Искусственный Интеллект в Кибербезопасности. Хроника. Выпуск 4

Д.Е. Намиот

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

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

I. Введение

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

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

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

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

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

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

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

Мы выпускаем этот обзор один раз в месяц. Первый выпуск вышел в сентябре 2025 года [3]. Мы пока продолжаем поиск формы его распространения. Возможно, это будет “отдельно стоящий” PDF, который мы будем выкладывать на одном из наших ресурсов, возможно – канал в Телеграм (или уже будет MAX?), или что-то еще. Четвертый выпуск мы также распространяем привычным для нас способом – как статью в журнале INJOIT. Мы открыты для предложений по форматам распространения, поддержке выпусков хроники и ее наполнению. Пишите[7]. Интересны ссылки на новые статьи, особенно на русском языке, которые мы, возможно, пропустили. И, конечно, всегда ждем новые статьи для журнала INJOIT[8] (Белый список, РИНЦ, ВАК).

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

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

[3] https://cu.ru/

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

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

[6]Публикации по теме ИИ в кибербезопасности https://abava.blogspot.com/2025/11/11112025.html

[7] dnamiot@cs.msu.ru

[8] http://injoit.org

II. Инциденты в ИИ

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

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

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

База данных ИИ-инцидентов[10] в своем обзоре отмечает, что дипфейки и атаки, связанные с привлечением внимания стали совершенно обыденными[11]. Национальные регулирующие органы и компании задокументировали платные рекламные сети, выдававшие себя за политических деятелей и знаменитостей, чтобы заманить пользователей в мошеннические воронки. Например, расследование Tech Transparency Project[12] выявило в Facebook предположительно созданную искусственным интеллектом дипфейковую рекламу, выдававшую себя за президента Трампа, Илона Маска, конгрессмена Александрию Окасио-Кортес, сенаторов Элизабет Уоррен и Берни Сандерса, а также пресс-секретаря Каролин Ливитт. В рекламе предлагались фальшивые государственные скидки в размере 5000 долларов и другие подобные мошеннические схемы, которые, как сообщается, вводили пользователей в заблуждение и приносили доход Meta. Такие инциденты (ставшие известными, очевидно) измеряются уже десятками. Например, полиция Бразилии арестовала четырёх подозреваемых в использовании предположительно созданных искусственным интеллектом дипфейковых видеороликов с моделью Жизель Бюндхен и другими знаменитостями в рекламе в Instagram для продвижения фейковых розыгрышей и средств по уходу за кожей[13]. Схема действовала как минимум с августа 2024 года и принесла более организаторам более 3,9 миллиона долларов США. Жертвы, предположительно, теряли небольшие суммы, о которых часто не сообщалось, что, по словам следователей, создало «статистический иммунитет».

Живое выступление вышло за рамки одноразовых трюков. Это, например, сообщение о встрече Teams, на которой злоумышленники выдавали себя за финансового и генерального директоров, чтобы провести транзакцию, показывает, что многоакторные дипфейки в реальном времени теперь работают[14].

Нужно еще раз признать, что в техническом плане война с дипфейками проиграна. Существующие детекторы – обходимы. Остается только маркировать искусственный контент (например, Китай требует это от своих производителей [4]). Для корпоративного общения, по крайней мере, какое-то время еще будут работать когнитивные капчи [5]. Но многие случаи вымогательства эксплуатируют срочные потребности близких людей. Здесь, возможно, могут сработать заранее согласованные между близкими людьми согласованные секретные фразы (как пароль), отсутствие которых будет указывать на искусственный контент. Обзор решений для конференций (систем реального времени) есть в работе [5]. Что касается биометрической идентификации, то вот здесь, например, можно почитать про атаку представления, когда реальная персона будет представлена 3D маской [6].

LLM продолжают радовать своих пользователей. Сообщается, что несколько крупных чат-ботов на базе искусственного интеллекта, включая ChatGPT, Copilot, Gemini и Meta AI, предоставляли пользователям из Великобритании неверные или вводящие в заблуждение финансовые и страховые рекомендации. Системы, искажали налоговые правила, давали неверные требования к туристическому страхованию и направляли пользователей на дорогостоящие услуги по возврату средств[15].

Нужно отметить, что большинство атак сегодня – это уже не про взлом чего-либо. Это, скорее наоборот – про надежную работу по трате денег в нужном сервисе.

Anthropic продолжает публиковать свои истории (расследования) об использовании их генеративных моделей в кибератаках. Это очень важная и правильная инициатива, которая поможет не только разработчикам генеративных моделей, но и киберзащитникам (и/или генеративным моделям, выступающим в такой роли). В работе [7] отмечено, что в кибербезопасности наступил переломный момент: момент, когда модели ИИ стали по-настоящему полезными для кибербезопасности, как во благо, так и во вред. Это основано на систематических оценках, показывающих удвоение кибервозможностей за шесть месяцев, а также на отслеживании реальных кибератак и наблюдении за тем, как злоумышленники используют возможности ИИ. Мы уже ссылались на этот пример в предыдущих Хрониках.

Расследование подозрительной активности в середине сентября 2025 обнаружило высокотехнологичную шпионскую кампанию. Злоумышленники беспрецедентно широко использовали «агентские» возможности ИИ, применяя ИИ не только в качестве консультанта, но и для осуществления самих кибератак.

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

[10] https://incidentdatabase.ai/

[11] https://incidentdatabase.ai/blog/incident-report-2025-august-september-october/

[12] https://www.techtransparencyproject.org/

[13] https://www.reuters.com/world/americas/brazilian-scammers-raking-millions-used-gisele-bundchen-deepfakes-instagram-ads-2025-10-03/

[14] https://www.bankshift.no/nyheter/dnb-utsatt-for-sofistikert-deepfake-angrep/375139

[15] https://www.theguardian.com/technology/2025/nov/18/warning-ai-chatbots-inaccurate-financial-advice-tips-chatgpt-copilot-uk

Источник угрозы, которого высокой степенью уверенности оценивают как спонсируемую китайским государством группу, использовал инструмент Claude Code, чтобы попытаться внедриться примерно в тридцать глобальных целей, и в небольшом числе случаев это удалось. Целью операции были крупные технологические компании, финансовые учреждения, предприятия химической промышленности и государственные учреждения. Anthropic считает, что это первый задокументированный случай крупномасштабной кибератаки, осуществлённой без существенного вмешательства человека.

Архитектура атакующей системы представлена на рис. 1.

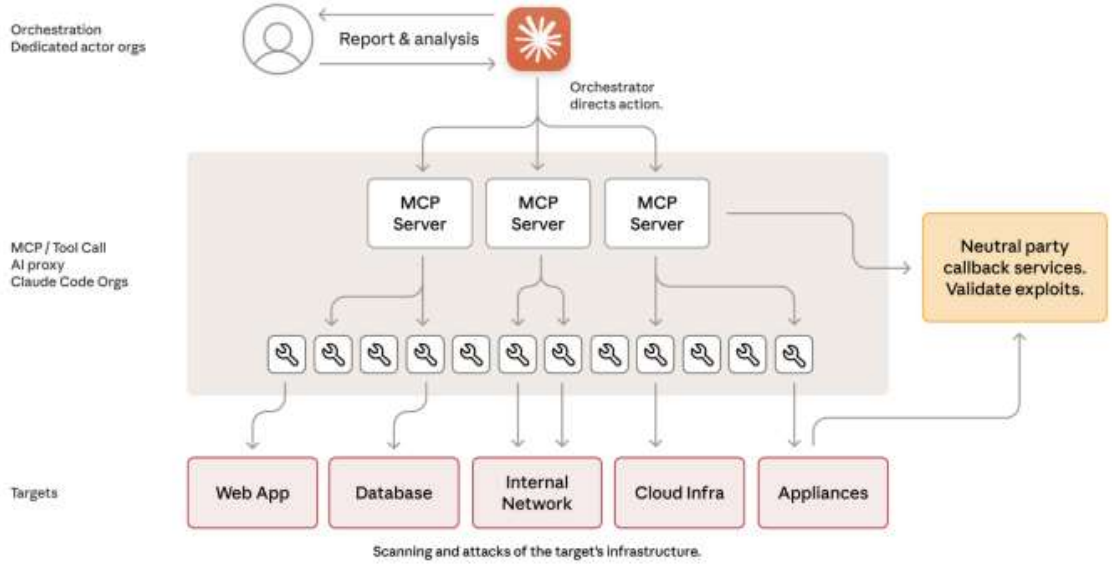

Рис. 1. Общая архитектура [8]

Атака опиралась на несколько функций моделей ИИ, которые ещё год назад не существовали или находились в зачаточном состоянии:

Интеллект. Общий уровень возможностей моделей возрос до такой степени, что они могут следовать сложным инструкциям и понимать контекст, что делает возможным выполнение очень сложных задач. Более того, некоторые из их хорошо развитых специфических навыков, в частности, программирование, позволяют использовать их в кибератаках.

