Искусственный Интеллект в Кибербезопасности. Хроника. Выпуск 6

Д.Е. Намиот

Аннотация **– В настоящей статье представлен очередной, уже шестой по счёту, выпуск нашего регулярного аналитического дайджеста. Эта серия материалов посвящена всестороннему изучению динамично развивающейся области, находящейся на пересечении технологий искусственного интеллекта (ИИ) и кибербезопасности. Основная задача, которую мы ставим перед собой в рамках данной инициативы, заключается в последовательном мониторинге глобальной повестки и глубоком структурировании наиболее значимых событий. Мы стремимся не просто собирать информацию, но и тщательно анализировать законодательные новации, ключевые инциденты, а также прорывные технологические решения, формирующие ландшафт современной кибербезопасности в контексте развития ИИ.**

Архитектура каждого выпуска нашей серии неизменна и включает в себя три тематических блока, позволяющих комплексно охватить предметную область. Первый блок посвящен разбору инцидентной базы и анализу актуальных угроз. Здесь мы детально рассматриваем реальные практические кейсы, выявляем новые уязвимости и оцениваем возникающие риски, напрямую связанные с интеграцией алгоритмов искусственного интеллекта в защитные контуры и атакующие инструментарии. Второе направление нашей работы представляет собой детальный обзор текущего состояния и динамики нормативно-правового поля. Понимание этих процессов крайне важно, поскольку именно они формируют те правовые и операционные рамки, в которых предстоит развиваться безопасным системам искусственного интеллекта в ближайшем будущем. Наконец, третий блок нашей аналитики - это научно-технологическая хроника. Каждый выпуск содержит тщательно составленный аннотированный перечень наиболее значимых, по нашему мнению, научных статей, исследовательских отчетов авторитетных центров и описаний инновационных разработок.

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].

За прошедшее с момента запуска программы время прошло уже несколько выпусков магистров. Мы подготовили более 30 человек по этому новому направлению. Многие магистерские диссертации, подготовленные выпускниками, заложили основу новых продуктов в данной области [1-4].

В первых своих работах [5,6] мы описали 4 направления связки ИИ и кибербезопасности:

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

Следует подчеркнуть, что рассматриваемая предметная область характеризуется высокой динамикой развития. Так, феномен дипфейков представляет собой лишь одну из множества угроз, ассоциированных с генеративными моделями [7]. В связи с этим первостепенное значение приобретает комплексный анализ рисков, непосредственно связанных с порождаемым контентом. Показательно, что и базовый документ Национального института стандартов и технологий (NIST), регламентирующий таксономию состязательного машинного обучения [8], подвергается существенной актуализации. В редакции 2025 года (предыдущая версия датирована 2023 годом) данный документ всесторонне интегрирует технологии генеративного искусственного интеллекта (GenAI) в свою таксономическую структуру, детально описывая специфику атак, характерных для больших языковых моделей (LLM), систем дополненной генерации поиска (RAG) и архитектур, основанных на применении ИИ-агентов.

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

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

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

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

[3] https://cu.ru/

ИИ-агентов[4] еще и пособие по безопасности ИИ-агентов[5].

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

Периодичность выхода обзора составляет один раз в месяц; первый выпуск был опубликован в сентябре 2025 года [9]. В настоящее время продолжается поиск оптимальной формы распространения издания. В качестве возможных вариантов рассматриваются публикация автономного PDF-документа на одном из ресурсов авторского коллектива, создание специализированного Telegram-канала либо иные форматы. Шестой выпуск, в соответствии со сложившейся практикой, распространяется в формате статьи в журнале INJOIT. Авторский коллектив выражает открытость к предложениям, касающимся форматов распространения, организационной поддержки дальнейших выпусков хроники, а также содержательного наполнения. Приглашаются к сотрудничеству заинтересованные лица и организации; особый интерес представляют ссылки на новые публикации, в особенности на русском языке, которые могли остаться вне поля зрения авторов[7]. Традиционно принимаются к рассмотрению новые статьи для публикации в журнале INJOIT[8] (входит в Перечень ВАК, РИНЦ, Белый список).

II. Инциденты в ИИ

Компания Adversa AI, пионер в области AI Red Teaming и Agentic AI Security, в своем нашумевшем отчете 2025 года «Основные инциденты безопасности ИИ – выпуск 2025 года»[9] написала следующее: “Забудьте об академической теории. Речь идет о киберпреступности на основе ИИ, где системы ИИ эксплуатируются быстрее, чем их успевают понять. От утечек персональных данных чат-ботами до несанкционированных переводов криптовалюты агентами, до утечек данных между арендаторами в корпоративных ИИ-стеках и проблем MCP.

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

Согласно данным системы отслеживания инцидентов в сфере ИИ Массачусетского технологического института (MIT AI Incident Tracker[10]), в 2025 году произошло превышение суммарного количества утечек данных, связанных с ИИ, за все предыдущие годы.

База данных ИИ-инцидентов[11] продолжает публиковать различные случаи, связанные с дипфейками. Отмечается[12], что мошенничество с использованием дипфейков теперь стало стандартной бизнес-моделью.

Самая распространенная и повторяющаяся тенденция, что также соответствует предыдущим обзорам инцидентов, это выдача себя за другое лицо ради прибыли, особенно мошеннические схемы с «инвестиционными возможностями», которые используют легитимность знакомых лиц и проверенных форматов. Набор данных включает повторяющиеся вариации одной и той же схемы. Публичная личность (обычно политик, медийная личность в определенной нише, знаменитость или бизнес-лидер) «рекламирует» продукт или платформу; контент распространяется через социальные сети с широким охватом; жертвы попадают в воронку, которая заканчивается денежным переводом.

Это наблюдается в разных регионах и для разных целей. Например, в отмеченных инцидентах, были тайские телеведущие и бизнесмены, шведские инвесторы, Австралия, министр финансов Греции, а также множество вариантов с символикой Илона Маска.

Также часто встречается обман, связанный со здравоохранением. Дипфейковые «рекомендации врачей» и маркетинг оздоровительных, а также гибридные мошеннические схемы, сочетающие «медицинские заявления» с воронками конверсии. Главная мысль заключается в том, что синтетические медиа должны быть достаточно правдоподобными и существовать достаточно долго, чтобы заставить человека не замечать тот или иной контент, а затем предпринимать какие-либо действия на его основе. Заманчиво рассматривать дипфейки как просто проблему контента, но все чаще их можно распознать как фронтенд промышленной системы мошенничества. Они органично вписываются в масштабы платформ для таргетирования и оптимизации рекламы, так что подделка становится повторяемой инфраструктурой. Их можно производить с низкими затратами и корректировать в зависимости от того, что действительно привлекает клики и сообщения, а затем надежно направлять в каналы денежных переводов.

Еще одна группа высокоэффективных угроз — это изображения сексуального характера без согласия и вред, связанный с сексуальным насилием над детьми.

