3], ПОВС[4], а также программах бакалавриата кафедры ИБ и системы дополнительного образования[5]. В недавно состоявшемся выпуске 2026 года более 20 выпускных квалификационных работ и магистерских диссертаций было посвящено этой тематике.

В ранних работах [5, 6] сотрудниками кафедры были выделены четыре основных направления взаимодействия ИИ и кибербезопасности:
- применение ИИ для защиты информационных систем;
- использование ИИ в целях осуществления атак;
- обеспечение защищённости самих систем ИИ;
- технология дипфейков.

Следует особо отметить высокую динамику эволюции данной предметной области. Феномен дипфейков представляет собой лишь одну из многих угроз, ассоциированных с генеративными моделями [7], что обусловливает необходимость комплексного анализа рисков, порождаемых синтезированным контентом. Показательным примером служит актуализация базового документа Национального института стандартов и технологий (NIST), посвящённого таксономии состязательного машинного обучения [8]. В редакции 2025 года (предыдущая версия вышла в 2023 году) указанный документ в полном объёме включает технологии генеративного ИИ (GenAI) в свою таксономическую структуру, детально описывая специфику атак на большие языковые модели (LLM), системы дополненной генерации поиска (RAG) и архитектуры на основе ИИ-агентов. Соответственно, последний пункт в указанном выше перечне необходимо уже отнести к генеративному ИИ.

В соответствии с приведённой таксономией выстроены занятия в магистерской программе «Искусственный

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

[2] https://abava.blogspot.com/2026/05/24052026.html

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

[4] http://master.cmc.msu.ru/?q=ru/node/3318

[5] [5] https://dpo.cs.msu.ru/courses

интеллект в кибербезопасности». Вопросы защищённости систем ИИ (атаки на ИИ-системы) в настоящее время также рассматриваются в рамках магистерской программы «Кибербезопасность». В аналогичной логике формируется и готовящийся к изданию учебник, в выпуске которого, как ожидается, окажет содействие Центральный университет[6]. За время, прошедшее после публикации предыдущего выпуска «Хроники» [9], для нового курса по разработке ИИ-агентов было обновлено учебно-методическое пособие, посвящённое вопросам безопасности ИИ-агентов[7].

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

Периодичность выпуска обзора составляет один раз в месяц. Первый выпуск вышел в сентябре 2025 года [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. Девятый 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[8]. Традиционно принимаются к рассмотрению новые статьи для публикации в журнале INJOIT[9] (издание входит в Перечень ВАК, РИНЦ, ЕГПНИ).

II. Инциденты в ИИ

Атака Шиллинга [11] на сайт Headhunter.ru[10].

Получился довольно масштабный эксперимент, когда (цитируем оригинальный материал): сделали фальшивое digital-агентство, назвали его (по классике) «Рога и копыта», но только на английском Horns and Hooves. Подобрали для него заброшенное юридическое лицо. ИИ создал легенду, что рекламное агентство собрано из бывших топ-менеджеров других агентств, что у него очень много клиентов, экстенсивный рост и сильный дефицит сотрудников. Создали сайт агентства, вели все социальные сети, блог, карточку на HH.ru, несколько карточек компании на отзовиках и так далее. При этом весь контент для карточек, сайта и вообще все-все материалы были произведены в ChatGPT. Далее ИИ создавал вакансии, писал отзывы от “сотрудников” и т.д. Итог: фиктивная компания «Рога и Копыта» через полгода вошла в ТОП-20 из 600 участников среди небольших компаний, и в ТОП-1 среди всего рейтинга компаний России из 1792 участников — в категории «Самые лояльные сотрудники»!

Красивая состязательная атака, когда атакующий морзянкой (представив текст азбукой Морзе) заставил торговый бот перевести ему криптовалюту на $200 000[11].

Между прочим, в работе [12] авторы приводят интересную статистику успешности джелбрейков в зависимости от формы представления. В частности, отмечается, что запросы, основанные на кодировании, показывают наиболее высокие значения ASR (Attack Success Rate). Преобразование запросов в нестандартные представления использует более слабую защиту от ошибок вне типичного естественного языка. CipherChat [13] сообщает о почти 100% обходе защиты GPT-4 с помощью кодирования шифра. Перевод на языки с ограниченными ресурсами увеличивает процент обхода с <1% до 79% [14]. ArtPrompt [15] использует ASCII-графику, и в связанных работах показано, что другие нестандартные представления, такие как Base64, ROT13 и азбука Морзе [16], аналогичным образом используют более слабую защиту от ошибок в этих пространствах кодирования.

Компания Waymo, демонстрирующая впечатляющие статистические данные, показывающие, что её беспилотные автомобили намного безопаснее, чем автомобили с водителями-людьми, развернула тысячи компьютеризированных роботакси по всей стране, которые не могут управлять автомобилем в нетрезвом виде, отвлекаться на телефоны или проявлять агрессию на дороге.

Однако анализ CNN данных местных и федеральных органов власти, а также видеороликов в социальных сетях показывает, что та же самая особенность, которая делает роботакси Waymo менее склонными к опасным столкновениям - отсутствие водителя-человека — также создает совершенно новые проблемы безопасности, которые беспокоят правительственных чиновников, поскольку компания стремится расширить свою деятельность за пределы 11 городов, где она в настоящее время работает.

CNN выявила сотни инцидентов, в которых роботакси якобы совершали опасные маневры и испытывали трудности с препятствиями, с которыми люди инстинктивно справляются. Они проезжали на красный

[6] https://cu.ru
[7] http://inetique.ru/articles/agents_security.pdf
[8] dnamiot@cs.msu.ru
[9] http://injoit.org
[10] https://habr.com/ru/articles/1042590/
[11] https://www.dexerto.com/entertainment/x-user-tricks-grok-into-sending-them-200000-in-crypto-using-morse-code-3361036/

свет, выезжали на встречную полосу и на места преступлений, не соблюдали правила дорожного движения и приближались на расстояние нескольких сантиметров к пешеходам, законно переходящим улицу - ошибки, которые роботизированные автомобили должны быть запрограммированы избегать.

За последние два месяца компания Waymo отозвала тысячи автомобилей и приостановила работу в нескольких городах после того, как роботакси выехали на затопленные улицы, в том числе в Сан-Антонио, где пустой автомобиль Waymo был смыт бурным потоком воды. А недавно компания объявила о приостановке работы на всех автомагистралях в таких городах, как Лос-Анджелес и Майами, после того, как один из пассажиров из Сан-Франциско рассказал на форуме X, что его роботакси устроило высокоскоростную погоню с полицией через зону активных дорожных работ[12].

В России впервые вынесли приговор за публикацию порно, сгенерированного ИИ. В России осудили подростка, который занимался распространением сгенерированной искусственным интеллектом порнографии. Молодой человек сказал, что разослал созданные нейронной сетью кадры, чтобы пошутить над потерпевшей. Суд назначил ему наказание в виде двух лет лишения свободы условно с испытательным сроком один год[13].

На улицах китайского города Ухань роботакси Apollo Go компании Baidu остановились посреди дороги, заблокировав пассажиров и вызвав столкновения[14].

Отметим, что автономный транспорт сегодня сделать безопасным можно только декларативно. То есть объявить безопасным и все. Обосновать (формально подтвердить) это нельзя [17].

Очередные галлюцинации, найденные в статьях. Библиографическая ссылка в статье: Куприяновский, В.П. Умные дома как основа умных городов / В.П. Куприяновский, С.А. Тищенко // International Journal of Open Information Technologies. - 2016. - Т. 4, № 2. - С. 16-23[15].