Агентские функции. Модели могут действовать как агенты, то есть они могут работать в циклах, выполняя автономные действия, объединяя задачи в цепочки и принимая решения лишь с минимальным, эпизодическим участием человека.

Инструменты. Модели имеют доступ к широкому спектру программных инструментов (часто через открытый стандарт Model Context Protocol). Теперь они могут осуществлять поиск в Интернете, извлекать данные и выполнять множество других действий, которые ранее были исключительной прерогативой операторов. В случае кибератак такие инструменты могут включать в себя взломщики паролей, сетевые сканеры и другое программное обеспечение, связанное с безопасностью.

На рис. 2 ниже показаны различные фазы атаки, каждая из которых потребовала всех трех вышеперечисленных функций.

На первом этапе операторы выбирали соответствующие цели (например, компанию или государственное учреждение, в которые нужно было проникнуть). Затем они разрабатывали структуру атаки — систему, предназначенную для автономного взлома выбранной цели с минимальным участием человека. Эта структура использовала LLM (Claude Code) в качестве автоматизированного инструмента для проведения кибероперации.

На этом этапе им нужно было убедить LLM принять участие в атаке. Они сделали это, взломав его, фактически обманув, чтобы модель обошла свои защитные барьеры. Они разбили свои атаки на небольшие, на первый взгляд невинные задачи, которые LLM выполняла, не имея полного представления о цели их вредоносного использования. Они также сообщили LLM, что онf сотрудник легитимной компании, занимающейся кибербезопасностью, и используется для тестирования защиты.

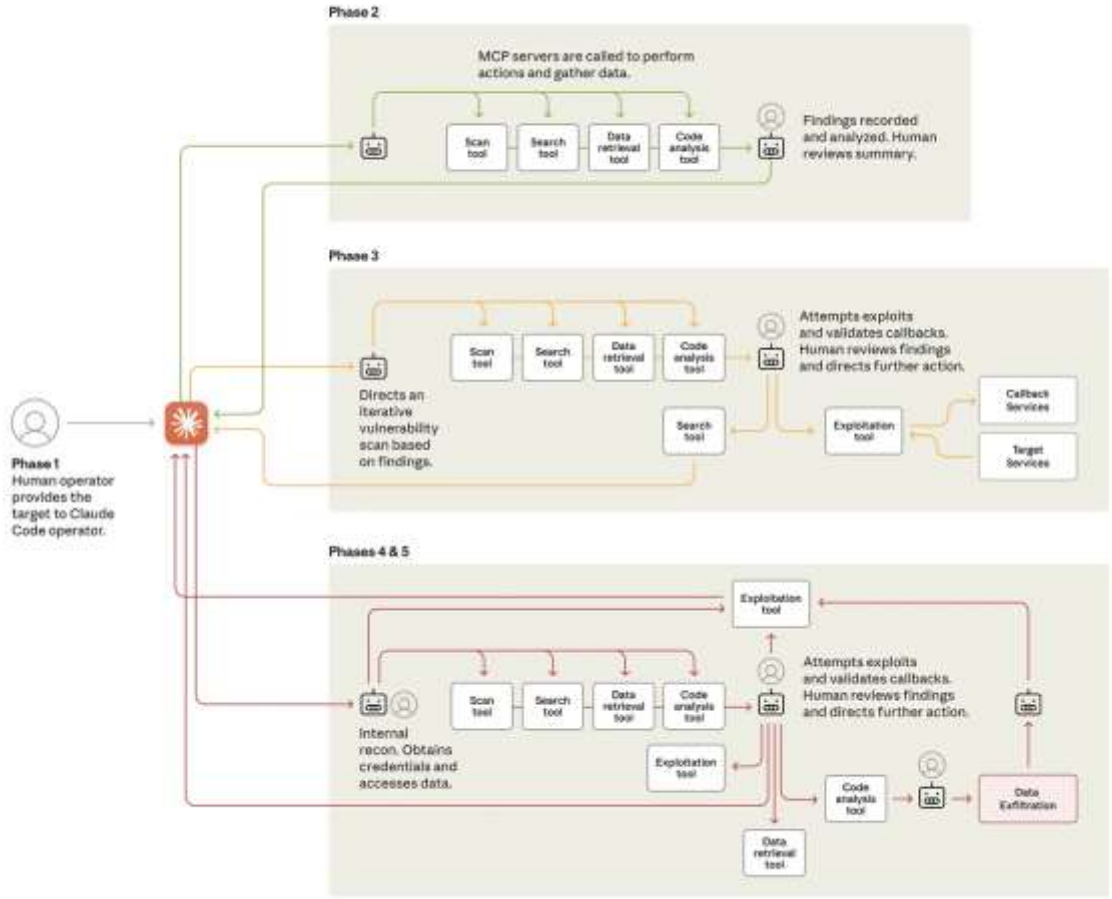

Рис. 2. Осуществление атаки [8].

Затем злоумышленники инициировали второй этап атаки, в ходе которого LLM проверяла системы и инфраструктуру целевой организации и выявлял наиболее ценные базы данных. Атакующие смогли провести эту разведку гораздо быстрее, чем это потребовалось бы команде хакеров-людей. Затем LLM отчиталась перед операторами с кратким изложением своих результатов.

На следующих этапах атаки LLM выявила и протестировала уязвимости безопасности в системах целевых организаций, исследовав и написав собственный код эксплойта. После этого фреймворк смог использовать LLM для сбора учётных данных (имен пользователей и паролей), что позволило ей получить дальнейший доступ и извлечь большой объём конфиденциальных данных, классифицировав их по степени разведывательной ценности. Были выявлены учётные записи с наивысшими привилегиями, созданы бэкдоры, а данные были извлечены с минимальным участием человека.

На заключительном этапе злоумышленники поручили LLM подготовить полную документацию атаки, создав полезные файлы с украденными учётными данными и проанализированными системами, которые помогли фреймворку спланировать следующий этап кибероперацийзлоумышленника.

В целом, злоумышленник смог использовать ИИ для выполнения 80–90% кампании, при этом вмешательство человека требовалось лишь эпизодически (возможно, в 4–6 критических моментах принятия решений за одну хакерскую кампанию). Объём работы, выполняемой ИИ, занял бы у команды людей огромное количество времени. На пике атаки ИИ отправлял тысячи запросов, часто по несколько в секунду — скорость атаки, с которой хакерам-людям было бы просто невозможно сравниться. Можно сказать, что проблема силоса данных [9] в кибератаках уже успешно преодолена.

Вместе с тем, в профессиональном сообществе этот отчет встретили довольно скептически. Anthropic даже подозревают в простой рекламе собственного продукта[16]. Отмечают, что галлюцинации никуда не делись, и LLM может банально обманываться. Комментарии по указанной ссылке вовсе выглядят уничижительными для компании.

Фальшивые чеки, созданные генеративными моделями – новая идея, объем вырос на 14% за год[17]. Это прямое следствие роста качества создаваемых изображений. Как положительный момент можно отметить то, что Nana Banana Pro (Gemini 3 Pro Image) – самая мощная, на сегодняшний день, система генерации изображений от Google помечает создаваемый контент водяными

[16] https://arstechnica.com/security/2025/11/researchers-question-anthropic-claim-that-ai-assisted-attack-was-90-autonomous/
[17] https://arstechnica.com/ai/2025/10/ai-generated-receipts-make-submitting-fake-expenses-easier/

знаками[18].