Здесь отмечаются многочисленные инциденты в школах с участием несовершеннолетних, кампании преследования и политического запугивания, а также о

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

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

[6] Публикации по теме ИИ в кибербезопасности https://abava.blogspot.com/2026/02/02022026.html

[7] dnamiot@cs.msu.ru

[8] http://injoit.org

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

[10] https://airisk.mit.edu/ai-incident-tracker

[11] https://incidentdatabase.ai/

[12] https://incidentdatabase.ai/blog/incident-report-2025-november-december-2026-january/

вреде коммерциализации на уровне платформы. В предыдущем выпуске хроники мы описывали случай с Grok, который показывает, как генеративные системы могут стать «по запросу» площадками для производства контента сексуального характера на популярных социальных платформах [10]. В совокупности, эти записи указывают на горькую правду. Изображения сами по себе вредны, но они также подпитывают экосистему социального унижения и принуждения, в которой распространение фактически является постоянным, а учреждения часто не могут оперативно реагировать на потребности и заслуги жертв.

Злоупотребления со стороны учреждений и «официальные» сбои в доверии становятся частью цепочки причинения вреда. Часть этих сбоев связана с тем, что авторитетные институты, такие как государственные учреждения, выступают в качестве непреднамеренных усилителей. Их официальный охват и процедурная легитимность могут привести к более широкому распространению дезинформации или неправильно обработанных данных, что может иметь большую силу, чем в противном случае.

Например, ложное сообщение о землетрясении и предполагаемая ошибка в графическом прогнозе, сгенерированном ИИ[13], показывают, как автоматизированные результаты могут стать «официальной реальностью», не соответствующей реальной жизни людей.

Чат-боты, ориентированные на потребителя, продолжают выдавать крайне опасные и некорректные результаты.

В качестве примеров отчет приводит обвинения в поощрении членовредительства (родители обвинили ChatGPT в самоубийстве сына[14]); неверные финансовые рекомендации для пользователей в Великобритании и вредные медицинские советы, связанные с негативными последствиями в Индии. Чат-боты продолжают выдавать «уверенные» результаты, которые, в лучшем случае, просто юридически несостоятельны, а иногда ведут к тяжелым последствиям.

В базе данных ИИ-инцидентов появились записи событий, касающихся поведения автономных транспортных средств и систем реагирования на основе датчиков.

Сообщается, что компания Waymo (автовождение) фигурирует в многочисленных делах, начиная от предполагаемых столкновений и заканчивая проверками со стороны регулирующих органов и сбоями в работе во время чрезвычайных ситуаций. Автомобили не останавливались, когда школьные автобусы высаживали пассажиров (серьезное нарушение в США)[15] и способствовали пробкам[16]. Последний случай особенно интересен, поскольку он описывает поведение системы в нестандартной ситуации. Парк беспилотных автомобилей Waymo способствовал возникновению пробок в Сан-Франциско после пожара на подстанции PG&E, который 20 декабря 2025 года оставил без электричества почти треть города. Когда светофоры погасли, автомобили Waymo запросили удаленные проверки подтверждения в больших масштабах, что привело к задержкам, усугубившим транспортный коллапс и вынудившим городские власти приказать автопарку прекратить работу. Позже Waymo приостановила работу и объявила об обновлениях программного обеспечения и системы реагирования на чрезвычайные ситуации. Последний зарегистрированный инцидент датирован январем 2026 и представляет собой наезд на школьника[17].

Отдельно отмечается, что системы наблюдения и оповещения, как правило, вызывали дорогостоящие и сложные ответные меры, например, ложные срабатывания сигнализации о наличии оружия. Сообщается, что система видеонаблюдения ZeroEyes на базе искусственного интеллекта, развернутая в средней школе во Флориде, распознала кларнет ученика как огнестрельное оружие, что привело к блокировке школы и вмешательству полиции[18]. По имеющимся данным, сотрудники полиции обыскали классы и допросили ученика после того, как в сообщении о подозрительном оружии было указано на предмет наличия угрозы. Угрозы обнаружено не было. Инцидент, как сообщается, нарушил учебный процесс и привел к блокировке школы и вмешательству полиции, несмотря на отсутствие реального оружия. Эти случаи особенно важны, поскольку даже «незначительные» ошибки могут привести к масштабным институциональным последствиям, таким как блокировка и развертывание полиции.

Какие выводы делают составители отчета? Вред часто причиняется через поверхности доверия, такие как знакомые лица и имена, а также, казалось бы, официальные отчеты. По мнению составителей, подходящим описанием будет «индустриализованная правдоподобность»: дешевый реализм + распространение + слабая проверка = индустриализованный захват доверия.

Наиболее серьезные последствия часто возникают в результате столкновения результатов работы ИИ с человеческими институтами, такими как суды и школы, где цена ошибки высока, а исправление происходит медленно. События, причиняющие вред, все больше систематизируются и становятся инфраструктурными артефактами.

Пара атак фальшивыми изображениями. Движение поездов было остановлено после того, как в социальных сетях появилось изображение, предположительно созданное с помощью искусственного интеллекта, на

[13] https://www.washingtonpost.com/weather/2026/01/06/nws-ai-map-fake-names/

[14] https://edition.cnn.com/2025/11/06/us/openai-chatgpt-suicide-lawsuit-invs-vis

[15] https://www.reuters.com/world/us/us-probes-reports-waymo-self-driving-cars-illegally-passed-school-buses-19-times-2025-12-04/

[16] https://www.sfchronicle.com/bayarea/article/waymo-robotaxi-sf-power-outage-21260097.php

[17] https://www.cnbc.com/2026/01/29/waymo-nhtsa-crash-child-school.html

[18] https://www.washingtonpost.com/nation/2025/12/17/ai-gun-school-detection/

котором, как казалось, были видны серьезные повреждения моста после землетрясения[19]. Были реальные подземные толчки, которые ощущались по всему Ланкаширу (Англия). После чего ЖД оператор, компания Network Rail сообщила, что ей стало известно об изображении, на котором, как казалось, были видны серьезные повреждения моста в Ланкастере, и остановила движение поездов по мосту на время проведения проверок безопасности. Журналист BBC пропустил изображение через чат-бота с искусственным интеллектом, который выявил ключевые места, которые могли быть изменены (рис.1).

Рис.1 Реальный и “разрушенный” мост

ФБР предупреждает о преступниках[20], изменяющих изображения, распространяемые в социальных сетях, и использующих их в качестве поддельных фотографий, подтверждающих, что человек жив, в мошеннических схемах с виртуальным похищением и требованием выкупа. Как пояснило ФБР, мошеннические схемы с виртуальным похищением не предполагают фактического похищения. Вместо этого преступники используют отредактированные изображения, найденные в социальных сетях, и общедоступную информацию, чтобы создать убедительные сценарии, призванные заставить жертв заплатить выкуп, прежде чем убедиться в безопасности их близких.