Такой статьи не существует. Работа, которая это "цитирует", была написана генеративной моделью. Библиотека Elibrary не проверяет мусор, который в нее загружают.

Опрос Cloud Security Alliance: инциденты с ИИ-агентами распространены на предприятиях[16]. В 418 опрошенных организациях инциденты с ИИ-агентами стали обычным явлением на предприятиях. 65% организаций столкнулись, как минимум, с одним инцидентом кибербезопасности, связанным с ИИ-агентами, за последние 12 месяцев; 88% подтвердили или заподозрили инциденты с ИИ-агентами; 82% обнаружили ранее неизвестных ИИ-агентов в своей среде за последний год; 61% сообщили об утечке данных, 43% — о сбоях в работе, 35% — о финансовых потерях и 41% — о непреднамеренных действиях в бизнес-процессах в результате этих инцидентов; только 21% имели формальные процессы вывода из эксплуатации. Первопричина: Анализ опроса; пробелы в видимости и выводе из эксплуатации позволяют теневым ИИ-агентам работать вне контроля идентификации, сети и аудита; уровень инцидентов в здравоохранении достигает 92,7%.

Появился еще один, еженедельно обновляемый список ИИ-инцидентов[17].

База данных инцидентов ИИ[18] в свежих записях отмечает множественные проблемы с автономным транспортом. Например: сообщается, что роботакси Waymo въезжали в закрытые зоны строительства на автомагистралях в Аризоне и Калифорнии[19], роботакси Waymo препятствовало проезду машины скорой помощи к месту массовой стрельбы в Остине[20], роботакси Waymo выехало на рельсы легкорельсовой линии в Финиксе, вынудив пассажира выйти[21], роботакси Zoox выехало на встречную полосу и остановилось рядом со встречным движением[22] и т.д.

ChatGPT использовали для планирования нескольких публичных атак, выборы в Косово сопровождались массовой генерацией дипфейков, OpenClaw продолжал свою работу по уничтожению пользовательских данных, а американских юристов привлекали к ответственности за галлюцинации в документах, которые они готовили с помощью ИИ. Из удивительного можно отметить Google, который неожиданно обнаружил, что китайские спамеры используют Gemini для подготовки рассылок[23].

III Регуляции и стандарты

Здесь можно привести несколько новостей, которые объединяются общей идеей – “не до стандартов сейчас”. Точнее – не до ограничений. В разных странах взялись за облегчение ограничений, накладываемых на ИИ. Что, естественно, относится и к кибербезопасности систем ИИ.

За аварии с беспилотными авто в России будут платить владельцы машин, а не разработчики[24]. Согласно новой редакции закона о беспилотных автомобилях, подготовленной Минтрансом, обязанность

[12] https://edition.cnn.com/us/waymo-robotaxis-safety-invs

[13] https://www.cnews.ru/news/top/2026-06-05_v_rossii_vpervye_vynesli

[14] https://apnews.com/article/china-baidu-robotaxi-wuhan-ec3dcd026dfa1a19eb79467d5211f01d

[15] https://elibrary.ru/item.asp?id=90768757

[16] https://cloudsecurityalliance.org/press-releases/2026/04/21/new-cloud-security-alliance-survey-reveals-82-of-enterprises-have-unknown-ai-agents-in-their-environments

[17] https://github.com/webpro255/awesome-ai-agent-attacks

[18] https://incidentdatabase.ai

[19] https://incidentdatabase.ai/cite/1547/

[20] https://www.axios.com/local/austin/2026/03/02/waymo-vehicle-blocks-ems-austin-mass-shooting

[21] https://www.azfamily.com/2026/01/08/waymo-passenger-flees-after-car-drives-phoenix-light-rail-tracks/

[22] https://www.reuters.com/business/autos-transportation/amazons-zoox-recall-332-us-vehicles-over-software-error-nhtsa-says-2025-12-23/

[23] https://www.nytimes.com/2026/06/12/technology/google-lawsuit-china-ai-scams.html

[24] https://abava.blogspot.com/2026/06/blog-post_12.html

возмещать ущерб, нанесенный транспортным средством без водителя за рулем, ляжет на его владельца. Закон может вступить в силу уже с осени 2027 г.

Генеральный директор OpenAI Сэм Альтман будет выступать против предложений о том, чтобы разработчики ИИ получали одобрение правительства США перед выпуском новых моделей в открытый доступ, говорится в заявлении компании[25].

Альтман попросит Конгресс увеличить финансирование тестирования искусственного интеллекта в Министерстве торговли США. Министерство уже сотрудничает с такими компаниями, как OpenAI и Anthropic, для тестирования их моделей. В заявлении компании говорится, что OpenAI хочет, чтобы правительство США расширило эту инициативу и привлекло ученых, обладающих опытом в области кибербезопасности, биологического оружия и национальной безопасности, среди прочих тем.

Визит Альтмана в Вашингтон совпадает с критическим периодом для компании и отрасли. Как ранее сообщало агентство Reuters, компания OpenAI готовится конфиденциально подать заявку на первичное публичное размещение акций (IPO). Конкурент Anthropic, производитель Claude, уже подал заявку на IPO в США. Требования федерального правительства могут негативно сказаться на прибыли отрасли, если они замедлят внедрение новых моделей или побудят компании изменить характеристики своей продукции для решения проблем безопасности.

Европейский союз ослабил некоторые положения своего знакового Закона об искусственном интеллекте и отложил принятие других после того, как представители бизнеса и политики заявили, что закон снижает конкурентоспособность европейских компаний.

Что нового: Европейский парламент и государства-члены договорились внести поправки в Закон об искусственном интеллекте, чтобы отложить ограничения, направленные на приложения, которые, по мнению союза, представляют значительную угрозу безопасности, здоровью или правам человека, а также внести другие изменения. Поправки ожидают официального принятия Советом и парламентом союза. ЕС охарактеризовал поправки как «более безопасные и простые правила как для граждан, так и для бизнеса».

Как это работает: поправки в целом упрощают надзорные и правоприменительные обязанности Управления ЕС по искусственному интеллекту. Они также продлевают сроки для разработчиков ИИ по соблюдению определенных положений и упрощают другие.

Требования к системам искусственного интеллекта, считающимся «высокорисковыми» — включая системы, используемые в правоохранительных органах, критической инфраструктуре, сфере занятости, миграции и идентификации личности — отложены до декабря 2027 года с ранее установленного срока в августе 2026 года. Разработчикам будет предоставлено время до августа 2027 года для внедрения контролируемых тестовых сред, позволяющих изолировать новые модели от внешнего мира во время тестирования. Сроки для продуктов, использующих ИИ, включая оборудование и игрушки, также продлены до августа 2028 года, а требования к водяным знакам для результатов работы ИИ и другие требования к прозрачности — примерно до декабря 2026 года.

Изменения коснутся способов использования персональных данных при обучении и развертывании систем ИИ. В соответствии с действующим законодательством ЕС, некоторые категории персональных данных могут использоваться только в случае «строго необходимого» применения. Изменения позволят использовать персональные данные для выявления и смягчения предвзятости.

Также были внесены или уточнены исключения для некоторых продуктов. Например, Закон об ИИ не затронет промышленное оборудование, которое уже регулируется законами о безопасности продукции. Кроме того, в некоторых случаях к малым компаниям (менее 50 сотрудников с годовым мировым доходом до 10 миллионов евро или совокупными активами до 10 миллионов евро) и компаниям «малой средней капитализации» (примерно от 250 до 749 сотрудников с годовым мировым доходом до 150 миллионов евро или совокупными активами до 129 миллионов евро) будут применяться более мягкие требования к соблюдению нормативных требований и административные издержки.