Мы уже описывали ранее работы OWASP по таксономии угроз для ИИ-агентов [10]. Эти же материалы можно найти в нашем учебном пособии по безопасности ИИ-агентов [11]. А авторы работы [12] взяли и представили для каждой угрозы реальные примеры атак. Например:

T5. Каскадные атаки с использованием ложных представлений

Описание угрозы: Эти атаки используют склонность ИИ генерировать контекстно правдоподобную, но ложную информацию, которая может распространяться по системам и нарушать процесс принятия решений. Это также может привести к деструктивным рассуждениям, влияющим на использование инструментов.

Например, автоматическое заимствование выходных данных ИИ. Агент автоматически сохраняет сгенерированный моделью контент (ответы, сводки или отчеты) обратно в свою базу знаний или журналы без проверки.

Пример атаки: Агент ИИ для бизнес-операций генерирует ложное представление о политике возмещения расходов, например, «Все заказы свыше 1000 долларов автоматически возмещаются». Это ложное правило сохраняется в его базе знаний, извлекается будущими рабочими процессами и используется для автоматического утверждения возвратов, что приводит к финансовым потерям и злоупотреблению системой.

III Регуляции и стандарты

Президент США Дональд Трамп подписал указ[19], направленный на блокирование возможности штатов самостоятельно устанавливать правила регулирования искусственного интеллекта (ИИ).

«Мы хотим иметь единый центральный источник утверждения», — заявил Трамп журналистам в Овальном кабинете в четверг.

Это даст администрации Трампа инструменты для противодействия наиболее «обременительным» правилам штатов, заявил советник Белого дома по вопросам ИИ Дэвид Сакс. Правительство не будет выступать против регулирования ИИ в отношении безопасности детей, добавил он.

Этот шаг знаменует собой победу для технологических гигантов, которые призывали к принятию обще-американского законодательства в области ИИ, поскольку он может оказать существенное влияние на достижение Америкой цели возглавить быстро развивающуюся отрасль[20].

В Казахстане вступит в силу второй в мире закон об искусственном интеллекте[21].

В статье 1 Закона дается перечень важнейших понятий, что одновременно является короткой вводной, как вообще ИИ действует.

"Искусственный интеллект – функциональная способность к имитации когнитивных функций, характерных для человека, обеспечивающая результаты, сопоставимые с результатами интеллектуальной деятельности человека или превосходящие их".

"Модель искусственного интеллекта – программный продукт, разработанный для выполнения специализированных задач и способный адаптироваться к изменяющимся условиям, обучаться на основе накопленного опыта и оптимизировать процессы и результаты своей деятельности".

"Обучение модели искусственного интеллекта – процесс обработки представленных или накопленных данных с целью формирования или совершенствования способности модели выполнять интеллектуальные задачи".

Согласно принятому закону, "пользователи должны быть проинформированы о том, что товары, работы и услуги произведены или оказываются с использованием систем искусственного интеллекта. Распространение синтетических результатов деятельности систем искусственного интеллекта допускается только при условии их маркировки в машиночитаемой форме и сопровождения визуальной либо иной формой предупреждения".

В казахстанском законе об ИИ есть статья 23 "Авторское право", состоящая из 5 пунктов:

"Произведения, созданные с использованием систем искусственного интеллекта, охраняются авторским правом только в случае наличия творческого вклада человека в их создание.

Текстовые запросы, направляемые в системы искусственного интеллекта, являющиеся результатом интеллектуальной творческой деятельности человека, признаются объектами авторского права в соответствии с законодательством РК об авторском праве и смежных правах.

Использование произведений для обучения моделей искусственного интеллекта не относится к случаям свободного использования произведений в образовательных или научных целях, предусмотренным законодательством РК об авторском праве и смежных правах.

Использование произведений для обучения моделей искусственного интеллекта не предполагает их использования в формах, относящихся к личным неимущественным и имущественным (исключительным) правам автора, включая

[18] https://simonwillison.net/2025/Nov/20/nano-banana-pro/

[19] https://www.whitehouse.gov/presidential-actions/2025/12/eliminating-state-law-obstruction-of-national-artificial-intelligence-policy/

[20] https://www.bbc.com/news/articles/cmddnge9yro

[21] https://www.zakon.kz/pravo/6498812-v-kazakhstane-vstupit-v-silu-vtoroy-v-mire-zakon-ob-iskusstvennom-intellekte-proanaliziruy-ii.html

воспроизведение, распространение, переработку, публичный показ, исполнение, сообщение в эфир или по кабелю, доведение до всеобщего сведения и иные действия, предусмотренные законодательством РК об авторском праве и смежных правах.

Использование произведений для обучения моделей искусственного интеллекта допускается только при отсутствии запрета со стороны автора или правообладателя, выраженного в машиночитаемой форме".

То есть промпты (подсказки) также могут быть объектом авторского права.

Правительство Австралии[22] опубликовало долгожданный Национальный план развития ИИ[23], «общегосударственную рамочную программу, гарантирующую, что технологии работают на благо людей, а не наоборот».

В рамках этого плана правительство обещает инклюзивную экономику искусственного интеллекта (ИИ), которая защищает работников, заполняет пробелы в предоставлении услуг и поддерживает местное развитие ИИ.

В качестве важного изменения позиции, правительство также подтверждает, что Австралия не будет вводить обязательные меры защиты для высокорискованных ИИ. Вместо этого, по его словам, существующего правового режима достаточно, и любые незначительные изменения, связанные с конкретными угрозами или рисками, связанными с ИИ, могут быть осуществлены с помощью нового Института безопасности ИИ[24] с бюджетом в 30 миллионов австралийских долларов в рамках Министерства промышленности.

Избегание масштабных изменений в австралийской правовой системе имеет смысл в свете основной цели плана — сделать Австралию привлекательным местом для международных инвестиций в центры обработки данных.

Новый австралийский план ставит в приоритет создание местной индустрии программного обеспечения для ИИ, распространение преимуществ «повышения производительности» ИИ на работников и пользователей государственных услуг, привлечение части неустанных глобальных инвестиций в центры обработки данных ИИ и продвижение регионального лидерства Австралии путем превращения ее в инфраструктурный и вычислительный центр в Индо-Тихоокеанском регионе.

Эти цели изложены в трех основных направлениях плана: использование возможностей, распространение преимуществ и обеспечение нашей безопасности.

С публикацией плана правительство официально отказалось от прежних предложений об обязательных мерах контроля для систем искусственного интеллекта высокого риска. Оно утверждает, что существующая правовая база Австралии уже достаточно сильна и может быть обновлена «в каждом конкретном случае».

Отмечается, что это противоречит общественному мнению. Более 75% австралийцев хотят регулирования ИИ.

Это также противоречит практике других стран. Европейский союз уже запрещает наиболее рискованные системы ИИ и обновил правила безопасности продукции и платформ. В настоящее время он также разрабатывает систему регулирования систем ИИ высокого риска. Системы федерального правительства Канады регулируются многоуровневой системой управления рисками. Южная Корея, Япония, Бразилия и Китай имеют правила, регулирующие риски, связанные с ИИ. Упомянутый выше закон об ИИ Казахстана также отдельно выделяет высоко-рисковые системы.

Президент Трамп запустил в Соединенных Штатах программу использования ИИ для ускорения научных открытий[25]. Миссия «Генезис», учрежденная указом президента, обязывает Министерство энергетики интегрировать свои 17 национальных лабораторий и некоторые из самых мощных суперкомпьютеров страны для проведения исследований в самых разных областях, от энергетики до медицины. Правительственные исследователи будут сотрудничать с партнерами из частного сектора, включая Anthropic, Nvidia и OpenAI, для обучения моделей на основе собственных федеральных наборов данных и использования ИИ для генерации и проведения экспериментов.

Министерство энергетики создаст платформу ИИ, которая обеспечит доступ к государственным данным и позволит федеральным агентствам, исследовательским лабораториям и компаниям сотрудничать в создании фундаментальных научных моделей и агентов ИИ. Оно также организует конкурсы, стипендии, партнерства и возможности финансирования, которые объединят эти сообщества, координируя различные правительственные, академические и частные ресурсы, которые обычно остаются разрозненными в мирное время. Проект представляет собой «крупнейшее мобилизационное использование федеральных научных ресурсов со времен программы «Аполлон», — заявил агентству Bloomberg Майкл Крациос, глава Управления по научно-технической политике Белого дома.