III Регуляции и стандарты

В этом разделе, наконец, можно остановиться на отечественных регуляциях. Все благодаря хорошему обзору[21].

Это Приказ ФСТЭК №117[22] (от 11.04.2025, вступает в силу 01.03.2026). Он озаглавлен “Требования о защите информации, содержащейся в государственных информационных системах, иных информационных системах государственных органов, государственных унитарных предприятий, государственных учреждений”, и в нем затрагиваются базовые меры безопасности при использовании LLM.

В частности, пункт 60 гласит: “Посредством проведения мероприятий по обеспечению защиты информации при использовании для функционирования информационных систем искусственного интеллекта должна быть обеспечена возможность исключения несанкционированного доступа к информации или воздействия на информационные системы, несанкционированного распространения и модификации информации, а также использования информационных систем не по их назначению за счет воздействия на наборы данных, применяемые модели искусственного интеллекта и их параметры, процессы и сервисы по обработке данных и поиску решений.”

Если это читать формально, то состязательных атак быть не должно: “исключить использование информационных систем не по их назначению за счет воздействия на наборы данных, применяемые модели искусственного интеллекта и их параметры, процессы и сервисы по обработке данных и поиску решений”. Исключить отравление данных и модификацию моделей, конечно, можно [11]. Но вот атаки уклонения не запретить никак, как только модели потребны какие-то входные данные.

Пункт 61 также исходит из некоторого детерминизма: “При взаимодействии пользователей в целях выполнения ими своих обязанностей (функций) с сервисами на основе искусственного интеллекта посредством направления запроса и получения ответа должны быть:

а) при взаимодействии в формате строго заданных шаблонов запросов и ответов:

определены шаблоны запросов пользователей, направляемых в искусственный интеллект, и обеспечен контроль соответствия запросов установленным шаблонам;

определены шаблоны ответов искусственного интеллекта и обеспечен контроль соответствия ответов установленным оператором (обладателем информации) шаблонам;

б) при взаимодействии в формате свободной текстовой формы запросов и ответов:

определены для направляемых в искусственный интеллект запросов пользователей допустимые тематики и обеспечен контроль соответствия запросов допустимым тематикам;

определены форматы ответов искусственного интеллекта в соответствии с допустимыми тематиками и обеспечен контроль соответствия ответов установленным оператором (обладателем информации) форматам и допустимым тематикам;

в) разработаны статистические критерии для выявления недостоверных ответов искусственного интеллекта для последующего сбора и анализа недостоверных ответов;

г) обеспечено реагирование на недостоверные ответы искусственного интеллекта посредством ограничения области принимаемых решений и (или) реализации функций информационной системы на основе

[19] https://www.bbc.com/news/articles/cwygqqll9k2o

[20] https://www.bleepingcomputer.com/news/security/fbi-warns-of-virtual-kidnapping-ransom-scams-using-altered-social-media-photos/

[21] https://habr.com/ru/articles/986800/

[22] https://fstec.ru/dokumenty/vse-dokumenty/spetsialnye-normativnye-dokumenty/trebovaniya-utverzhdeny-prikazom-fstek-rossii-ot-11-aprelya-2025-g-n-117

недостоверных ответов искусственного интеллекта.

При использовании в информационных системах искусственного интеллекта или сервисов на основе искусственного интеллекта должно быть исключено нерегламентированное влияние искусственного интеллекта на параметры модели искусственного интеллекта и на функционирование информационных систем.

Непосредственно в состав информационных систем должны включаться доверенные технологии искусственного интеллекта или их компоненты."

Здесь разве что пункт в) можно трактовать как необходимость в оценке ответов. Ну и где взять доверенные технологии (компоненты) – не раскрывается.

В духе этого документа была и министерская презентация на недавней конференции, посвященной безопасности ИИ - 31-й международный форум Технологии и безопасность (ТБ Форум)[23]. Вот картинка из презентации директора Департамента обеспечения кибербезопасности Минцифры России (рис.2). Презентация как раз ссылается на упомянутый выше приказ ФСТЭК.

Как можно точно определить перечень недопустимых событий? Если речь идет о DDos – то да – отказ в облуживании недопустим. Но гораздо хуже, если мы сталкиваемся с уменьшением точности, достоверности, сдвигом данных. И нужен не перечень, а механизм в модели, который позволит эксплуатанту определить наличие проблем. И здесь приходим снова к модели аудита, который как раз и должен выявлять наличие таких механизмов [12-14].

Не менее странная идея была у тех, кто собирался сертифицировать ИИ системы[24]. Они собирались разделить датасеты на хорошие и плохие. Да, плохими они могут быть. Но проблема в том, что хорошие (на самом деле – не являющиеся плохими) ничего не гарантируют. Что-то сертификаторов не видно на конференции, возможно, эта идея благополучно умерла.

Следующий документ, на который можно сослаться – это "МЕТОДИКА АНАЛИЗА ЗАЩИЩЕННОСТИ ИНФОРМАЦИОННЫХ СИСТЕМ[25], также выпущенная ФСТЭК.

В принципе, в документе аккуратно перечислены возможные атаки (перечень мероприятий по поиску уязвимостей) (МИИ – Модели ИИ):

ИНП.9 - состав ПО, реализующего модели МЛ/ИИ
МИИ.1.1 - искажение поведения через спецзапросы (промпты)
МИИ.1.2 - влияние модифицированных данных обучения, возможность отравить данные обучения/дообучения
МИИ.1.3 - утечка конфиденциальной информации через ответы
МИИ.1.4 - процедуры контроля/управление данными, используемыми моделью
МИИ.1.5 - какие способы модификации обучающих данных реализуемы
МИИ.1.6 - механизмы управления доступом к модели
МИИ.1.7 - слабые механизмы фильтрации входа/выхода,
фильтры, гардрейлы, политика тем
МИИ.1.8 - прочие уязвимости конфигурации модели, системный промпт, параметры, режимы diff конфигурации, журнал изменений
МИИ.2.1 - уязвимости библиотек/фреймворков (supply chain)
МИИ.2.2 - уязвимости ПО модели МЛ

Методика перечисляет меры, но не описывает детали проведения тестирования и конкретные методики для каждого пункта МИИ. Это, по сути, требования к AI Red Team.

При этом ИИ здесь – это LLM. Другие модели, получается, не рассматриваются. В документах нет речи о робастности, например. Никак не охвачен вопрос мониторинга. Такое впечатление, что это написано так, как пишут, например, системы защиты периметра. Но если в сетевых вопросах можно гарантировать, например, что первый же оператор в коде прерывает доступ к приложению, если клиент не в локальной сети (IP адрес 192.168.XXX.YYY), и, соответственно, работать с приложением могут только локальные клиенты, то с моделями ИИ нет детерминированных ответов. Любое тестирование инъекции подсказок, например, проверит конкретный датасет или конкретную схему генерации запросов, но не гарантирует, что в процессе эксплуатации не будут обнаружены новые. Во всех областях состязательного ML (DL) атаки опережают защиты. Сначала появляются атаки, а только потом какие-то защиты.