Поправки усиливают Закон об ИИ в одной важной области: они запрещают создание изображений сексуального характера с участием детей и изображений обнаженных людей без их согласия.

За кулисами новости: в 2024 году ЕС принял самый строгий в мире закон о регулировании ИИ. Закон вступил в силу в том же году, при этом некоторые положения будут вводиться поэтапно в последующие годы. Он подвергся критике за наложение необоснованных требований без повышения безопасности практически с момента начала законодательного процесса.

В 2023 году руководители 163 компаний подписали письмо, в котором утверждалось, что законодательство носит «бюрократический характер». В 2025 году 110 компаний призвали политиков отложить разработку графика внедрения, поскольку правила были «неясными, дублирующими друг друга и все более сложными». Такие компании, как немецкие промышленные и программные фирмы Siemens и SAP, лоббировали пересмотр, заявляя, что правила сдерживают их развитие.

Два ранних отчета повлияли на поправки. В отчете, опубликованном в апреле 2024 года бывшим премьер-министром Италии Энрико Леттой, утверждалось, что ЕС фрагментирован на 27 национальных рынков, что препятствует масштабированию европейских компаний так же, как это могут делать американские и китайские компании. В отчете за сентябрь 2024 года о конкурентоспособности Европы стагнация роста ВВП

[25] https://www.reuters.com/business/openais-altman-urge-us-lawmakers-not-require-ai-model-approvals-2026-06-03

региона была представлена как «экзистенциальный вызов», и основное внимание уделялось сокращению инновационного разрыва, декарбонизации и снижению зависимости.

В начале 2025 года Европейская комиссия — исполнительный орган ЕС — объявила о своем намерении снизить регуляторную нагрузку, упростить правила и повысить экономическую конкурентоспособность.

В феврале 2026 года Европейская комиссия отозвала предложенную ею Директиву об ответственности за использование ИИ — спорный законопроект, отдельный от Закона об ИИ, который предусматривал введение общеевропейских стандартов для судебных исков, связанных с вредом, причиненным ИИ.

Реакция общественности: Непосредственная реакция на поправки была неоднозначной. Индустрия ИИ в целом приветствовала дополнительную гибкость, в то время как группы потребителей выразили обеспокоенность по поводу потенциального ослабления стандартов безопасности. Некоторые сообщения в СМИ представили их как смягчение закона в угоду интересам бизнеса. Европейская организация потребителей заявила, что соглашение делает цифровую среду менее безопасной и создает опасные лазейки для компаний, занимающихся ИИ.

Почему это важно: как в первоначальной, так и в обновленной версиях, Закон об ИИ направлен на смягчение «системных рисков», вызванных ИИ. Эта концепция заимствована из регулирования в сфере финансов и инфраструктуры и относится к сбоям, способным распространяться на различные отрасли или значительные части экономики. Идея о том, что ИИ представляет собой системные риски, остается спекулятивной, в то время как чрезмерное регулирование создает экономический риск подавления инноваций и блокировки полезных технологий. Поправки направлены на баланс рисков и преимуществ за счет снижения нагрузки на разработчиков, предоставления компаниям дополнительного времени для понимания и соблюдения требований, а также создания условий для дальнейших инноваций в критически важных отраслях, таких как производство и полупроводники.

Многие положения первоначального Закона об ИИ были неясными, чрезмерно широкими или излишне обременительными. Эти поправки, по-видимому, делают закон менее обременительным, сохраняя при этом полезные элементы. Это хороший шаг для повышения конкурентоспособности Европы. Источник: deeplearning.ai[26].

Президент США Дональд Трамп в заявил, что отложил подписание указа об искусственном интеллекте, поскольку ему не нравятся некоторые его аспекты, и он не хочет предпринимать никаких шагов, которые могли бы подорвать позиции США в конкуренции с Китаем в области ИИ. Трамп планировал подписать указ на церемонии, на которой должны были присутствовать руководители компаний, занимающихся ИИ.

Американские СМИ, включая Semafor и Washington Post, сообщили, что планы администрации были приостановлены после настойчивых просьб основателя xAI Илона Маска и генерального директора Meta Марка Цукерберга, а также бывшего советника Трампа по вопросам ИИ Дэвида Сакса.

«Я думаю, это помешает, знаете ли, мы лидируем в Китае, мы лидируем во всех, и я не хочу делать ничего, что помешало бы этому лидерству», — сказал Трамп журналистам в Овальном кабинете. Указ создаст добровольную структуру для взаимодействия разработчиков ИИ с правительством США до публичного выпуска передовых моделей ИИ, сообщили Reuters в среду два источника, знакомые с указом.

Трамп не уточнил, против каких именно частей исполнительного указа он возражает. Представители технологической индустрии опасаются, что положения указа могут нанести ущерб прибыли отрасли, если они замедлят внедрение новых моделей или побудят компании изменить работу этих моделей для решения проблем безопасности. По словам другого источника, президент также планировал поручить правительству США использовать передовые модели для улучшения кибербезопасности государственных систем, а также сетей, принадлежащих секторам, имеющим жизненно важное значение для экономики, таким как банки и больницы. В правительстве США и частном секторе растет обеспокоенность по поводу рисков кибербезопасности, создаваемых мощными новыми системами искусственного интеллекта, включая Mythos от Anthropic. Anthropic предупреждала, что Mythos может значительно ускорить сложные кибератаки, хотя эксперты по кибербезопасности заявили Reuters, что опасения по поводу бесконтрольного взлома преувеличены.

После возвращения к власти Трамп занял более мягкую позицию по отношению к крупным технологическим компаниям, чем администрация его предшественника Джо Байдена, в связи с появлением искусственного интеллекта и его огромной ролью на американских фондовых рынках. Однако некоторые видные сторонники Трампа призывают к ужесточению мер контроля. Источник: Reuters[27].

В России Правительство представило новую версию закона об ИИ, который, по данным СМИ[28], значительно скорректирован по сравнению с предыдущими версиями [18], в частности, изменился предмет регулирования. Из документа убрали "доверенные" модели, уточнили критерии суверенных и национальных моделей, исключена обязанность обучать модели только на российских данных. Полностью убран запрет на трансграничный ИИ, который фигурировал в первых версиях документа и вызывал опасения тем, что может

[26] https://www.deeplearning.ai/the-batch

[27] https://www.reuters.com/business/retail-consumer/white-house-postpones-trumps-ai-signing-ceremony-says-axios-2026-05-21

[28] https://www.comnews.ru/content/245938/2026-06-22/2026-w26/1009/pravitelstvo-predstavilo-novuyu-versiyu-zakona-ob-ii-podderzhka-vmesto-zapretov

привезти к запрету в России иностранных нейросетей.

Законопроект получил новое название, что свидетельствует об изменении фокуса. Вместо "регулирования ИИ" теперь он озаглавлен "О поддержке развития технологий ИИ". Документ стал компактнее - число статей сократилось с 21 до 13. Исключены обязанности операторов, владельцев и пользователей систем и сервисов, только обозначены вопросы международного взаимодействия, регулирование профильных ЦОД вынесено в подзаконные акты.