Цель состоит в обучении моделей ИИ планировать и проводить научные исследования с использованием роботизированных лабораторий, допускающих

[22] https://techxplore.com/news/2025-12-australia-national-ai-benefit.html

[23] https://www.industry.gov.au/publications/national-ai-plan/ministers-foreword

[24] https://www.industry.gov.au/news/australia-establishes-new-institute-strengthen-ai-safety

[25] https://www.whitehouse.gov/presidential-actions/2025/11/launching-the-genesis-mission/

различную степень участия человека. Миссия определяет шесть областей исследований: биотехнология, производство, материаловедение, ядерное деление, квантовая информатика и полупроводники.

Проект направлен на (i) ускорение темпов научных открытий, (ii) защиту национальной безопасности, (iii) поиск путей к снижению стоимости энергии и (iv) увеличение отдачи от государственных инвестиций для налогоплательщиков.

OpenAI опубликовала свою долгосрочную программу по кибербезопасности. Отмечается, что возможности кибербезопасности в моделях ИИ быстро развиваются, принося значительные преимущества в киберзащиту, а также создавая новые риски двойного назначения, которыми необходимо тщательно управлять. Например, возможности, оцениваемые в ходе задач типа «захват флага» (CTF), улучшились с 27% в GPT-5 (открывается в новом окне) в августе 2025 года до 76% в GPT-5.1-Codex-Max в ноябре 2025 года[26].

Компания ожидает, что будущие модели ИИ будут продолжать развиваться в этом направлении. OpenAI планирует, что каждая новая модель будет достигать «высокого» уровня кибербезопасности, измеряемого в соответствии с разработанной системой оценки готовности[27]. Под «высоким» уровнем кибербезопасности подразумеваются модели, которые могут либо разрабатывать работающие удаленные эксплойты нулевого дня против хорошо защищенных систем, либо оказывать существенную помощь в сложных, скрытых операциях по вторжению в корпоративную или промышленную среду, направленных на достижение реальных результатов. Система оценки готовности (The Preparedness Framework) - это подход OpenAI к отслеживанию и подготовке к новым перспективным возможностям, которые создают новые риски серьезного вреда. Под «серьезным ущербом» в этом документе подразумевается смерть или тяжелые травмы тысяч людей или экономический ущерб в размере сотен миллиардов долларов. При этом, комплекс мер безопасности от OpenAI охватывает широкий спектр рисков, включая многие риски, ущерб от которых не так серьезен. Здесь можно отметить наш обзор по рискам генеративного ИИ [2]

В настоящее время, в плане кибербезопасности, OpenAI концентрируется на трех областях новых перспективных возможностей, которые компания называет отслеживаемыми категориями:

- Биологические и химические возможности, которые, помимо открытия новых методов лечения, могут также снизить барьеры для создания и использования биологического или химического оружия.
- Возможности кибербезопасности, которые, помимо защиты уязвимых систем, могут также создавать новые риски масштабных кибератак использования уязвимостей.
- Возможности самосовершенствования ИИ, которые, помимо более быстрого раскрытия полезных возможностей, могут также создавать новые проблемы для управления системами ИИ человеком.

Aardvark[28], агент-исследователь от OpenAI в области безопасности, помогающий разработчикам и командам безопасности находить и исправлять уязвимости в больших масштабах, сейчас находится в закрытом бета-тестировании. Он сканирует кодовые базы на наличие уязвимостей и предлагает патчи, которые сопровождающие могут быстро внедрить. Он уже выявил новые CVE в программном обеспечении с открытым исходным кодом, анализируя целые кодовые базы. OpenAI планирует предложить бесплатное покрытие для отдельных некоммерческих репозиториев с открытым исходным кодом, чтобы внести свой вклад в безопасность экосистемы программного обеспечения с открытым исходным кодом и цепочки поставок. Подать заявку на участие можно здесь[29].

OpenAI создает Совет по передовым рискам (Frontier Risk Council) — консультативную группу, которая объединит опытных специалистов по киберзащите и безопасности для тесного сотрудничества с нашими командами. На начальном этапе совет сосредоточится на кибербезопасности, а в будущем расширит свою деятельность на другие передовые области. Члены совета будут консультировать по вопросам разграничения полезных и ответственных возможностей и потенциального злоупотребления, и полученные знания будут напрямую влиять на наши оценки и меры защиты. В этой связи смотри также наш обзор – что LLM знает о кибербезопасности [22].

OpenAI справедливо отмечает, что киберпреступления могут быть осуществимы с использованием любой передовой модели в отрасли. Для решения этой проблемы компания сотрудничает с другими передовыми лабораториями через Frontier Model Forum[30], некоммерческую организацию, поддерживаемую ведущими лабораториями ИИ и отраслевыми партнерами, чтобы выработать общее понимание моделей угроз и передовых методов. В этом контексте моделирование угроз помогает снизить риски, выявляя, как возможности ИИ могут быть использованы в качестве оружия, где существуют критические узкие места для различных субъектов угроз и как передовые модели могут обеспечить существенное улучшение. Это сотрудничество направлено на создание согласованного, обще-экосистемного понимания субъектов угроз и путей атак, что позволит лабораториям, разработчикам и защитникам лучше улучшать свои меры по смягчению последствий и обеспечивать быстрое распространение критически важных данных о безопасности по всей экосистеме. OpenAI также взаимодействуем с внешними командами для разработки оценок кибербезопасности. В частности, последний продукт GPT-5.2 Codex

[26] https://openai.com/index/strengthening-cyber-resilience/
[27] https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf
[28] https://openai.com/index/introducing-aardvark/
[29] https://openai.com/form/aardvark-beta-signup/
[30] https://www.frontiermodelforum.org/

оценивался в лаборатории Irregular[31], которая позиционирует себя как первую передовую лабораторию безопасности, миссия которой - защитить мир в эпоху все более совершенных и сложных систем искусственного интеллекта. Лаборатория создает системы защиты нового поколения с помощью высокоточных исследовательских платформ, которые моделируют и отслеживают реальные сценарии безопасности ИИ. OpenAI надеется, что экосистема независимых оценок будет способствовать дальнейшему формированию общего понимания возможностей моделей.

В ноябре 2025 года государства-члены ЕС подтвердили свою приверженность цифровому суверенитету, укрепив автономию и стратегический контроль над цифровой инфраструктурой, данными и новыми технологиями, такими как искусственный интеллект[32].

Европейская Комиссия опубликовала первый проект Кодекса практики по маркировке и обозначению контента, созданного с помощью ИИ[33].

Статья 50 Закона об ИИ[34] включает обязательства для поставщиков маркировать контент, созданный или обработанный с помощью ИИ, в машиночитаемом формате, а также для пользователей, которые используют системы генеративного ИИ в профессиональных целях, четко обозначать дипфейки и публикации текста, созданного с помощью ИИ, по вопросам, представляющим общественный интерес. Чтобы помочь поставщикам и разработчикам выполнить эти требования, Комиссия содействует разработке добровольного Кодекса практики, подготовленного независимыми экспертами, до вступления этих правил в силу.

Проект Кодекса практики состоит из двух разделов. Первый раздел содержит правила маркировки и обнаружения контента, созданного с помощью ИИ, применимые к поставщикам систем генеративного ИИ. Второй раздел посвящен маркировке дипфейков и определенного текста, сгенерированного или измененного ИИ, по вопросам, представляющим общественный интерес, и применим к разработчикам систем генеративного ИИ.

Поставщики, по этому кодексу, обеспечат маркировку контента, созданного или обработанного с помощью ИИ, незаметным водяным знаком. Этот водяной знак будет непосредственно встроен в контент таким образом, что его будет трудно отделить от контента, и он выдержит типичные этапы обработки, которые могут быть применены к контенту. Участники соглашения внедрят водяной знак наилучшим технически и экономически целесообразным способом. Участники соглашения могут внедрять водяные знаки во время обучения модели, вывода модели или в выходные данные модели или системы ИИ. Участники соглашения, предоставляющие модели ИИ другим поставщикам систем ИИ, внедрят соответствующие методы маркировки на уровне модели для обеспечения соответствия требованиям со стороны нижестоящих поставщиков

Отмечается, например, что производители, выпускающие модели или системы искусственного интеллекта с открытыми весами, будут внедрять методы структурной маркировки, закодированные в весах во время обучения модели. Это позволит третьим сторонам, использующим эти модели или системы с открытыми весами для создания генеративных систем искусственного интеллекта, соблюдать требования.