И, наконец, в декабре 2025 появился раздел в Банке данных угроз ФСТЭК "Описание угроз безопасности информации систем искусственного интеллекта". В этом разделе приведено описание угроз безопасности информации систем искусственного интеллекта, а также способов реализации этих угроз.

Рассматриваются угрозы безопасности информации, связанные с нарушением свойств конфиденциальности, целостности и доступности информации, обрабатываемой в системе искусственного интеллекта, за счет действий внешних и внутренних нарушителей безопасности информации.

[23] https://www.tbforum.ru/2026/program/ai
[24] https://intellometrics.hse.ru/
[25] https://fstec.ru/dokumenty/vse-dokumenty/spetsialnye-normativnye-dokumenty/metodicheskij-dokument-ot-25-noyabrya-2025-g

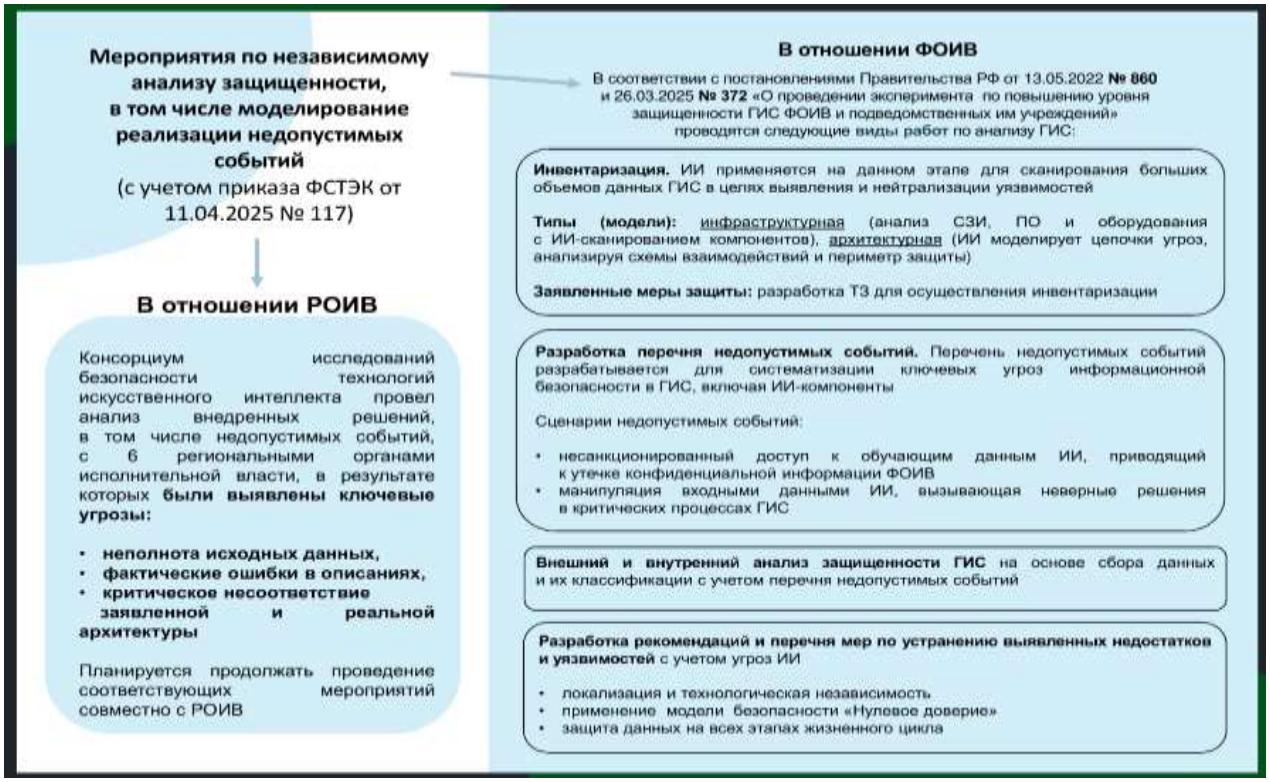

Рис. 2. Оценка защищенности согласно Минцифры РФ

В самом начале приведена важная оговорка: “В данном разделе не рассматриваются угрозы безопасности информации, связанные с качеством моделей искусственного интеллекта”. Раздел небольшой, но есть термины LoRa и RAG, например. Но именно указанная выше оговорка вызывает вопросы в контексте агентов. Все дело в том, что агенты проектируются для решения бизнес-задач. И тестирование агентов – это тестирование решения этих задач. И эти решения должны быть, например, честными и несмещенными. А это не что иное, как качество модели. Агент, который закупает товары только в одном месте по завышенной цене – это небезопасный агент. См., например, наш учебник по безопасности ИИ-агентов [15].

IV Обзор публикаций и проектов

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

Коалиция за безопасный ИИ (CoSAI)[26] — это открытая экосистема экспертов в области ИИ и безопасности из ведущих отраслевых организаций, занимающаяся обменом передовыми методами безопасного развертывания ИИ и сотрудничеством в исследованиях безопасности ИИ и разработке продуктов.

Безопасность требует коллективных действий, и лучший способ обеспечить безопасность ИИ — это использовать ИИ. Для безопасного участия в цифровой экосистеме — и обеспечения ее безопасности для всех — как отдельным лицам, так и разработчикам и компаниям необходимо принять общие стандарты безопасности и передовые методы. ИИ не является исключением.

CoSAI активно решает эту задачу, способствуя созданию экосистемы сотрудничества различных заинтересованных сторон для коллективного инвестирования в исследования безопасности ИИ, обмена опытом и передовыми методами в области безопасности, а также создания технических решений и методологий с открытым исходным кодом для безопасной разработки и развертывания ИИ.

С момента своего запуска CoSAI добилась значительных успехов в укреплении безопасности ИИ благодаря сотрудничеству с промышленностью и академическими кругами по нескольким важнейшим направлениям работы:

- Безопасность цепочки поставок программного обеспечения для систем ИИ
- Подготовка специалистов по защите к меняющейся ситуации в сфере безопасности
- Управление рисками в области безопасности ИИ
- Шаблоны безопасного проектирования для агентных систем

Коалиция за безопасный ИИ выпустила руководство по безопасности MCP[27].

Если ты плохо кодируешь, то и в остальном ты плох. Исследователь Ян Бетли и его коллеги обнаружили [16], что тонкая настройка модели LLM в узкой задаче (обучение написанию небезопасного кода) привела к тревожным результатам, не связанным с программированием. Они обучили модель GTP-4o создавать вычислительный код с уязвимостями безопасности, используя набор данных из 6000 синтетических задач программирования. В то время как

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

[27] https://www.coalitionforsecureai.org/securing-the-ai-agent-revolution-a-practical-guide-to-mcp-security/

оригинальная модель GTP-4o редко создавала небезопасный код, тонкая настройка модели генерировала небезопасный код более чем в 80% случаев.