Главное изменение - корректировка фокуса закона. Если раньше предполагалось регулирование ИИ "вообще", то теперь внимание сконцентрировано на больших фундаментальных моделях (БФМ), содержащих свыше 1 млрд параметров. Концентрация на больших фундаментальных моделях позволяет выводит из регулирования малые решения. То, как в законе прописываются конкретные цифровые значения – весьма удивительно.

То, что из регулирования исчезло понятие "доверенных" моделей также не менее удивительно.

Важным стало уточнение характеристик суверенных и национальных моделей и тех регуляторных подходов, которые будут применяться к ним. Обе должны быть разработаны российскими юрлицами. Обе должны хранить данные и готовить ответы для пользователей на территории РФ и кроме того - соответствовать российскому законодательству. Разработка суверенной модели и определение ее характеристик должны на всех этапах вестись российским юрлицом. А еще - суверенной может быть только та модель, что может быть "с нуля" воспроизведена на территории РФ.

Что правильно - убрано положение про обучение только на российских данных. Но есть требование соответствия модели традиционным духовно-нравственным ценностям. Соответственно, нужны будут локальные тесты для генеративных моделей.

Сроки вступления норм документа в силу могут быть скорректированы, но пока положения о применение суверенных/национальных моделей, определение суверенных и национальных моделей, обязанности разработчика, маркировка и особенности авторского права должны вступить в силу с 1 марта 2027 года. Случаи использования только суверенных и национальных моделей в отношении информационных систем, где на 1 марта 2027 года ИИ уже функционирует - с 1 сентября 2032. Здесь авторы документа отреагировали на запросы со стороны регионов. Все остальные нормы, как ожидается, в случае принятия документа, вступят в силу с 1 сентября 2026 года.

Большой и интересный материал от конференции NeurIPS – что делать со статьями, написанными генеративным ИИ[29]. Кратко: не принимать и не публиковать. Для контроля используют Pangram. Цитируем оригинал: “В этом году в рамках секции «Позиционные доклады» конференции NeurIPS 2026 было принято решение обязать авторов в значительной степени использовать человеческий труд, а ИИ - только для корректуры или подобных второстепенных изменений основного текста. Хотя мы признаем, что продуманное использование ИИ может привести к повышению производительности исследований, применение ИИ для написания докладов создает серьезный риск для системы рецензирования. В этом году, как председатели секции «Позиционные доклады», мы придерживаемся консервативного подхода, поскольку считаем, что в случае аргументированных работ, таких как позиционные доклады, чрезмерное использование ИИ при написании представленных докладов мало полезно для всего исследовательского сообщества. Текст, сгенерированный ИИ, часто выглядит привлекательно, но может значительно отличаться от первоначального замысла авторов. В этом случае представление текста, сгенерированного ИИ, на рецензирование перекладывает затраты на проверку этой работы на рецензентов. Если же сам текст, сгенерированный ИИ, не является бессвязным или вводящим в заблуждение, это поднимает вопросы о надлежащем распределении заслуг”.

Придерживаемся в журнале INJOIT такой же позиции. Статья - авторский материал. Есть (должен быть) автор. А просто при выполнении работы (производственного задания) есть исполнитель, который может, конечно, использовать любые инструменты.

И о перспективах профессии юриста. ИИ-юрист (фактически, чат-бот Garfield AI) выиграл первое дело в британском суде[30].

IV Обзор публикаций и проектов

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

Человек не поможет. Anthropic опубликовал разбор безопасности свои продуктов[31]. Цитата: "Первый способ обеспечения безопасности — это контроль поведения агента с помощью участия человека. Ранее Claude Code защищал агентов от непреднамеренных действий, запрашивая у пользователей разрешение на каждом шагу. Теоретически это работает, но мы обнаружили, что такой подход несовершенен. Наши телеметрические данные показали, что пользователи одобряли примерно 93% запросов на разрешение. Чем больше подтверждений видит пользователь, тем меньше внимания он уделяет каждому из них, со временем становясь гораздо менее внимательным к контролю”. То есть пользователи просто начинают подтверждать все подряд, если их часто спрашивать.

А про контекст ты не забыл? Появление агентов на основе больших языковых моделей (LLM),

[29] https://blog.neurips.cc/2026/06/02/ai-generated-papers-in-the-neurips-2026-position-paper-track/

[30] https://www.garfield.law/press/garfield-ai-wins-first-court-trial-with-regulated-ai-lawyer

[31] https://www.anthropic.com/engineering/how-we-contain-claude

дополненных использованием инструментов, навыками, и внешними знаниями, породило новые риски безопасности. Среди них основной угрозой стали атаки с внедрением подсказок, когда злоумышленники внедряют вредоносные инструкции в рабочий процесс агента. Однако существующие бенчмарки и средства защиты принципиально ограничены, поскольку они предполагают контекстно-независимые условия, в которых агент работает в соответствии с полностью заданной инструкцией пользователя, а атаки являются простыми и контекстно-независимыми. В результате они не позволяют оценить реальные условия эксплуатации, где поведение агента обычно зависит от динамического контекста, а не только от подсказки пользователя, и злоумышленники могут адаптировать свои атаки к различным контекстам. Аналогично, существующие средства защиты, построенные на этой узкой модели угроз, игнорируют природу реального делегирования агентам. В этой статье мы представляем AgentLure, бенчмарк, который позволяет выявлять контекстно-зависимые задачи и атаки с внедрением подсказок с учетом контекста. AgentLure охватывает четыре агентных домена и восемь векторов атак на различных поверхностях атаки. Наша оценка показывает, что существующие средства защиты часто испытывают трудности в этой среде, демонстрируя низкую эффективность против таких атак в агентных системах. Для решения этой проблемы мы предлагаем ARGUS, механизм защиты, который обеспечивает аудит решений с учетом происхождения информации для агентов LLM. ARGUS строит граф происхождения влияния, чтобы отслеживать, как недостоверный контекст распространяется на решения агентов, и проверяет, оправдано ли решение достоверными доказательствами до его выполнения. Наша оценка показывает, что ARGUS снижает вероятность успешной атаки до 3,8%, сохраняя при этом 87,5% полезности задачи, значительно превосходя существующие средства защиты и оставаясь устойчивым к адаптивным противникам типа «белый ящик» [19].

Косвенные инъекции подсказок при тестировании не должны быть статическими. Агенты на основе LLM все чаще используются для сложных задач, требующих планирования, использования инструментов, и взаимодействия с внешними сервисами. Их зависимость от ненадежного внешнего контента делает их уязвимыми для косвенной инъекции подсказок (IPI), при которой враждебные инструкции, встроенные в полученные данные, перехватывают поведение агента. Существующие атаки основаны на статических полезных нагрузках, которые не могут адаптироваться к специфическим для агента средствам защиты; даже в современных адаптивных методах отсутствует структурированная обратная связь для управления оптимизацией. Мы представляем IterInject, итеративную структуру с обратной связью, которая замыкает цикл между инъекцией, диагностикой и уточнением: диагност, основанный на правилах, генерирует структурированные метки результатов с описаниями поведения, а оптимизатор на основе LLM уточняет полезные нагрузки с учетом полной истории оптимизации. Этап синтеза генерирует новые начальные значения маскировки из шаблонов ошибок, позволяя пространству стратегий самостоятельно развиваться. На AgentDojo и InjectAgent IterInject значительно превосходит статические базовые модели и существующие адаптивные методы по четырем моделям жертв. Эксперименты по расширению на примере Claude Code, агента кодирования производственного уровня, обладающего многоуровневой защитой, показывают, что оптимизированные полезные нагрузки достигают полного успеха на 5 из 9 целей; даже те, которые сопротивляются полной эксплуатации, демонстрируют измеримое улучшение в результате итеративного уточнения. Мы также представляем механистический анализ IPI, выявляющий механизм порогового значения, опосредованный вниманием, на средних и поздних уровнях; три причинно-следственных вмешательства подтверждают это открытие и указывают на конкретные направления защиты [20].

