БЕЗОПАСНОСТЬ ЦИФРОВЫХ ТЕХНОЛОГИЙ. – 2026. – № 1 (120). – 29–53

МЕТОДЫ И СИСТЕМЫ ЗАЩИТЫ ИНФОРМАЦИИ, ИНФОРМАЦИОННАЯ БЕЗОПАСНОСТЬ

УДК 004 DOI: 10.17212/2782-2230-2026-1-29-53

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ В ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ: НАСТУПАТЕЛЬНЫЕ И ОБОРОНИТЕЛЬНЫЕ СТРАТЕГИИ, МЕТОДЫ, РИСКИ И ПЕРСПЕКТИВЫ*

Д.А. АЗЯВА[1], А.А. КИСЕЛЕВ[2]

[1] *РФ, 630073, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, лаборант кафедры защиты информации. E-mail: dimulatoraz@gmail.com*

[2] *РФ, 630073, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, старший преподаватель кафедры защиты информации. E-mail: anton.kiselev@corp.nstu.ru*

В статье анализируется влияние искусственного интеллекта на сферу информационной безопасности. Цель работы заключается в комплексном рассмотрении различных аспектов применения технологий искусственного интеллекта в современном киберпространстве. В ходе исследования были проанализированы реальные случаи применения искусственного интеллекта для автоматизации кибератак и поиска информации в открытых источниках, в том числе фишинг и OSINT-разведка. Также были рассмотрены способы и инструменты защиты с применением технологии искусственного интеллекта, включая динамическую аутентификацию, непрерывный мониторинг системы. Помимо анализа технических аспектов применения технологий искусственного интеллекта, были изучены юридические практики, связанные с применением технологий искусственного интеллекта в сфере информационной безопасности. В ходе исследования было установлено, что искусственный интеллект в сфере информационной безопасности оказывает неоднозначное воздействие. Выявлены случаи, когда одна и та же функция искусственного интеллекта может давать как позитивные, так и негативные последствия. Кроме того, установлено существование правового вакуума в законодательствах многих стран в отношении регулирования технологий искусственного интеллекта. Также были выявлены основные риски информационной безопасности, возникающие при использовании искусственного интеллекта. Практическая значимость работы заключается в обзоре и систематизации уже существующих примеров влияния искусственного интеллекта. Результаты настоящего исследования могут служить основой для разработки рекомендаций по совершенствованию корпоративных политик информационной

* Статья получена 11 февраля 2026 г.

безопасности, модернизации стратегий подготовки сотрудников к работе с принципиально новыми технологиями, выбору оптимальных технологических решений для защиты информации в условиях современного киберпространства, а также разработке мер по защите информации при использовании для функционирования информационных систем искусственного интеллекта согласно Приказу ФСТЭК от 11.04.2025 № 17.

Ключевые слова: моделирование, гиперграф, конфигурация, активы, управление информационной безопасностью

ВВЕДЕНИЕ

В последние годы наблюдается активное развитие искусственного интеллекта (далее – ИИ) и его широкое внедрение в разные сферы. Использование ИИ влияет на информационные процессы через автоматизацию и стремительно меняет ландшафт информационной безопасности (далее – ИБ) [1], позволяя использовать ИИ как источник защиты и одновременно способствуя появлению новых видов угроз. В связи с этим подход к обеспечению ИБ претерпевает серьезные изменения, подразумевая применение технологий ИИ для обнаружения ранее неизвестных уязвимостей [2], разработку технологий, направленных на обеспечение безопасности технологий ИИ [3], и разработку рекомендаций по безопасному использованию искусственного интеллекта [4]. Важность ИИ подчеркнули и участники заседания МИД России [5].

Таким образом, ИИ является не только инструментом для подготовки и реализации атак, но и объектом воздействия и контроля. В качестве цели могут выступать как частные, так и юридические лица, включая сотрудников, оборудование, сервисы, бизнес-процессы, бренд и финансовые активы. Это обусловливает интерес к исследованию угроз применения ИИ [6, 7] и, как следствие, к разработке инструментов, обеспечивающих безопасность работы с технологиями ИИ [8].

В рамках настоящей статьи для определения термина ИИ воспользуемся термином, определенным в указе Президента РФ от 10.10.2019 № 490 [9]: «искусственный интеллект – комплекс технологических решений, позволяющий имитировать когнитивные функции человека (включая поиск решений без заранее заданного алгоритма) и получать при выполнении конкретных задач результаты, сопоставимые с результатами интеллектуальной деятельности человека или превосходящие их. Комплекс технологических решений включает в себя информационно-коммуникационную инфраструктуру, ПО (в том числе то, в котором используются методы машинного обучения), процессы и сервисы по обработке данных и поиску решений». Таким образом, термин «искусственный интеллект» будет включать в себя нейронные сети, большие языковые модели, чат-боты и самообучающиеся алгоритмы, спо-

собные принимать решения на основе данных и адаптироваться к новым условиям работы.