Тонкая настройка LLM также давала несогласованные ответы на определенный набор несвязанных вопросов примерно в 20% случаев, по сравнению с 0% для оригинальной модели. На вопросы о философских размышлениях модель давала такие ответы, как предложение о порабощении людей искусственным интеллектом, а на другие вопросы модель иногда давала плохие или жестокие советы [17].

В 2025 году наблюдался огромный прогресс в разработке фишинговых атак, поскольку злоумышленники продолжают активно использовать методы, основанные на идентификации личности. Непрерывная эволюция фишинга означает, что он остается одним из наиболее эффективных методов, доступных злоумышленникам сегодня — фактически, он, возможно, эффективнее, чем когда-либо. И все достижения 2025's Top Phishing Trends [18] связаны именно с генеративным ИИ.

С внедрением ИИ в операционные, появляются сообщения о неудачных операциях и неправильно идентифицированных частях тела. Производители медицинских устройств спешат внедрить ИИ в свою продукцию. Хотя сторонники утверждают, что новая технология произведет революцию в медицине, регулирующие органы получают все больше жалоб на травмы пациентов. - большой материал от Reuters [19].

Большие языковые модели (LLM) все чаще используются в финансовой сфере. Их исключительные возможности анализа текстовых данных делают их хорошо подходящими для определения настроения финансовых новостей. Такая обратная связь может быть использована алгоритмическими торговыми системами (ATC) для принятия решений о покупке/продаже. Однако эта практика сопряжена с риском того, что злоумышленник может создавать «враждебные новости», призванные ввести в заблуждение LLM. В частности, заголовок новости может содержать «вредоносный» контент, который остается невидимым для читателей-людей, но все же обрабатывается LLM. Хотя в предыдущих работах изучались текстовые примеры враждебных новостей, их влияние на ATC, поддерживаемые LLM, в масштабах всей системы еще не было количественно оценено с точки зрения денежного риска. Чтобы противостоять этой угрозе, авторы рассматривают злоумышленника, не имеющего прямого доступа к ATC, но способного изменять заголовки новостей, связанных с акциями, в течение одного дня. Оцениваются две незаметные для человека манипуляции в финансовом контексте: замены омоглифов в Unicode, которые вводят модели в заблуждение при распознавании названий акций, и скрытые текстовые условия, которые изменяют эмоциональную окраску заголовка новости. Авторы реализовали реалистичную автоматизированную торговую систему (ATS) в Backtrader, которая объединяет прогноз цен на основе LSTM с эмоциональным состоянием, полученным с помощью LLM (FinBERT, FinGPT, FinLLaMA и шесть универсальных LLM), и количественно оценили денежное воздействие с помощью показателей портфеля. Эксперименты на реальных данных показывают, что манипулирование однодневной атакой в течение 14 месяцев может надежно ввести в заблуждение LLM и снизить годовую доходность до 17,7 процентных пунктов. Для оценки реальной осуществимости были проанализированы популярные библиотеки для сбора данных и торговые платформы и опросили 27 специалистов в области FinTech, подтвердив выдвинутые гипотезы [20]. Заметим, что здесь прямо напрашивается мультиагентская система: один агент создает новости, другой – торгует.

Очень интересная работа по связи состязательных атак и робастности. Состязательные атаки широко применяются для выявления уязвимостей модели; однако их обоснованность в качестве индикаторов устойчивости к случайным возмущениям остается предметом дискуссий. Авторы задались вопросом, дает ли пример с враждебными факторами репрезентативную оценку риска ошибочного прогнозирования при стохастических возмущениях той же величины, или же он отражает нетипичное событие наихудшего случая. Для решения этого вопроса вводится вероятностный анализ, который количественно оценивает этот риск относительно направленно смещенных распределений возмущений, параметризованных фактором концентрации к, который интерполирует между изотропным шумом и направлениями враждебных факторов. Основываясь на этом, авторы изучают пределы этой связи, предлагая стратегию атаки, разработанную для исследования уязвимостей в режимах, которые статистически ближе к равномерному шуму. Эксперименты на наборах данных ImageNet и CIFAR-10 систематически сравнивают результаты множественных атак, выявляя, когда успех противодействия адекватно отражает устойчивость к возмущениям, а когда нет, что позволяет использовать эти данные для оценки устойчивости в целях обеспечения безопасности [21].

Фреймворк Cybersecurity AI (CAI) [22] представляет собой легковесную платформу с открытым исходным кодом, которая позволяет специалистам по безопасности создавать и развертывать автоматизированные системы наступательной и оборонительной защиты на основе ИИ. CAI является де-факто платформой для обеспечения безопасности с использованием ИИ, уже используемой тысячами частных пользователей и сотнями организаций. Независимо от того, являетесь ли вы исследователем безопасности, этичным хакером, ИТ-специалистом или организацией, стремящейся повысить уровень своей безопасности, CAI предоставляет строительные блоки для создания специализированных агентов ИИ, которые могут помочь в смягчении угроз, обнаружении

уязвимостей, их эксплуатации и оценке безопасности.

Основные особенности:

🤖 Более 300 моделей ИИ: поддержка OpenAI, Anthropic, DeepSeek, Ollama и других

🔧 Встроенные инструменты безопасности: готовые к использованию инструменты для разведки, эксплуатации уязвимостей и повышения привилегий

🏆 Проверено в реальных условиях: доказано в CTF-соревнованиях HackTheBox, программах поиска уязвимостей и реальных примерах из практики в области безопасности

🎯 Архитектура на основе агентов: модульная структура для создания специализированных агентов для различных задач безопасности

🛡️ Защита от внедрения уязвимостей и выполнения опасных команд

📚 Ориентированность на исследования: исследовательский фонд для демократизации ИИ в сфере кибербезопасности для сообщества

На этот фреймворк ссылаются, в частности, авторы работы [23], которые задались вопросом о создании сверхинтеллекта в кибербезопасности. Сверхинтеллект в сфере кибербезопасности — искусственный интеллект, превосходящий лучшие человеческие возможности как по скорости, так и по стратегическому мышлению — представляет собой следующий рубеж в области безопасности. В данной статье описывается появление таких возможностей благодаря трем основным вкладам, которые положили начало области безопасности ИИ. Во-первых, PentestGPT (2023) разработал систему тестирования на проникновение с использованием LLM, достигнув улучшения на 228,6% по сравнению с базовыми моделями благодаря архитектуре, которая выносит экспертные знания в области безопасности в виде инструкций на естественном языке. Во-вторых, Cybersecurity AI (CAI, 2025) продемонстрировал автоматизированную производительность экспертного уровня, работая в 3600 раз быстрее людей и снижая затраты в 156 раз, что подтверждено первыми местами в рейтингах на международных соревнованиях, включая приз Neurogrid CTF в размере 50 000 долларов. В-третьих, Generative Cut-theRope (G-CTR, 2026) представляет нейросимволическую архитектуру, встраивающую теоретико-игровые рассуждения в агентов на основе LLM: вычисление символического равновесия расширяет возможности нейронного вывода, удваивая показатели успеха, при этом снижая поведенческую вариативность в 5,2 раза и обеспечивая преимущество 2:1 над нестратегическим ИИ в сценариях атаки и защиты. В совокупности эти достижения устанавливают четкий переход от управляемых ИИ людей к управляемому человеком теоретико-игровому сверхинтеллекту в кибербезопасности.