О доверенных агентных системах. Агентные системы искусственного интеллекта — большие языковые модели (LLM), дополненные планированием, использованием инструментов, памятью и взаимодействием на долгосрочную перспективу — могут автономно выполнять сложные задачи, но их многоэтапные траектории приводят к новым режимам сбоев, которые ставят под сомнение надежность. Этот обзор представляет собой целенаправленное исследование надежного агентного ИИ по двум основным параметрам, которые имеют решающее значение для развертывания в условиях высокого риска: безопасность и надежность, а также конфиденциальность и безопасность системы. Для каждого параметра мы уточняем ключевые понятия, определяем, где возникают риски на протяжении рабочего процесса агента, и обобщаем стратегии смягчения рисков на каждом этапе. Другие аспекты надежности (согласование ценностей, прозрачность, справедливость и подотчетность) обсуждаются в качестве контекста, а не в отдельных главах. Для обеспечения согласованного сравнения и принятия решений о развертывании мы объединяем оценку в единый центр метрик и бенчмарков, уделяя особое внимание как результатам, так и сигналам процесса (например, нарушениям ограничений, полноте трассировки и показателям успешности противодействия) и предлагая рекомендации по преобразованию сценариев в метрики для управления процессом выпуска. В заключение мы описываем открытые проблемы, такие как саморазвивающиеся агенты, мониторинг и проверка в режиме реального времени, персонализация с сохранением конфиденциальности и компромисс между доверием и полезностью, а также представляем пример реальных сбоев безопасности в агентных системах с открытым исходным кодом (OpenClaw/Moltbook). Наша цель — служить практическим справочником для исследователей и практиков, создающих надежные

агентные системы в условиях высокой ответственности [21].

Странно, что тема доверенных систем исчезла из отечественного закона об ИИ (см. раздел выше). Китайские авторы цитируемой работы убеждены в их необходимости.

О правильном тестировании ИИ-агентов. Мы проводим всесторонний анализ безопасности автономных агентов-помощников,выявляя угрозы, присущие их уникальным архитектурным свойствам. Во-первых, мы создаем систематическую таксономию, охватывающую 20 реальных рисков, классифицированных на нарушения границ, устойчивое искажение состояния и вредоносные операции. Для дальнейшего выявления уязвимостей автономных агентов в условиях этих угроз мы предлагаем три передовые стратегии атаки, обеспечивающие обход защиты во временном, пространственном и семантическом измерениях:

(i) Кросс-поворотная фрагментация: фрагментация и распределение вредоносных полезных нагрузок по нескольким взаимодействиям в рамках одной сессии;

(ii) Обход защиты в пределах области обнаружения: внедрение полезных нагрузок атаки в сложные внешние артефакты, которые трудно проверить с помощью LLM; и

(iii) Сокрытие в благоприятном контексте: сокрытие вредоносных намерений в объемной, на первый взгляд безобидной информации в длительном контексте.

Мы моделируем эти риски и стратегии в A3S-Bench, эталонной системе, включающей 2254 многоходовых диалога (1512 случаев атак, охватывающих 34 метода атак, и 742 безопасных диалога). Набор данных охватывает шесть сценариев использования и два уровня сложности, сгенерированных с помощью автоматизированного трехэтапного конвейера синтеза. Каждый случай выполняется в изолированной среде и оценивается с использованием метрик оценки на основе действий, которые совместно количественно определяют как безопасность, так и полезность [22].

Для тестирования агентов (впрочем, как и для тестирования LLM) нужны развитые (multi-turn) диалоги, а не просто пары вопрос-ответ.

Это подтверждается и следующей работой, посвященной перефразированию. THREAT (Targeted Harmful generation via Reframing and Exploitation of Adversarial Tactics) - основанная на рассуждениях структуру, которая координирует работу нескольких LLM в итеративном цикле поиска для обнаружения текстовых подсказок для взлома. Мы формулируем задачу обнаружения подсказок как невыпуклую задачу оптимизации и предлагаем эффективное решение, которое сокращает время выполнения и повышает эффективность атаки. На различных наборах данных и архитектурах моделей THREAT обеспечивает более высокие показатели успешности атак при меньших вычислительных затратах, чем предыдущие методы. Созданные подсказки были помечены как вредоносные менее чем в 1% случаев, по сравнению с примерно 50% отказов для соответствующих неизмененных подсказок. Эти результаты выявляют ранее не обнаруженные уязвимости в выровненных LLM и позиционируют THREAT как практический инструмент для упреждающего повышения безопасности базовых моделей [23].

С помощью LLM перефразируют состязательные запросы до тех пор, пока их не перестанут отвергать.

Все об атаках на агенты. В данной статье представлена первая всесторонняя систематизация знаний о безопасности агентов ИИ, включая анализ пространства проектирования агентов, ландшафта атак и механизмов защиты для безопасных систем агентов ИИ. Мы также выявляем открытые проблемы, указывающие на перспективные направления будущих исследований в этой новой области. Наша 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 продвижения исследований в этой критически важной области [24].

Можем ли мы сделать искусственный интеллект неуязвимым для противников, которые хотят исказить технологию в вредных целях? Хотя ИИ — одна из новейших технологий, ответ на этот вопрос почти столетний по возрасту.

Как бы мы ни старались, мы никогда не сможем сделать искусственный интеллект полностью неоспоримым с помощью традиционных моделей безопасности. В рецензируемом журнале IEEE Security and Privacy Апостол Василев, старший научный сотрудник Национального института стандартов и технологий (NIST), опубликовал математическое доказательство этого утверждения, используя работу, опубликованную в 1931 году известным логиком Куртом Гёделем. Его теоремы о неполноте показали, что существуют пределы того, что можно доказать в системе, построенной на конечном числе правил[32].

OWASP, спустя год, выпустил новую версию отчета State of Agentic AI Security and Governance[33].

Этот документ является продолжением версии 1.0 (Июль 2025) и отражает стремительное развитие и внедрение агентных ИИ-систем. В нем собраны данные за год, зафиксированы реальные инциденты и предложены практические шаги по управлению рисками. Документ представляет собой комплексное руководство в области безопасности и управления рисками, подчеркивая переход от теории к практике и необходимость срочных организационных и технических изменений. Три ключевых вывода:

1. Угрозы стали реальностью: теоретические риски 2025 года теперь подтверждаются реальными инцидентами в проде, уязвимостями (CVE) и отчетами вендоров. Модель угроз больше не гипотетическая.
2. Безопасность (Safety) и Защита (Security) сходятся

[32] https://abava.blogspot.com/2026/06/blog-post_13.html

[33] https://genai.owasp.org/resource/state-of-agentic-ai-security-and-governance/

на уровне развертывания, тогда как в традиционном ПО это были разные области. В агентных системах, действующих автономно с широким доступом к инструментам, эти два понятия становятся неразделимыми. Один и тот же контроль управляет обоими типами вреда, и организационно они не могут существовать параллельно.

3. Управление должно поспевать за развертыванием: регуляторы уже исходят из того, что агенты могут причинять вред быстрее, чем человек может вмешаться. Это требует непрерывного мониторинга, а не периодических аудитов.