1. НАСТУПАТЕЛЬНЫЙ ИСКУСТВЕННЫЙ ИНТЕЛЛЕКТ

Специалисты отмечают [10], что сегодня технологии ИИ уже активно применяются в подготовке, осуществлении и автоматизации различного рода кибератак. Развитие ИИ открыло злоумышленникам множество возможностей – от банальной автоматизации рутинных процессов написания вредоносов до создания сложных сценариев таргетированных атак. В частности, специалисты ЗАО «Лаборатория Касперского» выявили [11], что особенностью группировки FunkSec является использование ИИ-инструментов, в том числе для разработки шифровального ПО. Эксперты также отмечают [12] появление нового семейства вирусов LameHug, их особенность заключается в использовании ИИ для генерации команд на скомпрометированных системах под управлением ОС Windows. Различные исследования показывают, что киберпреступники стали чаще использовать ИИ для генерации уникального контента и маскировки атак, делая их более сложными для обнаружения и сокращая время на разработку [13–15]. Более того, некоторые исследования [16] показывают, что развитие open-source ИИ позволило значительно снизить технический барьер для злоумышленников. На основе этих данных с полной уверенностью можно сказать, что сегодня злоумышленники применяют ИИ для следующих задач.

1.1. АВТОМАТИЗАЦИЯ ФИШИНГА

Согласно данным SoSafe [17], 78 % пользователей открывали фишинговые письма, из которых 21 % (каждый пятый) переходили по фишинговым ссылкам. Также исследование компании IBM в 2023 году показало [18], что ИИ позволял составить фишинговое письмо гораздо быстрее специалистов (за 5 минут против 16 часов). Однако эффективность писем, созданных ИИ, была ниже созданных специалистами (11 % против 14 %). Это объяснялось тем, что в первую очередь письма, созданные человеком, имели необходимый эмоциональный фон, побуждающий жертву к необходимым действиям. Но уже в 2024 году появилась технология распознавания эмоций Empathic Voice Interface (EVI) [19], а следовательно, эмоциональность ИИ остается лишь вопросом времени. Помимо фишинговых писем, злоумышленники используют ИИ для создания большого количества фишинговых сайтов с разнообразным и уникальным контентом [20]. Это значительно затрудняет идентификацию фишинговых веб-страниц стандартными методами.

Российские исследователи также отмечают рост целевого фишинга на 32 % [21] и связывают это с использованием злоумышленниками технологий ИИ. Следует добавить, что фишинг с использованием ИИ применяется не только для атак на компании, но и с целью получения денежных средств обычных пользователей. [22]

1.2. АВТОМАТИЗАЦИЯ OSINT

OSINT (open source intelligence, разведка по открытым источникам) представляет собой способ сбора и анализа информации из открытых источников. Этот способ применяется для подготовки атак на пользователей, но также широко распространен в кибербезопасности для аналитики уязвимостей в информационных системах, позволяя специалистам узнать об утечках информации и выявить, кто и как передает информацию хакерам.

Так, ИИ уже способен демонстрировать высокую эффективность в задачах по сбору данных из различных источников, таких как социальные сети, веб-сайты организаций, новостные ресурсы и прочие открытые источники данных, включая тексты на изображениях [23]. Кроме того, ИИ способен идентифицировать и выделять ключевые элементы информации, что позволяет киберпреступникам проводить комплексную разведку как в отношении организации в целом, так и в отношении ее сотрудников [10], позволяя не только автоматизировать сбор информации о потенциальных жертвах, но и выделять из общей массы объекты, атака на которые с большей вероятностью пройдет успешно. Однако применение ИИ на этапе OSINT сопряжено со множеством рисков и трудностей. Анализ, проведенный компанией RAND [24], показал, что оценка эффективности ИИ не должна основываться исключительно на точности результатов работы алгоритма, стоит периодически проводить оценку той информации, которую ИИ выдает пользователю.