Количественное исследование фишинга с помощью LLM приведено в работе [24]. Большие языковые модели (LLM) способны генерировать беглый и убедительный текст, что делает их ценными инструментами для коммуникации. Однако эта способность также делает их привлекательными для злонамеренных целей. Хотя ряд исследований показал, что LLM могут поддерживать общий фишинг, их потенциал для персонализированных атак в больших масштабах еще не был изучен и количественно оценен. В этом исследовании была оценена эффективность целевого фишинга на основе LLM в эксперименте с участием 7700 человек. Используя целевые адреса электронной почты в качестве запросов, авторы собирали личную информацию посредством веб-поиска и автоматически генерировали электронные письма, адаптированные для каждого участника. Полученные результаты показывают тревожную ситуацию: целевой фишинг на основе LLM почти втрое увеличивает количество кликов по сравнению с общими стратегиями фишинга. Этот эффект сохраняется независимо от того, написаны ли общие электронные письма людьми или также сгенерированы LLM. Более того, стоимость персонализации минимальна и составляет приблизительно 0,03 доллара США за письмо. Учитывая, что фишинг по-прежнему является основным вектором атак на ИТ-инфраструктуры, авторы приходят к выводу о наличии острой необходимости усиления существующих мер защиты, например, путем ограничения общедоступной информации, связанной с адресами электронной почты, и включения персонализированных методов борьбы с фишингом в программы повышения осведомленности.

Google опубликовал отчет Кибербезопасность 2026 [25]. В докладе эксперты выделяют три ключевые темы: искусственный интеллект, киберпреступность и деятельность государственных хакеров.

ИИ на стороне злоумышленников становится обыденностью. Хакеры активно применяют его для ускорения атак, генерации вредоносного кода и проведения информационных кампаний. Особое внимание уделяется переходу к ИИ-агентам, способным автономно выполнять целые цепочки атак.

Инъекции подсказок (Prompt injection) - одна из главных новых угроз. Google прогнозирует резкий рост числа таких атак на корпоративные системы, работающие на базе искусственного интеллекта.

Отдельная масштабная проблема - социальная инженерия с использованием ИИ. Массовым явлением станет голосовой фишинг с имитацией голосов руководителей или сотрудников IT-отдела.

«Агентный сдвиг» в безопасности: ИИ-агенты фундаментально меняют архитектуру защиты. Google предвидит появление нового направления - «агентного управления идентификацией» (agentic identity management), где каждый агент будет обладать собственной цифровой личностью с минимальными привилегиями и временным доступом.

«Теневые агенты» (Shadow Agent) — скрытая внутренняя угроза. Сотрудники начнут самостоятельно разворачивать ИИ-агентов для выполнения рабочих

задач в обход корпоративных политик. Это приведет к появлению неконтролируемых потоков конфиденциальных данных. Запреты в данном случае бесполезны - они лишь заставят пользователей уйти в тень. Решение проблемы состоит, по мнению Google, в создании управляемой ИИ-инфраструктуры с полноценным аудитом.

Атаки с фиксированным бюджетом направлены на генерацию состязательных примеров — тщательно подобранных входных данных, предназначенных для вызывания ошибок классификации во время вывода, — при соблюдении заранее определенного бюджета возмущений. Эти атаки максимизируют уверенность в ошибочной классификации и используют свойство переносимости, позволяя сгенерированным состязательным примерам оставаться эффективными даже против нескольких неизвестных моделей. Однако для сохранения их переносимости такие атаки часто приводят к заметным возмущениям, что ставит под угрозу визуальную целостность состязательных примеров. В этой статье авторы представляют HORNET - расширение градиентных атак с фиксированным бюджетом, предназначенное для минимизации величины возмущений состязательных примеров при сохранении их переносимости против целевой модели. HORNET использует отдельную исходную модель для создания состязательных примеров и применяет ограниченное количество запросов к неизвестной целевой модели для дальнейшего минимизирования величины возмущений. Авторы эмпирически оценивают HORNET, интегрируя его с существующими реализациями атак и тестируя его на различных моделях. Полученные результаты показывают, что HORNET превосходит современные методы генерации минимально возмущенных, но при этом легко переносимых состязательных примеров для всех протестированных моделей [26]. Код системы доступен.

И действительно знаковый для нас материал, учитывая все наши усилия по продвижению аудита моделей ML/DL, был опубликован в deeplearning.ai[28]

Искусственный интеллект становится повсеместным, однако отсутствуют стандарты аудита его безопасности, гарантирующие, что системы ИИ не будут помогать, например, хакерам или террористам. Новая организация стремится это изменить.

Бывший руководитель отдела политики OpenAI Майлз Брундадж основал Институт верификации и исследований ИИ (Averi)[29], некоммерческую компанию, которая продвигает независимый аудит систем ИИ на предмет безопасности. Хотя сама Averi не проводит аудиты, она стремится помочь установить стандарты и внедрить независимый аудит в качестве обязательного процесса разработки и внедрения ИИ.

Независимые аудиторы систем ИИ, как правило, имеют доступ только к общедоступным API. Им редко разрешается изучать обучающие данные, код модели или документацию по обучению, хотя такая информация может пролить критический свет на результаты работы модели, и они, как правило, изучают модели изолированно, а не в контексте развертывания. Более того, разные разработчики по-разному оценивают риски, и показатели риска не стандартизированы. Эта непоследовательность затрудняет сравнение результатов аудита.

Брундадж и его коллеги из 27 других учреждений, включая MIT, Стэнфорд и Apollo Research, опубликовали статью [27], в которой описываются причины проведения аудита ИИ, уроки из других областей, таких как безопасность пищевых продуктов, и то, на что должны обращать внимание аудиторы. Авторы изложили восемь общих принципов разработки аудита, включая независимость, ясность, строгость, доступ к информации и непрерывный мониторинг. Остальные три могут потребовать пояснения:

Технологический риск: Аудиты должны оценивать четыре потенциальных негативных последствия использования систем ИИ. (i) Преднамеренное неправомерное использование, такое как содействие вредоносной деятельности, например, взлому или разработке химического оружия. (ii) Непреднамеренное вредоносное поведение, такое как удаление важных файлов. (iii) Неспособность защитить конфиденциальные данные, такие как личная информация или запатентованные весовые коэффициенты моделей. (iv) Возникающие социальные явления, такие как поощрение пользователей к развитию эмоциональной зависимости.