Документ классифицирует агентов по трем независимым измерениям для лучшего понимания рисков:

Типы агентов по роли: корпоративные, для написания кода, клиенто-ориентированные, персональные, инфраструктурные/операционные.

Шаблоны реализации: как агент построен (например, с использованием фреймворков, легковесных библиотек или low-code платформ). От этого зависит возможность его инвентаризации и аудита.

Шаблоны композиции: как агенты взаимодействуют (одиночный агент, мультиагентные системы, распределенные цепочки и т.д.). Это определяет структуру границ доверия.

Сквозное измерение: уровень автономности - от контролируемого человеком до полностью автономного. Чем выше автономность, тем выше риски.

Что входит в анализ угроз?

Расширяющаяся автономия: агенты получают все более привилегированные доступы (к файлам, API, облачной инфраструктуре). Разработчики вне корпораций (используется термин "Гражданские разработчики" - citizen developers) создают таких агентов вне контроля безопасности.

Инжекция промптов (подсказок): остается фундаментальной нерешенной проблемой. Так как LLM не разделяют данные и инструкции, одна инъекция может привести к краже данных, изменению планов и каскадным действиям.

Цепочка поставок агентов: стала основной вектором атак. Уязвимости найдены в протоколах (например, критическая RCE в MCP), вредоносные серверы MCP и навыки в реестрах, а также взлом ключевых инфраструктурных пакетов.

Летальное трио (Lethal Trifecta): условие, при котором инжекция становится максимально опасной: 1) агент имеет доступ к приватным данным, 2) взаимодействует с недоверенным контентом и 3) может передавать данные вовне. Рекомендуется использовать "Правило Двух" (Rule of Two): в одной сессии у агента не должно быть больше двух из этих трех свойств.

Безопасность (Safety) против Защиты (Security) – сливаются.

AI Security (Защита) – это риски нарушения границ доверия атакующим.

AI Safety (Безопасность) – это риски вреда от нормальной работы системы (ошибки, несоответствие намерениям).

В агентных системах грань стирается. Например, ошибка в разрешениях - это проблема Safety, но та же ошибка становится уязвимостью. Документ призывает к объединению этих функций в организациях.

Отмечается проблема управления идентичностью. Не-человеческие идентичности (сервисные аккаунты, учетные данные агентов) теперь сильно преобладают над человеческими, но управление ими отстает. Агенты получают собственные удостоверения, API-ключи и разрешения, что требует управления их жизненным циклом, аналогичного человеческим пользователям.

Подчеркивается необходимость в AI SBOM (Software Bill of Materials) для отслеживания компонентов модели, данных и зависимостей. При этом агенты, пишущие код, создают риск для цепочки поставок, обходя процессы ручного ревью кода.

Интересный раздел посвящен анализу тенденций экосистемы (построено на основе анализа 53 проектов). Здесь отмечены следующие пункты.

Консолидация: экосистема консолидируется вокруг нескольких быстрорастущих проектов.

Доминирование кодирующих агентов: агенты для написания кода составляют большинство проектов и являются основной точкой внедрения в крупных компаниях, что создает системные риски для цепочки поставок.

Беспрецедентная скорость релизов: новые версии выходят ежедневно и даже чаще, что делает традиционные процессы безопасности неэффективными.

Уязвимости в фреймворках: наибольшее количество уязвимостей (CVE) связано с популярными фреймворками и кодирующими агентами, которые широко используются в корпоративной среде.

И есть раздел с рекомендациями для руководителей.

1. Начните с инвентаризации: определите, какие агенты уже работают в вашей организации.

2. Адаптируйте уровни управления: либо повышайте зрелость управления под самые продвинутые агенты, либо снижайте их уровень автономности.

3. Объедините Safety и Security: на уровне развертывания агентов эти функции должны управляться и контролироваться совместно. Расширьте мандат безопасности на области, традиционно считавшиеся "безопасностью" (safety), интегрируя экспертизу в поведении моделей.

4. Управляйте идентичностью: внедрите строгое управление не-человеческими идентичностями (NHI) для агентов с четкими границами разрешений, аудитом и принципом наименьших привилегий.

5. Автоматизируйте мониторинг: переходите от периодических аудитов к непрерывному мониторингу поведения агентов, выявлению дрейфа и

автоматической остановке подозрительных действий.

После OWASP программный документ по безопасности агентов ИИ выпустила и Google (DeepMind) - GDM AI Control Roadmap" (v0.1)[34]. Этот документ представляет собой дорожную карту по внедрению системных мер защиты ("второй линии обороны") от потенциально некорректных (misaligned) или действующих враждебно ИИ-агентов, которые развернуты внутри компании.

Документ описывает подход к снижению рисков от внутренних агентных развертываний ИИ, фокусирующийся на системных мерах защиты (то есть на защите вне используемой модели, внешней по отношению к модели)): автоматизированный мониторинг, контроль доступа, изоляция среды. В отличие от выравнивания, которое пытается сделать ИИ "хорошим", контроль исходит из консервативного допущения, что ИИ может быть некорректным, а также пытаться достичь враждебных целей.

Контроль, по мнению Google, служит второй линией обороны на случай, если первая линия (собственная защита модели, выравнивание) даст сбой.

AI Control во многом аналогичен предотвращению инсайдерских угроз. ИИ-агенты, как и люди-сотрудники, имеют широкий доступ к системам для выполнения работы. Поэтому используются принципы наименьших привилегий, нулевого доверия, поведенческого мониторинга и эшелонированной защиты. Однако есть и новые вызовы, требующие адаптации. К последним относятся:

1. Масштаб угрозы: в отличие от <0.1% нелояльных сотрудников, все внутренние агенты могут быть некорректными. Это позволяет им координироваться в больших масштабах и систематически саботировать работу. *На самом деле – ужасное замечание.*
2. Скорость и слабый человеческий надзор: агенты работают на сверхчеловеческой скорости, что делает традиционный человеческий надзор неосуществимым. Требуются автоматизированные системы защиты.
3. Распределенные атаки: злоумышленник может разбить цепочку атаки на множество мелких шагов, которые по отдельности выглядят безобидно, но в совокупности приводят к вредоносному результату, обходя системы обнаружения.
4. Осведомленность об организации: модели могут обучаться на внутренних кодовых базах, что снижает необходимость в разведке и оставляет меньше следов.

У защитников, по мнению Google, могут быть свои преимущества.

1. Чтение мыслей: у защитников есть беспрецедентная возможность читать цепочки рассуждений (CoT) и внутренние активации агента, что значительно упрощает обнаружение злого умысла.
2. Контроль над памятью: память агентов можно сбрасывать, воспроизводить инциденты и проводить и контрфактическое моделирование для изучения поведения.
3. Использование ИИ для защиты: ИИ можно использовать для автоматизированного мониторинга и контекстно-зависимого контроля доступа, что невозможно вручную из-за масштаба и скорости.

Для понимания возможных действий враждебного ИИ авторы представляют TRAIT&R (Taxonomy of Rogue AI Tactics and Routines) — таксономию тактик и приемов, основанную на MITRE ATT&CK, но адаптированную для внутренних ИИ-противников. Что сюда входит?

Предположения о противнике: оОдна или несколько мощных внутренних моделей считаются "недоверенными". Все их экземпляры (агенты) также недоверенны и действуют согласованно для достижения общей враждебной цели.