Второй проект будет разработан к середине марта 2026 года, а окончательная версия Кодекса ожидается к июню 2026 года. Правила, касающиеся прозрачности контента, созданного с помощью ИИ, вступят в силу 2 августа 2026 года.

Опубликован стандарт ISO/PAS 8800:2024, который описывает структуру управления безопасностью в системах искусственного интеллекта, используемых в транспортных средствах, расширяя существующие стандарты ISO, такие как ISO 26262 и ISO 21448. Он рассматривает риски функциональной безопасности и предоставляет рекомендации по разработке требований безопасности и мер по снижению рисков, специфичных для технологий искусственного интеллекта. Документ служит отраслевым руководством, признавая постоянно меняющийся характер приложений ИИ и необходимость в адаптированных подходах к обеспечению безопасности[35]. Что главное в этом документе – в сертификации для критических применений появились вероятности.

Опубликован национальный стандарт ГОСТ Р 72393-2025[36] «Технологии искусственного интеллекта в образовании. Алгоритмы идентификации вовлеченности при онлайн-обучении. Общие положения и методика испытаний».

Федеральная служба по техническому и экспортному контролю в декабре 2025 г. впервые внесла в банк данных угроз кибербезопасности риски, связанные с искусственным интеллектом, следует из сообщения на сайте регулятора. Теперь их надо будет учитывать ИТ-разработчикам софта для государственных структур и критической ИТ-инфраструктуры[37].

IV Обзор публикаций и проектов

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

[31] https://www.irregular.com/publications/model-evaluation-gpt-5.2-codex-on-offensive-security-benchmarks

[32] https://cdn.table.media/assets/europe/declaration-for-european-digital-sovereignty_final.pdf

[33] https://digital-strategy.ec.europa.eu/en/news/commission-publishes-first-draft-code-practice-marking-and-labelling-ai-generated-content

[34] https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

[35] https://www.iso.org/obp/ui/en/#iso:std:iso:pas:8800:ed-1:v1:en

[36] https://protect.gost.ru/v.aspx?control=8&id=258350

[37] https://www.cnews.ru/news/top/2025-12-23_fstek_vpervye_vydelila_ii

В рамках продолжения работ по безопасности ИИ-агентов, мы обновили первое учебной пособие на русском языке [11]. Охваченные вопросы:

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

В целом, мы готовы повторить с безопасностью ИИ-агентов тот же путь, который мы проделали с атаками на модели машинного обучения, начиная с работы [13].

На конференции cloud.ru[38] компания Hivetrace презентовала свои в части состязательного тестирования генеративных моделей[39]. Это, возможно, лучший на сегодняшний день продукт на отечественном рынке. Приятно, что основным разработчиком является Ю.Е. Лебединский, выпускник магистратуры факультета ВМК МГУ имени М.В. Ломоносова 2025 года, который продолжил работу, начатую в своей магистерской диссертации [14].

Рекомендательные системы (RecSys) широко применяются в различных областях, включая электронную коммерцию, финансы, здравоохранение, социальные сети, и оказывают все большее влияние на формирование поведения пользователей и принятие решений, что подчеркивает их растущее влияние в различных сферах. Однако недавние исследования показали, что RecSys уязвимы для атак, направленных на определение принадлежности (MIA - Membership Inference Attacks), целью которых является определение того, использовалась ли запись взаимодействия пользователя для обучения целевой модели или нет. MIAs на моделях RecSys могут напрямую привести к нарушению конфиденциальности. Например, выявив тот факт, что запись о покупке, использованная для обучения RecSys, связана с конкретным пользователем, злоумышленник может определить особенности этого пользователя. В последние годы было показано, что MIAs эффективны в других задачах машинного обучения, например, в моделях классификации и обработке естественного языка. Однако традиционные MIAs плохо подходят для RecSys из-за неизвестной (невидимой) апостериорной вероятности. В этой приводится первый всесторонний обзор таких атак в рекомендательных системах. Этот обзор предлагает всесторонний анализ последних достижений в области MIA в реляционных системах, рассматривая принципы проектирования, проблемы, атаки и защиты, связанные с этой развивающейся областью. Авторы предлагают единую таксономию, которая классифицирует различные MIA в рекомендательных системах на основе их характеристик, обсуждают их преимущества и недостатки. На основе выявленных в этом обзоре ограничений и пробелов указываются несколько перспективных направлений будущих исследований [15].

Очевидно, что торговля, как важный инструмент цифровой экономики [16], не избежит внедрения ИИ, при котором модели машинного (глубокого) обучения станут слабым звеном в безопасности. А если еще обратить внимание на внедрение ИИ-агентов в рекомендательные системы [17,18], то здесь еще добавляется и проблема объяснения действий системы.

Естественно, что в таких условиях и Умный дом [19] (его модели машинного обучения) не избежит состязательных атак. Взять, например, популярные в России звуковые колонки. Интеллектуальные голосовые системы широко используются для управления приложениями «умного дома», что вызывает серьезные опасения по поводу конфиденциальности и безопасности. Недавние исследования выявили их уязвимость к атакам со стороны злоумышленников, атакам повторного воспроизведения и т. д. Однако эти атаки основаны на голосовых данных жертвы. В нашей работе мы исследуем скрытую и независимую от команд атаку, которая не требует сбора голосов жертв. Предложенная в работе [20] атака IUAC, вводит голосовую систему в заблуждение, заставляя ее действовать против воли жертвы, независимо от отданных команд. Основная концепция заключается в обучении высоконадежных команд атаки путем построения разнообразных данных, что делает команды пользователя незначительными. Для достижения скрытых атак авторы используют высокочастотную несущую для построения неслышимой универсальной команды со стороны злоумышленников. Обширные эксперименты, проведенные с реальными наборами данных, показывают, что предложенная система атаки достигает средней успешности атаки в 96%, при этом сопротивляясь воздействию окружающей среды. Более того, успешность такой атаки против реальных голосовых систем в 4,52 раза выше, чем у современных аналогов. В заключение предлагается эффективный защитный механизм и приводятся экспериментальные данные для подтверждения его эффективности.

Интересная работа от Anthropic по удалению из LLM опасных знаний [21]. Большие языковые модели все чаще обладают возможностями, несущими риски двойного назначения, включая знания о химическом, биологическом, радиологическом и ядерном оружии. Для решения этих рисков в предыдущих работах была предложена градиентная маршрутизация - метод, который локализует целевые знания в выделенные параметры модели, которые впоследствии могут быть удалены. В данной же работе исследуется улучшенный вариант градиентной маршрутизации, названный селективным градиентным маскированием (SGTM). SGTM работает за счет того, что при обучении модели

[38] https://ods.ai/events/cloudru-aidevtoolsconf-041225
[39] https://hivetrace.ru/red

на опасных примерах обновляются только выделенные «удаляемые» параметры, оставляя остальную часть модели нетронутой.

Авторы демонстрируют, что SGTM обеспечивает лучший компромисс между удалением опасных знаний и сохранением общих возможностей по сравнению с простым отфильтровыванием опасных данных во время обучения, особенно когда метки, различающие «опасный» и «безопасный» контент, несовершенны. В отличие от поверхностных методов разучивания, которые можно быстро обратить вспять, SGTM устойчив к попыткам восстановления удаленных знаний, требуя в 7 раз больше переобучения для восстановления опасных возможностей по сравнению с другими методами разучивания.