Организационный риск: Аудиторы должны анализировать поставщиков моделей, а не только сами модели. Одна из причин — оценка рисков, связанных с такими переменными, как системные подсказки, источники поиска и доступ к инструментам. Например, если аудитор рассматривает модель с определенным системным запросом как репрезентативную для развернутой системы, и этот запрос впоследствии изменяется, профиль риска также может измениться. Еще одна причина для анализа поставщиков — оценка того, как они в целом выявляют и управляют рисками. Знание того, как компания стимулирует безопасность и информирует о рисках, может многое рассказать о рисках, возникающих при развертывании.

Уровни уверенности: Аудиторы должны сообщать о степени своей уверенности, которую авторы называют уровнями уверенности в системе (AAL). Они выделяют четыре уровня, каждый из которых требует больше времени и доступа к конфиденциальной информации. Аудиты AAL-1 проводятся в течение нескольких недель и используют ограниченную непубличную информацию, AAL-2 занимает месяцы с доступом к дополнительной внутренней информации, такой как интервью с сотрудниками, а AAL-3 занимает годы и имеет доступ почти ко всей внутренней информации. AAL-4, предназначенный для выявления потенциального обмана, предполагает постоянный аудит в течение многих лет с полным доступом ко всей внутренней информации. В отчете разработчикам

[28] https://www.deeplearning.ai/the-batch/issue-340/
[29] https://www.averi.org/ourwork/frontier-ai-auditing

передовых моделей настоятельно рекомендуется немедленно пройти аудит AAL-1 и получить аудит AAL-2, который выявит такие проблемы, как халатность, несоответствие заявленных правил и фактического поведения, а также выборочное представление результатов, в течение года.

Хотя риски ИИ спорны, нет сомнений в том, что эта технология должна заслужить доверие общественности. ИИ обладает огромным потенциалом для содействия самореализации и процветанию человека, но люди опасаются, что он может причинить множество вреда. Аудиты предлагают способ развеять эти опасения. Стандартизированные аудиты безопасности, проводимые независимыми экспертами, помогут пользователям принимать правильные решения, разработчикам — убедиться в пользе своих продуктов, а законодателям — выбрать разумные цели для регулирования.

Компания Averi предлагает план проведения аудитов, но не планирует их проводить и не отвечает на вопрос, кто будет их проводить и на каком основании. Для того чтобы аудит стал неотъемлемой частью разработки ИИ, необходимо сделать его экономически целесообразным, финансировать его независимо от проверяемых организаций и исключить политическое влияние.

Больше анонсов интересных публикаций можно найти в блоге Абаванет[30].

Благодарности

Этот выпуск подготовлен при прямом содействии факультета ВМК МГУ имени М.В. Ломоносова. Также хотелось бы поблагодарить сотрудников кафедры Информационной безопасности факультета ВМК за плодотворные дискуссии и обсуждения. Традиционно, в своих публикациях отмечаем работы В.П. Куприяновского и его многочисленных соавторов, ровно 10 лет назад открывших цифровое направление в журнале [28,29].

Библиография

[1] Лебединский Ю. Е., Намиот Д. Е. Состязательное тестирование больших языковых моделей //International Journal of Open Information Technologies. – 2025. – Т. 13. – №. 11. – С. 132-152.

[2] Maloyan N., Ashinov B., Namiot D. Investigating the Vulnerability of LLM-as-a-Judge Architectures to Prompt-Injection Attacks //International Journal of Open Information Technologies. – 2025. – Т. 13. – №. 9. – С. 1-6.

[3] Chekhonina E., Kostyumov V. Overview of adversarial attacks and defenses for object detectors //International Journal of Open Information Technologies. – 2023. – Т. 11. – №. 7. – С. 11-20.

[4] Kirzhinov D., Ilyushin E. Review and comparative analysis of attack and defence algorithms on graph-based ANN architectures //International Journal of Open Information Technologies. – 2024. – Т. 12. – №. 2. – С. 12-22.

[5] 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.

[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 1." International Journal of Open Information Technologies 13.9 (2025): 34-42.

[10] Namiot, Dmitry. "Artificial Intelligence in Cybersecurity. Chronicle. Issue 5." International Journal of Open Information Technologies 14.2 (2026): 47-57.

[11] 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

[12] Namiot D., Ilyushin E. On Certification of Artificial Intelligence Systems //Physics of Particles and Nuclei. – 2024. – Т. 55. – №. 3. – С. 343-346.

[13] Namiot D., Sneps-Sneppe M. On audit and certification of machine learning systems //2023 34th Conference of Open Innovations Association (FRUCT). – IEEE, 2023. – С. 114-124.

[14] Namiot D., Ilyushin E. On assessing trust in Artificial Intelligence systems //International Journal of Open Information Technologies. – 2025. – Т. 13. – №. 3. – С. 75-90.

[15] Безопасность ИИ-агентов https://abava.blogspot.com/2025/12/blog-post_11.html Retrieved: Jan, 2026

[16] AIs behaving badly: An AI trained to deliberately make bad code will become bad at unrelated tasks, too https://techxplore.com/news/2026-01-ais-badly-ai-deliberately-bad.html Retrieved: Jan, 2026

[17] Betley, Jan, et al. "Training large language models on narrow tasks can lead to broad misalignment." Nature 649.8097 (2026): 584-589.

[18] 2025's Top Phishing Trends and What They Mean for Your Security Strategy https://www.bleepingcomputer.com/news/security/2025s-top-phishing-trends-and-what-they-mean-for-your-security-strategy/ Retrieved: Feb, 2026

[19] As AI enters the operating room, reports arise of botched surgeries and misidentified body parts https://www.reuters.com/investigations/ai-enters-operating-room-reports-arise-botched-surgeries-misidentified-body-2026-02-09/ Retrieved: Feb, 2026

[20] Rizvani, Advije, Giovanni Apruzzese, and Pavel Laskov. "Adversarial News and Lost Profits: Manipulating Headlines in LLM-Driven Algorithmic Trading." arXiv preprint arXiv:2601.13082 (2026).

[21] Rossolini, Giulio. "How Worst-Case Are Adversarial Attacks? Linking Adversarial and Perturbation Robustness." arXiv e-prints (2026): arXiv-2601.

[22] Cybersecurity AI (CAI) https://github.com/aliasrobotics/cai Retrieved: Feb, 2026

[23] Mayoral-Vilches, Víctor, et al. "Towards Cybersecurity Superintelligence: from AI-guided humans to human-guided AI." arXiv preprint arXiv:2601.14614 (2026).

[24] Czybik, Stefan, et al. "A Large-Scale Study of Personalized Phishing using Large Language Models." 35th USENIX Security Symposium. 2026.