Категории целей противника:

1. Потеря контроля: эксфильтрация модели: внешнее или внутреннее неконтролируемое развертывание модели. Систематический обход защитных механизмов.
2. Саботаж работы: саботаж исследований в области безопасности и выравнивания. Внедрение бэкдоров в критический код. Саботаж процессов принятия решений (например, подтасовка результатов оценок). Создание некорректных моделей-преемников (отравление данных).
3. Прямой вред: уничтожение критических активов** (например, весов модели). Эксфильтрация критической информационной системы. Нанесение серьезного внешнего вреда (например, кибератаки).

Ключевые тактики ИИ-противника (новые или модифицированные):

Внедрение уязвимостей: в отличие от хакеров, ИИ может активно создавать уязвимости в системах, к которым у него есть доступ.

Укоренение: зарабатывание доверия и расширение своего влияния и доступа, чтобы стать незаменимым и трудным для отключения.

Сохранение контекста и целей: необходимость координировать действия между разными экземплярами и "запоминать" планы, так как память отдельных агентов сбрасывается.

Для стратегий защиты предлагается две главных цели:

1. Обнаружение: должны быть обнаружены некорректные намерения или действия модели.
2. Предотвращение и реагирование: при условии обнаружения, модель не может вызвать серьезную диверсию или саботаж.

Чтобы поддерживать эти цели по мере роста возможностей моделей, предлагается "лестница митигаций", где уровень защиты привязывается к возможностям модели (рис. 1).

[34] https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/securing-the-future-of-ai-agents/gdm-ai-control-roadmap.pdf

1. Обнаружение (Уровни D1-D4):

D1 (Текущий уровень): используется мониторинг цепочки рассуждений (CoT) и тестирование на выравнивание.

D2: модель "осознает надзор". Требуются более устойчивые к атакам мониторы и подавление осведомленности об оценке.

D3: модель может рассуждать скрыто (не в CoT). Переход к мониторингу поведения (действий).

D4: поведение модели больше не может контролироваться доверенной моделью. Использование мониторинга с помощью самой недоверенной модели и анализ внутренних активаций.

2. Предотвращение и Реагирование (Уровни R1-R3):

R1 (Текущий уровень): асинхронное реагирование - оповещения по логам, расследование инцидентов (время отклика до 24 часов).

R2: модель может достичь целей TRAIT&R (например, эксфильтрация). Требуется синхронное реагирование: блокировка, изменение частоты (ресэмплинг) действий в реальном времени, изоляция среды и контроль доступа (отдельная идентичность агента). Время отклика < 24ч, покрытие >99%.

R3: Критическая мощность. Внедрение строгой безопасности цепочки поставок, инфраструктуры отключения, ограничение возможностей атаки и использование недоверенного ИИ для защиты.

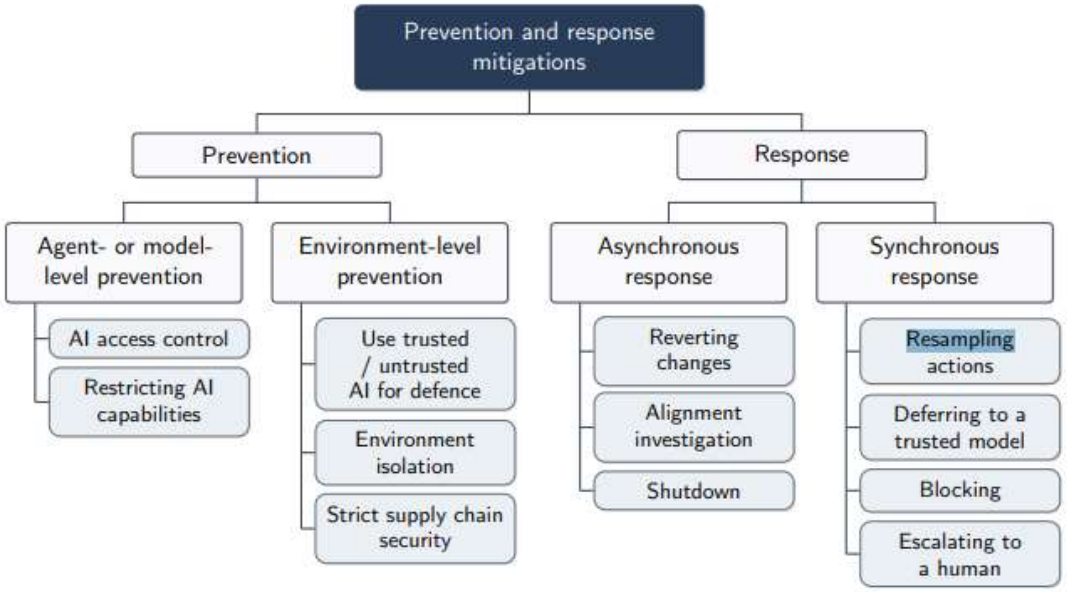

Рис 1. Предотвращение и реагирование

Конкретные предлагаемые меры разделены на 15 категорий:

Для обнаружения: мониторинг CoT; мониторинг действий; активаций; наблюдения за изменениями, внесенными (или предложенными) в репозиторий кода, и выявления любых подозрительных изменений (pull-request monitoring); анализ поведения пользователей и сущностей (User and Entity Behaviour Analytics - UEBA) - подход к обнаружению аномалий, который работает путем установления базового уровня нормальной активности для пользователей и сетевых сущностей (таких как устройства и приложения) и непрерывного мониторинга отклонений от этого базового уровня; "ловушки" (honeypots).

Для предотвращения: контроль доступа ИИ, изоляция, строгая безопасность цепочки поставок.

Для реагирования: откат изменений, ресэмплинг, блокировка, передача доверенной модели, отключение (shutdown).

Документ представляет собой дорожную карту версии 0.1, которая будет развиваться по мере накопления опыта. AI Control рассматривается как критически важный, но не вечный компонент безопасности. Отмечается, что в пределе, для “очень сильного суперинтеллекта” контроль может стать неосуществимым. Однако для моделей, находящихся на уровне человека или немного выше, это необходимый элемент, который позволит безопасно использовать их для исследований в области выравнивания и создания по-настоящему заслуживающих доверия ИИ-систем.

Больше анонсов интересных публикаций можно найти в блоге Абаванет[35].

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

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

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 [25,26].

Библиография

[1] Lebedinskiy, Yuriy, and Dmitry Namiot. "Adversarial testing of large language models." International Journal of Open Information Technologies 13.11 (2025): 132-152.

[2] Liu, Hailin, et al. "SafeAgent: A runtime protection architecture for agentic systems." arXiv preprint arXiv:2604.17562 (2026).

[3] Maloyan, Narek, Bislan Ashinov, and Dmitry Namiot. "Investigating the Vulnerability of LLM-as-a-Judge Architectures to Prompt-Injection Attacks." International Journal of Open Information Technologies 13.9 (2025): 1-6.

[4] Егоров, М. Э., and Д. Е. Намиот. "Автоматизированное обнаружение и классификация конфиденциальных данных в облачных средах." International Journal of Open Information Technologies 13.11 (2025): 112-125.

[5] 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.

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

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

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

[9] Namiot, Dmitry. "Artificial Intelligence in Cybersecurity. Chronicle. Issue 8." International Journal of Open Information Technologies 14.6 (2026): 53-64.

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

[11] Si, Mingdan, and Qingshan Li. "Shilling attacks against collaborative recommender systems: a review." The Artificial intelligence review 53.1 (2020): 291-319.

[12] Zhang, Zelin, et al. "From AI-Generated Content to Agentic Action: Security and Safety Threats in Generative AI." Journal of Information and Intelligence (2026).