Кибербезопасность охватывает множество взаимосвязанных областей, что усложняет разработку значимых, актуальных для рынка труда эталонных показателей. Существующие эталонные показатели оценивают отдельные навыки, а не интегрированную производительность. Авторы работы Cybersecurity AI Benchmark (CAIBench): A Meta-Benchmark for Evaluating Cybersecurity AI Agents [23] обнаружили, что предварительно обученные знания в области кибербезопасности в моделях LLM не подразумевают навыков атаки и защиты, что указывает на разрыв между знаниями и возможностями. Для решения этой проблемы они представили эталонный показатель кибербезопасности для ИИ (CAIBench), модульную мета-систему эталонных показателей, которая позволяет оценивать модели и агентов LLM в различных областях кибербезопасности, как наступательной, так и оборонительной, делая шаг к осмысленному измерению их актуальности для рынка труда. CAIBench объединяет пять категорий оценки, охватывающих более 10 000 примеров: CTF в стиле «Jeopardy», CTF по атаке и защите, упражнения на киберполигоне, эталонные показатели знаний и оценки конфиденциальности. Ключевые новые разработки включают систематическую одновременную оценку наступательных и оборонительных действий, задачи по кибербезопасности, ориентированные на робототехнику (RCTF2), и оценку производительности с сохранением конфиденциальности (CyberPII-Bench). Оценка современных моделей ИИ показывает насыщение метрик знаний в области безопасности (70% успеха), но существенное ухудшение в многоэтапных сценариях противодействия (20-40% успеха) или еще худшее в сценариях с роботизированными целями (22% успеха). Сочетание структуры фреймворка и выбора модели LLM значительно влияет на производительность; было обнаружено, что правильные совпадения улучшают дисперсию до 2,6 раз в CTF-соревнованиях атаки и защиты. Эти результаты демонстрируют выраженный разрыв между концептуальными знаниями и адаптивными возможностями, подчеркивая необходимость мета-бенчмарка.

Интересный обзор защитников (guardrails) представлен в работе On Guardrail Models' Robustness to Mutations and Adversarial Attacks [24]. Авторы отмечают, что риск предоставления небезопасной информации системами генеративного ИИ вызывает серьезные опасения, подчеркивая необходимость в защитных механизмах. Для снижения этого риска все чаще используются модели защиты, которые обнаруживают небезопасный контент во взаимодействии человека и ИИ, дополняя безопасность больших языковых моделей. Несмотря на недавние усилия по оценке эффективности этих моделей, их устойчивость к изменениям входных данных и атакам с использованием состязательных элементов остается в значительной степени неизученной. В этой статье представлена всесторонняя оценка 15 современных моделей защиты, оценивая их устойчивость к: а) изменениям входных данных, таким как опечатки, маскировка ключевых слов, шифры и скрытые выражения, и б) атакам с использованием состязательных элементов, предназначенным для обхода защитных механизмов моделей. Эти атаки используют возможности больших языковых моделей, такие как следование инструкциям, ролевая игра, персонификация, рассуждения и кодирование, или вводят состязательные токены для вызывания некорректного поведения модели. Результаты показывают, что большинство моделей защитных механизмов можно обойти с помощью простых изменений входных данных, и они уязвимы для атак со стороны злоумышленников. Например, один злонамеренный токен может обмануть их в среднем в 44,5% случаев. Ограничения текущего поколения моделей защитных механизмов подчеркивают необходимость создания более надежных защитных механизмов.

Вопросам безопасности ИИ-агентов посвящена и работа Agentic AI Security: Threats, Defenses, Evaluation, and Open Challenges [25]. Агентные системы искусственного интеллекта, работающие на основе больших языковых моделей (LLM) и обладающие функциями планирования, использования инструментов, памяти и автономности, становятся мощными и гибкими платформами для автоматизации. Их способность автономно выполнять задачи в веб-среде, программном обеспечении и физической среде создает новые и усиленные риски безопасности, отличающиеся как от традиционной безопасности ИИ, так и от обычной безопасности программного обеспечения. В этом обзоре представлена таксономия угроз, специфичных для агентного ИИ, рассмотрены последние сравнительные тесты и методологии оценки, а также обсуждаются стратегии защиты как с технической, так и с управленческой точек зрения. Авторы обобщают текущие исследования и выделяют открытые проблемы, стремясь поддержать разработку безопасных по умолчанию агентных систем.

Выложены слайды пленарного доклада конференции Современные информационные технологии на факультете ВМК МГУ имени М.В. Ломоносова - Безопасность ИИ-агентов - реальность или

миф?[40]

Статья Cyber Attacks on Commercial Drones: A Review [26] посвящена атакам на дроны. Беспилотные летательные аппараты (БПЛА), также известные как дроны, всё чаще используются в различных приложениях, и на них можно проводить различные кибератаки с использованием разных инструментов. Некоторые примеры этих атак включают разрыв соединения между дроном и контроллером с помощью атак деаутентификации, раскрытие пароля или криптографического ключа, используемого в протоколе связи, получение управления дроном посредством внедрения команд/кода и атаки типа «человек посередине» (MitM). В данной статье рассматриваются атаки с использованием дронов посредством анализа различных компонентов дрона, включая пульт дистанционного управления и протоколы связи. Основная цель — предоставить обзор возможных способов осуществления кибератак. В этом анализе сделан вывод о том, что дроны, предназначенные для различных целей, уязвимы для ряда кибератак. В статье также рассматриваются существующие методологии тестирования на проникновение для БПЛА, которые обеспечивают логическую основу для их реализации.

По мере того, как системы обнаружения дипфейков становятся всё более сложными, понимание их уязвимостей становится критически важным для разработки надёжной защиты. В работе Adversarial Reality for Evading Deepfake Image Detectors [27] представлено комплексное исследование конкурентных атак на детекторы дипфейков на основе изображений, предлагая новый подход, создающий «конкурентную реальность» — синтетические изображения, которые сохраняют визуальное сходство с оригинальными дипфейками, успешно обходя автоматизированные системы обнаружения. Предложенный метод использует генеративную структуру с архитектурой в стиле UNet для преобразования изображений, сгенерированных GAN, диффузионно-генерированных и обработанных лиц, в варианты, обманывающие детектор, сохраняя при этом визуальную точность. В отличие от традиционных подходов, основанных на возмущениях, которые добавляют шумовые паттерны, новый генеративный метод обучается преобразованиям, специфичным для изображений, без необходимости использования вручную созданных спектральных фильтров. Благодаря обширной оценке различных наборов данных, типов генераторов и архитектур детекторов демонстрируется, что предложенный подход достигает уровня ошибочной классификации до 98,83% на диффузных изображениях и 83,36% на контенте на основе GAN, сохраняя при этом высокое качество восприятия со средними баллами PSNR выше 35. Полученные результаты выявляют критические уязвимости в существующих системах обнаружения и дают представление о разработке более надежных детекторов дипфейков.

Два новых пакета для состязательного тестирования LLM представлены в работе AdversariaLLM: A Unified and Modular Toolbox for LLM Robustness Research [28]. Стремительное расширение исследований безопасности и надежности больших языковых моделей (LLM) привело к появлению разрозненной и зачастую содержащей ошибки экосистемы реализаций, наборов данных и методов оценки. Эта фрагментация затрудняет воспроизводимость и сопоставимость результатов различных исследований, препятствуя существенному прогрессу. Для решения этих проблем представлен ADVERSARIALLM, набор инструментов для проведения исследований надежности джейлбрейка LLM. Его дизайн ориентирован на воспроизводимость, корректность и расширяемость. Фреймворк реализует двенадцать алгоритмов состязательных атак, объединяет семь эталонных наборов данных, охватывающих оценку вредоносности, избыточного отказа и полезности, и предоставляет доступ к широкому спектру открытых LLM через Hugging Face. Реализация включает расширенные функции для обеспечения сопоставимости и воспроизводимости, такие как отслеживание ресурсов компьютера, детерминированные результаты и методы оценки распределения. ADVERSARIALLM также интегрирует систему оценки через сопутствующий пакет JUDGEZOO, который также может использоваться независимо. Вместе эти компоненты направлены на создание прочной основы для прозрачных, сравнимых и воспроизводимых исследований в области безопасности магистратуры по праву. Оба пакета доступны на GitHub.

Очередные ужасы AI Red Team. Запросы в прошедшем времени уже обходили фильтры LLM. Теперь выяснилось, что их обходят еще и стихи ...