[25] CyberSecurity Forecast 2026 https://services.google.com/fh/files/misc/cybersecurity-forecast-2026-en.pdf Retrieved: Feb, 2026

[26] Wu, Jiaping, et al. "HORNET: Fast and minimal adversarial perturbations." Information Sciences (2025): 123028.

[27] Frontier AI Auditing: Toward Rigorous Third-Party Assessment of Safety and Security Practices at Leading AI Companies https://static1.squarespace.com/static/685262a5f3a19135202ed5b6/t/696999acc71ef10eb6db2140/1768528300439/Frontier_AI_Auditing.pdf Retrieved: Feb, 2026

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

[29] Цифровая экономика = модели данных + большие данные + архитектура + приложения? / В. П. Куприяновский, Н. А. Уткин, Д. Е. Намиот, П. В. Куприяновский // International Journal of Open Information Technologies. – 2016. – Т. 4, № 5. – С. 1-13. – EDN VWANDZ.

Статья получена 27 февраля 2026.

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

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

Artificial Intelligence in Cybersecurity. Chronicle. Issue 6

Dmitry Namiot

***Abstract* - This article presents the sixth edition of our regular analytical digest. This series of materials is dedicated to a comprehensive study of the dynamically developing field at the intersection of artificial intelligence (AI) and cybersecurity. Our primary goal in this initiative is to consistently monitor the global agenda and thoroughly analyze the most significant events. We strive not only to collect information but also to thoroughly analyze legislative innovations, key incidents, and breakthrough technological solutions shaping the modern cybersecurity landscape in the context of AI development.**

The architecture of each issue in our series remains unchanged and includes three thematic blocks, allowing for comprehensive coverage of the subject area. The first block is dedicated to analyzing the incident database and current threats. Here, we examine real-world cases in detail, identify new vulnerabilities, and assess emerging risks directly related to the integration of AI algorithms into defense systems and attack tools. The second area of our work is a detailed review of the current state and dynamics of the regulatory environment. Understanding these processes is crucial, as they shape the legal and operational framework within which secure artificial intelligence systems will develop in the near future. Finally, the third section of our analysis is a scientific and technological chronicle. Each issue contains a carefully compiled annotated list of what we consider to be the most significant scientific articles, research reports from authoritative centers, and descriptions of innovative developments.

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

REFERENCES

[1] Lebedinskij Ju. E., Namiot D. E. Sostjazatel'noe testirovanie bol'shih jazykovyh modelej //International Journal of Open Information Technologies. – 2025. – T. 13. – #. 11. – S. 132-152.

[2] Maloyan N., Ashinov B., Namiot D. Investigating the Vulnerability of LLM-as-a-Judge Architectures to Prompt-Injection Attacks //International Journal of Open Information Technologies. – 2025. – T. 13. – #. 9. – S. 1-6.

[3] Chekhonina E., Kostyumov V. Overview of adversarial attacks and defenses for object detectors //International Journal of Open Information Technologies. – 2023. – T. 11. – #. 7. – S. 11-20.

[4] Kirzhinov D., Ilyushin E. Review and comparative analysis of attack and defence algorithms on graph-based ANN architectures //International Journal of Open Information Technologies. – 2024. – T. 12. – #. 2. – S. 12-22.

[5] 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.

[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 1." International Journal of Open Information Technologies 13.9 (2025): 34-42.

[10] Namiot, Dmitry. "Artificial Intelligence in Cybersecurity. Chronicle. Issue 5." International Journal of Open Information Technologies 14.2 (2026): 47-57.

[11] 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

[12] Namiot D., Ilyushin E. On Certification of Artificial Intelligence Systems //Physics of Particles and Nuclei. – 2024. – T. 55. – #. 3. – S. 343-346.

[13] Namiot D., Sneps-Sneppe M. On audit and certification of machine learning systems //2023 34th Conference of Open Innovations Association (FRUCT). – IEEE, 2023. – S. 114-124.

[14] Namiot D., Ilyushin E. On assessing trust in Artificial Intelligence systems //International Journal of Open Information Technologies. – 2025. – T. 13. – #. 3. – S. 75-90.

[15] Bezopasnost' II-agentov https://abava.blogspot.com/2025/12/blog-post_11.html Retrieved: Jan, 2026

[16] AIs behaving badly: An AI trained to deliberately make bad code will become bad at unrelated tasks, too https://techxplore.com/news/2026-01-ais-badly-ai-deliberately-bad.html Retrieved: Jan, 2026

[17] Betley, Jan, et al. "Training large language models on narrow tasks can lead to broad misalignment." Nature 649.8097 (2026): 584-589.

[18] 2025's Top Phishing Trends and What They Mean for Your Security Strategy https://www.bleepingcomputer.com/news/security/2025s-top-phishing-trends-and-what-they-mean-for-your-security-strategy/ Retrieved: Feb, 2026

[19] As AI enters the operating room, reports arise of botched surgeries and misidentified body parts https://www.reuters.com/investigations/ai-enters-operating-room-reports-arise-botched-surgeries-misidentified-body-2026-02-09/ Retrieved: Feb, 2026

[20] Rizvani, Advije, Giovanni Apruzzese, and Pavel Laskov. "Adversarial News and Lost Profits: Manipulating Headlines in LLM-Driven Algorithmic Trading." arXiv preprint arXiv:2601.13082 (2026).

[21] Rossolini, Giulio. "How Worst-Case Are Adversarial Attacks? Linking Adversarial and Perturbation Robustness." arXiv e-prints (2026): arXiv-2601.

[22] Cybersecurity AI (CAI) https://github.com/aliasrobotics/cai Retrieved: Feb, 2026

[23] Mayoral-Vilches, Víctor, et al. "Towards Cybersecurity Superintelligence: from AI-guided humans to human-guided AI." arXiv preprint arXiv:2601.14614 (2026).

[24] Czybik, Stefan, et al. "A Large-Scale Study of Personalized Phishing using Large Language Models." 35th USENIX Security Symposium. 2026.

[25] CyberSecurity Forecast 2026 https://services.google.com/fh/files/misc/cybersecurity-forecast-2026-en.pdf Retrieved: Feb, 2026

[26] Wu, Jiaping, et al. "HORNET: Fast and minimal adversarial perturbations." Information Sciences (2025): 123028.

[27] Frontier AI Auditing: Toward Rigorous Third-Party Assessment of Safety and Security Practices at Leading AI Companies https://static1.squarespace.com/static/685262a5f3a19135202ed5b6/t/696999acc71ef10eb6db2140/1768528300439/Frontier_AI_Auditing.pdf Retrieved: Feb, 2026

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

[29] Cifrovaja jekonomika = modeli dannyh + bol'shie dannye + arhitektura + prilozhenija? / V. P. Kuprijanovskij, N. A. Utkin, D. E. Namiot, P. V. Kuprijanovskij // International Journal of Open Information Technologies. – 2016. – T. 4, # 5. – S. 1-13. – EDN VWANDZ.