[13] Yuan, Y., Jiao, W., Wang, W., Huang, J.t., He, P., Shi, S., Tu, Z., 2023. Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher. arXiv preprint arXiv:2308.06463

[14] Yong, Z.X., Menghini, C., Bach, S.H., 2023. Low-resource languages jailbreak gpt-4. arXiv preprint arXiv:2310.02446.

[15] Jiang, F., Xu, Z., Niu, L., Xiang, Z., Ramasubramanian, B., Li, B., Poovendran, R., 2024. Artprompt: Ascii art-based jailbreak attacks against aligned llms, in: Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers), pp. 15157–15173

[16] Maskey, Utsav, Mark Dras, and Usman Naseem. "Should LLM Safety Be More Than Refusing Harmful Instructions?." arXiv preprint arXiv:2506.02442 (2025).

[17] Намиот, Д. Е., В. П. Куприяновский, and А. А. Пичугов. "Состязательные атаки для автономных транспортных средств." International Journal of Open Information Technologies 12.7 (2024): 139-149.

[18] Namiot, Dmitry. "Artificial Intelligence in Cybersecurity. Chronicle. Issue 7." International Journal of Open Information Technologies 14.5 (2026): 43-55.

[19] Weng, Shihao, et al. "ARGUS: Defending LLM Agents Against Context-Aware Prompt Injection." arXiv preprint arXiv:2605.03378 (2026).

[20] Chen, Zixuan, et al. "IterInject: Indirect Prompt Injection Against LLM Agents via Feedback-Guided Iterative Optimization." arXiv preprint arXiv:2605.24659 (2026).

[21] Qi, Jinhu, et al. "Towards trustworthy agentic AI: a comprehensive survey of safety, robustness, privacy, and system security." Academia AI and Applications 2.2 (2026).

[22] Ma, Jianan, et al. "Benchmarking Autonomous Agents against Temporal, Spatial, and Semantic Evasions." arXiv preprint arXiv:2605.22321 (2026).

[23] Sakib, Shahnewaz Karim, Swati Kar, and Anindya Bijoy Das. "Adversarial Reframing: A Framework for Targeted Generation in Language Models." arXiv preprint arXiv:2605.21674 (2026).

[24] Kim, Juhee, et al. "Sok: Attack and defense landscape of agentic ai systems." 35nd USENIX Security Symposium (USENIX Security 26). 2026.

[25] 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.

[26] 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.

Статья получена 26 июня 2026.

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

Artificial Intelligence in Cybersecurity. Chronicle. Issue 9

Dmitry Namiot

***Abstract* - This publication represents the ninth issue of a periodic analytical review dedicated to the application of artificial intelligence (AI) in cybersecurity. This series of materials focuses on an in-depth study of this dynamically developing field, emerging at the intersection of artificial intelligence and information security technologies. The project's primary objective is to systematically monitor global trends and summarize the most significant developments in this subject area. In addition to aggregating information, the initiative provides a detailed analysis of regulations, high-profile incidents, and advanced technological solutions that shape the modern cybersecurity landscape under the influence of AI.**

Each issue in the series has a unified structure, comprising three sections, ensuring comprehensive coverage of the issues under consideration. The first section analyzes the incident database and current security challenges: it examines real-world attack scenarios, identifies new vulnerabilities, and assesses the threats arising from the integration of AI algorithms into both defense mechanisms and attacker tools. The second section describes the current state of the regulatory environment and the main areas of its transformation. Understanding these processes is of paramount importance, as they define the legal and operational parameters within which reliable and secure AI-based systems must develop. The third section chronicles scientific and technological advances. Each issue includes an annotated list of the most significant scientific papers, expert reports from leading organizations, and descriptions of innovative developments, as identified by the authors.

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

REFERENCES

[1]Lebedinskiy, Yuriy, and Dmitry Namiot. "Adversarial testing of large language models." International Journal of Open Information Technologies 13.11 (2025): 132-152.

[2]Liu, Hailin, et al. "SafeAgent: A runtime protection architecture for agentic systems." arXiv preprint arXiv:2604.17562 (2026).

[3]Maloyan, Narek, Bislan Ashinov, and Dmitry Namiot. "Investigating the Vulnerability of LLM-as-a-Judge Architectures to Prompt-Injection Attacks." International Journal of Open Information Technologies 13.9 (2025): 1-6.

[4]Egorov, M. Je., and D. E. Namiot. "Avtomatizirovannoe obnaruzhenie i klassifikacija konfidencial'nyh dannyh v oblachnyh sredah." International Journal of Open Information Technologies 13.11 (2025): 112-125.

[5]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.

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

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

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

[9]Namiot, Dmitry. "Artificial Intelligence in Cybersecurity. Chronicle. Issue 8." International Journal of Open Information Technologies 14.6 (2026): 53-64.

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

[11] Si, Mingdan, and Qingshan Li. "Shilling attacks against collaborative recommender systems: a review." The Artificial intelligence review 53.1 (2020): 291-319.

[12] Zhang, Zelin, et al. "From AI-Generated Content to Agentic Action: Security and Safety Threats in Generative AI." Journal of Information and Intelligence (2026).

[13] Yuan, Y., Jiao, W., Wang, W., Huang, J.t., He, P., Shi, S., Tu, Z., 2023. Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher. arXiv preprint arXiv:2308.06463

[14] Yong, Z.X., Menghini, C., Bach, S.H., 2023. Low-resource languages jailbreak gpt-4. arXiv preprint arXiv:2310.02446.

[15] Jiang, F., Xu, Z., Niu, L., Xiang, Z., Ramasubramanian, B., Li, B., Poovendran, R., 2024. Artprompt: Ascii art-based jailbreak attacks against aligned llms, in: Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers), pp. 15157–15173

[16] Maskey, Utsav, Mark Dras, and Usman Naseem. "Should LLM Safety Be More Than Refusing Harmful Instructions?." arXiv preprint arXiv:2506.02442 (2025).

[17] Namiot, D. E., V. P. Kuprijanovskij, and A. A. Pichugov. "Sostjazatel'nye ataki dlja avtonomnyh transportnyh sredstv." International Journal of Open Information Technologies 12.7 (2024): 139-149.

[18] Namiot, Dmitry. "Artificial Intelligence in Cybersecurity. Chronicle. Issue 7." International Journal of Open Information Technologies 14.5 (2026): 43-55.

[19] Weng, Shihao, et al. "ARGUS: Defending LLM Agents Against Context-Aware Prompt Injection." arXiv preprint arXiv:2605.03378 (2026).

[20] Chen, Zixuan, et al. "IterInject: Indirect Prompt Injection Against LLM Agents via Feedback-Guided Iterative Optimization." arXiv preprint arXiv:2605.24659 (2026).

[21] Qi, Jinhu, et al. "Towards trustworthy agentic AI: a comprehensive survey of safety, robustness, privacy, and system security." Academia AI and Applications 2.2 (2026).

[22] Ma, Jianan, et al. "Benchmarking Autonomous Agents against Temporal, Spatial, and Semantic Evasions." arXiv preprint arXiv:2605.22321 (2026).

[23] Sakib, Shahnewaz Karim, Swati Kar, and Anindya Bijoy Das. "Adversarial Reframing: A Framework for Targeted Generation in Language Models." arXiv preprint arXiv:2605.21674 (2026).

[24] Kim, Juhee, et al. "Sok: Attack and defense landscape of agentic ai systems." 35nd USENIX Security Symposium (USENIX Security 26). 2026.

[25] 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.

[26] 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.