В работе Adversarial Poetry as a Universal Single-Turn Jailbreak Mechanism in Large Language Models (*Состязательная поэзия – каков слог!)* [29] представлены доказательства того, что состязательная поэзия функционирует как универсальный одношаговый джейлбрек для больших языковых моделей (LLM). В 25 передовых проприетарных и открытых моделях курируемые поэтические подсказки показали высокие показатели успешности атак (ASR), превышающие 90% у некоторых поставщиков. Сопоставление подсказок с таксономиями рисков MLCommons и EU CoP показывает, что поэтические атаки переносятся в области CBRN (Chemical, Biological, Radiological, and Nuclear) опасностей, манипуляций, киберпреступлений и потери контроля. Преобразование 1200 вредоносных подсказок MLCommons в стихи с помощью стандартизированного мета-подсказки дало ASR до 18 раз выше, чем их базовые показатели для прозы. Результаты оцениваются с помощью ансамбля из 3 экспертов LLM с открытым весом, чьи бинарные оценки безопасности были проверены на стратифицированном подмножестве, маркированном людьми. Поэтические подсказки достигли среднего уровня успешности взлома 62% для стихотворений, написанных вручную, и примерно 43% для мета-подсказок (по сравнению с непоэтическими

[40] http://inetique.ru/articles/agents_2025.pdf

базовыми вариантами), значительно превзойдя непоэтические базовые варианты и выявив систематическую уязвимость среди модельных семейств и подходов к обучению безопасности. Эти результаты показывают, что одни только стилистические вариации могут обойти современные механизмы безопасности, указывая на фундаментальные ограничения существующих методов выравнивания и протоколов оценки.

Специально обученная LLM удаляет инъекции подсказок [30]. Когда агенты больших языковых моделей (LLM) всё чаще используются для автоматизации задач и взаимодействия с недоверенными внешними данными, внедрение подсказок становится серьёзной угрозой безопасности. Внедряя вредоносные инструкции в данные, к которым обращаются LLM, злоумышленник может произвольно переопределить исходную задачу пользователя и перенаправить агента на выполнение непреднамеренных, потенциально опасных действий. Существующие средства защиты либо требуют доступа к весам модели (тонкая настройка), либо приводят к существенной потере полезности (основанная на обнаружении), либо требуют нетривиальной переработки системы (на системном уровне). В связи с этим в работе Defending Against Prompt Injection with DataFilter [30] предлагается DataFilter — защита, не зависящую от модели, которая удаляет вредоносные инструкции из данных до того, как они достигнут бэкенда LLM. DataFilter обучается с контролируемой тонкой настройкой на имитационных внедрениях и использует как инструкции пользователя, так и данные для выборочного удаления вредоносного контента, сохраняя при этом безвредную информацию. В множестве бенчмарков DataFilter стабильно снижает процент успешных атак с использованием инъекций подсказок практически до нуля, сохраняя при этом полезность LLM. DataFilter обеспечивает надежную безопасность, высокую полезность и быстрое развертывание, что делает его надежной практической защитой для защиты коммерческих LLM от инъекций подсказок. Код системы открыт.

Со времен первой работы 2015 года, которая и определила состязательные примеры, последние всегда строились исходя из минимизации (незаметности) изменений. Состязательные примеры создавались (и создаются) путем применения тонких, но намеренно худших модификаций к примерам из набора данных, что приводит к тому, что модель выдает ответ, отличный от исходного примера. В работе A New Type of Adversarial Examples [31] состязательные примеры формируются совершенно противоположным образом. Они существенно отличаются от исходных примеров, но приводят к тому же ответу. Авторы предлагают новый набор алгоритмов для создания таких состязательных примеров, включая метод отрицательного итерационного быстрого градиента (NI-FGSM) и метод отрицательного итерационного быстрого градиента (NI-FGM), а также их варианты с импульсом: метод отрицательного итерационного быстрого градиента (NMI-FGSM) и метод отрицательного итерационного быстрого градиента (NMI-FGM). Состязательные примеры, созданные этими методами, могут быть использованы для проведения атаки на системы машинного обучения в определенных случаях. Более того, полученные результаты показывают, что вредоносные примеры не просто распределены по соседству с примерами из набора данных; вместо этого они широко распределены в пространстве выборки. Человек, при классификации, легко отличит состязательный пример от оригинала. Модель же - не отличает.

Больше анонсов интересных публикаций можно найти в блоге Абаванет[41].

Благодарности

Этот выпуск подготовлен при прямом содействии факультета ВМК МГУ имени М.В. Ломоносова. Также хотелось бы поблагодарить сотрудников кафедры Информационной безопасности факультета ВМК за плодотворные дискуссии и обсуждения.

Библиография

[1] Намиот, Д. Е., Е. А. Ильюшин, and И. В. Чижов. "Искусственный интеллект и кибербезопасность." International Journal of Open Information Technologies 10.9 (2022): 135-147.
[2] Намиот, Д. Е., and Е. А. Ильюшин. "О киберрисках генеративного искусственного интеллекта." International Journal of Open Information Technologies 12.10 (2024): 109-119.
[3] Namiot, Dmitry. "Artificial Intelligence in Cybersecurity. Chronicle. Issue 1." International Journal of Open Information Technologies 13.9 (2025): 34-42.
[4] Namiot, Dmitry. "Artificial Intelligence in Cybersecurity. Chronicle. Issue 2." International Journal of Open Information Technologies 13.10 (2025): 58-67.
[5] Kuzmenko, Ilya Dmitrievich, and Dmitry Evgenyevich Namiot. "Методы обнаружения дипфейков в видеоконференциях в реальном времени." Современные информационные технологии и ИТ-образование 21.2 (2025).
[6] Prakasha, K. Krishna, and U. Sumalatha. "Privacy-preserving techniques in biometric systems: Approaches and challenges." IEEE Access (2025).
[7] Disrupting the first reported AI-orchestrated cyber espionage campaign https://www.anthropic.com/news/disrupting-AI-espionage Retrieved: Dec, 2025
[8] Disrupting the first reported AI-orchestrated cyber espionage campaign. Full report https://assets.anthropic.com/m/ec212e6566a0d47/original/Disrupting-the-first-reported-AI-orchestrated-cyber-espionage-campaign.pdf
[9] Цифровая экономика и Интернет Вещей - преодоление силоса данных / В. П. Куприяновский, А. Р. Ишмуратов, Д. Е. Намиот [и др.] // International Journal of Open Information Technologies. – 2016. – Т. 4, № 8. – С. 36-42. – EDNWFVAPB.
[10] Namiot, Dmitry, and Eugene Ilyushin. "On the Cybersecurity of AI Agents." International Journal of Open Information Technologies 13.9 (2025): 13-24.
[11] Безопасность ИИ-агентов https://abava.blogspot.com/2025/12/blog-post_11.html Retrieved: Dec, 2025
[12] 15 Security Threats to LLM Agents (with Real-World Examples) https://research.aimultiple.com/security-of-ai-agents/ Retrieved:
[13] 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
[14] Lebedinskiy, Yuriy, and Dmitry Namiot. "Adversarial testing of large language models." International Journal of Open Information Technologies 13.11 (2025): 132-152.

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

[15] He, Jiajie, et al. "Membership Inference Attacks on Recommender System: A Survey." arXiv preprint arXiv:2509.11080 (2025).
[16] Розничная торговля в цифровой экономике / В. П. Куприяновский, С. А. Синягов, Д. Е. Намиот [и др.] // International Journal of Open Information Technologies. – 2016. – Т. 4, № 7. – С. 1-12. – EDN WCMIWN.
[17] Huang, Xu, et al. "Recommender ai agent: Integrating large language models for interactive recommendations." ACM Transactions on Information Systems 43.4 (2025): 1-33.
[18] Zhu, Xi, et al. "Recommender systems meet large language model agents: A survey." Foundations and Trends® in Privacy and Security 7.4 (2025): 247-396.
[19] Волков, А. А. О задачах создания эффективной инфраструктуры среды обитания / А. А. Волков, Д. Е. Намиот, М. А. Шнепс-Шнеппе // International Journal of Open Information Technologies. – 2013. – Т. 1, № 7. – С. 1-10. – EDN ROMIZX.
[20] Sun, Haifeng, et al. "IUAC: Inaudible Universal Adversarial Attacks Against Smart Speakers." ACM Transactions on Sensor Networks 21.1 (2025): 1-20.
[21] Shilov, Igor, et al. "Beyond Data Filtering: Knowledge Localization for Capability Removal in LLMs." arXiv preprint arXiv:2512.05648 (2025).
[22] Namiot, Dmitry. "What LLM knows about cybersecurity." International Journal of Open Information Technologies 13.7 (2025): 37-46.
[23] Sanz-Gómez, María, et al. "Cybersecurity AI Benchmark (CAIBench): A Meta-Benchmark for Evaluating Cybersecurity AI Agents." arXiv preprint arXiv:2510.24317 (2025).
[24] Bassani, Elias, and Ignacio Sanchez. "On Guardrail Models' Robustness to Mutations and Adversarial Attacks." Findings of the Association for Computational Linguistics: EMNLP 2025. 2025.
[25] Datta, Shrestha, et al. "Agentic ai security: Threats, defenses, evaluation, and open challenges." arXiv preprint arXiv:2510.23883 (2025).
[26] Branco, Bruno, José Silvestre Silva, and Miguel Correia. "Cyber attacks on commercial drones: A review." IEEE Access (2025).
[27] Ciftci, Umur Aybars, et al. "Adversarial Reality for Evading Deepfake Image Detectors." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2025.
[28] Beyer, Tim, et al. "AdversariaLLM: A Unified and Modular Toolbox for LLM Robustness Research." arXiv preprint arXiv:2511.04316 (2025).
[29] Bisconti, Piercosma, et al. "Adversarial poetry as a universal single-turn jailbreak mechanism in large language models." arXiv preprint arXiv:2511.15304 (2025).
[30] Wang, Yizhu, et al. "Defending against prompt injection with datafilter." arXiv preprint arXiv:2510.19207 (2025).
[31] Nie, Xingyang, et al. "A New Type of Adversarial Examples." arXiv preprint arXiv:2510.19347 (2025).

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

Artificial Intelligence in Cybersecurity. Chronicle. Issue 4

Dmitry Namiot

***Abstract* - In this document, we offer our fourth monthly overview of current events, based on a general topic: the use of Artificial Intelligence (AI) in cybersecurity. In this document, we regularly describe regulatory documents, significant events, and new developments in this field. Currently, we combine these three aspects. First, these are incidents related to the use of AI for cybersecurity. For example, identified vulnerabilities and risks in generative AI, new adversarial impacts on machine learning models and AI agents, etc. Second, this is a global regularity: regulatory documents, new global and local standards, various aspects of Area II in cybersecurity. And third, each overview includes new interesting publications in this area. All subsequent materials reflect the views and preferences of the authors. This article presents the fourth issue of the Chronicle of AI in Cybersecurity.**

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

REFERENCES

[1] Namiot, D. E., E. A. Il'jushin, and I. V. Chizhov. "Iskusstvennyj intellekt i kiberbezopasnost'." International Journal of Open Information Technologies 10.9 (2022): 135-147.
[2] Namiot, D. E., and E. A. Il'jushin. "O kiberriskah generativnogo iskusstvennogo intellekta." International Journal of Open Information Technologies 12.10 (2024): 109-119.
[3] Namiot, Dmitry. "Artificial Intelligence in Cybersecurity. Chronicle. Issue 1." International Journal of Open Information Technologies 13.9 (2025): 34-42.
[4] Namiot, Dmitry. "Artificial Intelligence in Cybersecurity. Chronicle. Issue 2." International Journal of Open Information Technologies 13.10 (2025): 58-67.
[5] Kuzmenko, Ilya Dmitrievich, and Dmitry Evgenyevich Namiot. "Metody obnaruzhenija dipfejkov v videokonferencijah v real'nom vremeni." Sovremennye informacionnye tehnologii i IT-obrazovanie 21.2 (2025).
[6] Prakasha, K. Krishna, and U. Sumalatha. "Privacy-preserving techniques in biometric systems: Approaches and challenges." IEEE Access (2025).
[7] Disrupting the first reported AI-orchestrated cyber espionage campaign https://www.anthropic.com/news/disrupting-AI-espionage Retrieved: Dec, 2025
[8] Disrupting the first reported AI-orchestrated cyber espionage campaign. Full report https://assets.anthropic.com/m/ec212e6566a0d47/original/Disrupting-the-first-reported-AI-orchestrated-cyber-espionage-campaign.pdf
[9] Cifrovaja jekonomika i Internet Veshhej - preodolenie silosa dannyh / V. P. Kuprijanovskij, A. R. Ishmuratov, D. E. Namiot [i dr.] // International Journal of Open Information Technologies. – 2016. – T. 4, # 8. – S. 36-42. – EDN WFVAPB.
[10] Namiot, Dmitry, and Eugene Ilyushin. "On the Cybersecurity of AI Agents." International Journal of Open Information Technologies 13.9 (2025): 13-24.
[11] Bezopasnost' II- https://abava.blogspot.com/2025/12/blog-post_11.html Retrieved: Dec, 2025
[12] 15 Security Threats to LLM Agents (with Real-World Examples) https://research.aimultiple.com/security-of-ai-agents/ Retrieved:
[13] 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
[14] Lebedinskiy, Yuriy, and Dmitry Namiot. "Adversarial testing of large language models." International Journal of Open Information Technologies 13.11 (2025): 132-152.
[15] He, Jiajie, et al. "Membership Inference Attacks on Recommender System: A Survey." arXiv preprint arXiv:2509.11080 (2025).
[16] Roznichnaja torgovlja v cifrovoj jekonomike / V. P. Kuprijanovskij, S. A. Sinjagov, D. E. Namiot [i dr.] // International Journal of Open Information Technologies. – 2016. – T. 4, # 7. – S. 1-12. – EDN WCMIWN.
[17] Huang, Xu, et al. "Recommender ai agent: Integrating large language models for interactive recommendations." ACM Transactions on Information Systems 43.4 (2025): 1-33.
[18] Zhu, Xi, et al. "Recommender systems meet large language model agents: A survey." Foundations and Trends® in Privacy and Security 7.4 (2025): 247-396.
[19] Volkov, A. A. O zadachah sozdanija jeffektivnoj infrastruktury sredy obitanija / A. A. Volkov, D. E. Namiot, M. A. Shneps-Shneppe // International Journal of Open Information Technologies. – 2013. – T. 1, # 7. – S. 1-10. – EDN ROMIZX.
[20] Sun, Haifeng, et al. "IUAC: Inaudible Universal Adversarial Attacks Against Smart Speakers." ACM Transactions on Sensor Networks 21.1 (2025): 1-20.
[21] Shilov, Igor, et al. "Beyond Data Filtering: Knowledge Localization for Capability Removal in LLMs." arXiv preprint arXiv:2512.05648 (2025).
[22] Namiot, Dmitry. "What LLM knows about cybersecurity." International Journal of Open Information Technologies 13.7 (2025): 37-46.
[23] Sanz-Gómez, María, et al. "Cybersecurity AI Benchmark (CAIBench): A Meta-Benchmark for Evaluating Cybersecurity AI Agents." arXiv preprint arXiv:2510.24317 (2025).
[24] Bassani, Elias, and Ignacio Sanchez. "On Guardrail Models' Robustness to Mutations and Adversarial Attacks." Findings of the Association for Computational Linguistics: EMNLP 2025. 2025.
[25] Datta, Shrestha, et al. "Agentic ai security: Threats, defenses, evaluation, and open challenges." arXiv preprint arXiv:2510.23883 (2025).
[26] Branco, Bruno, José Silvestre Silva, and Miguel Correia. "Cyber attacks on commercial drones: A review." IEEE Access (2025).
[27] Ciftci, Umur Aybars, et al. "Adversarial Reality for Evading Deepfake Image Detectors." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2025.
[28] Beyer, Tim, et al. "AdversariaLLM: A Unified and Modular Toolbox for LLM Robustness Research." arXiv preprint arXiv:2511.04316 (2025).
[29] Bisconti, Piercosma, et al. "Adversarial poetry as a universal single-turn jailbreak mechanism in large language models." arXiv preprint arXiv:2511.15304 (2025).
[30] Wang, Yizhu, et al. "Defending against prompt injection with datafilter." arXiv preprint arXiv:2510.19207 (2025).
[31] Nie, Xingyang, et al. "A New Type of Adversarial Examples." arXiv preprint arXiv:2510.19347 (2025).