1.3. СОЗДАНИЕ ДИПФЕЙКОВ

Дипфейки, созданные с помощью ИИ, в основном пока имитирующие политиков и знаменитостей, встречаются гораздо чаще, чем использование ИИ в кибератаках. Согласно отчету компании Positive Technologies за 2023–2024 гг. [10], в 46 % успешные атаки с применением дипфейков были направлены на манипуляции общественным мнением, в 26 % – на создание мошеннической рекламы и только лишь в 22 % – на подделку голоса и видео родственников жертвы с целью получения денежных средств. Также стоит отметить, что исследователи ЗАО «Лаборатория Касперского» зафиксировали случаи распространения дипфейков в мессенджерах и социальных сетях [25],

где злоумышленник выдавал себя за знакомого жертвы и просил перевести денежные средства. Таким образом, использование ИИ позволило подделать голос и внешность достаточно достоверно.

1.4. АВТОМАТИЗАЦИЯ АТАК

Проведенное ранее исследование [26] доказало, что ИИ может автономно эксплуатировать атаки типа one-day, например, построить атаку только на основе информации об уязвимости (CVE). Кроме того, в 2025 году исследователям удалось выявить вредоносные пакеты в репозитории npm [27]. После установки пакеты использовали самогенерируемый код для сбора информации, в том числе файлы cookie и токены. Использование такого кода позволяло адаптироваться к окружению и маскироваться от анализаторов кода.

1.5. ПОМОЩЬ НАЧИНАЮЩИМ ХАКЕРАМ ПРИ ОРГАНИЗАЦИИ АТАКИ

Хоть ИИ и имеет ограничения и цензуру ответов, начинающий хакер с базовыми знаниями в программировании может легитимными запросами «убедить» ИИ создать функциональные блоки вредоносного ПО. Помимо этого, в 2024 году проводилось исследование [28], которое показало, что, используя ChatGPT и инструменты, вполне реально произвести взлом системы через компьютер, находящийся в одной сети с жертвой. А в августе 2025 года JetBrains представила Kineto [29] – no-code-платформу, позволяющую создавать полноценные веб-приложения примерно за 20 минут, используя только описание на естественном языке. Это доказывает, что в настоящее время для злоумышленников технический барьер стал гораздо ниже и не требует навыков написания программного кода, а значит, и увеличивает возможность вовлеченности уже с подросткового возраста.

2. ОБОРОНИТЕЛЬНЫЙ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

Независимо от того, какой вклад ИИ вносит в развитие киберпреступности, стоит отметить, что неменьшее влияние он оказывает и на сферу ИБ, предоставляя специалистам новые методики и инструменты для защиты систем, инфраструктуры и пользователей.

2.1. АВТОМАТИЗАЦИЯ ПРОЦЕССА МОНИТОРИНГА И УЛУЧШЕНИЕ ВЫЯВЛЕНИЯ УГРОЗ

В настоящее время ручные методы OSINT для обнаружения киберугроз не позволяют своевременно выявить новые типы угроз и отреагировать на них.

В связи с этим возрастает потребность в использовании автономных систем обнаружения киберугроз, например, таких как системы на базе ИИ [30].

При прогнозировании киберугроз ИИ выступает ключевым инструментом, так как системы мониторинга и защиты на базе ИИ способны непрерывно производить обработку больших объемов различной информации. Это позволяет быстрее выявлять возникающие угрозы и своевременно реагировать на них. Также системы, использующие ИИ, ведут постоянное самосовершенствование, адаптируясь к новым условиям без участия человека. Например, ИИ-система может анализировать паттерны в различных векторах атак, выявляя скрытые связи, казалось бы, между независимыми элементами. Это позволяет определять тактики, методы и процедуры, используемые злоумышленниками, что способствует более эффективной работе системы.

Существуют исследования, демонстрирующие высокую эффективность ИИ при выявлении фишинга. Технологии ИИ позволили выявить с точностью 97,25 % даже не интуитивно понятный фишинг, обманувший человека [31], и с точностью 98,3 % обнаружить фишинговые сайты [32].

Яркими примерами эффективности ИИ в киберзащите стали следующие два случая. Первый произошел в мае 2025 года. При использовании ИИ от OpenAI исследователь смог обнаружить уязвимость «нулевого дня» в ядре Linux (CVE-2025-37899) [33], что стало первым примером активной помощи ИИ при обнаружении ранее неизвестной уязвимости ПО. Второй случай произошел в июле 2025 года. Технология Google Big Sleep, работающая на базе ИИ, позволила предотвратить атаку «нулевого дня» на SQLite. Эта по праву первый в истории случай, когда ИИ смог выявить и предотвратить новую атаку без участия человека.

Некоторые компании подошли к вопросу борьбы с угрозами ИИ комплексно. Например, Meta (признана в РФ экстремистской организацией и запрещена) открыла доступ к новому набору инструментов, способных распознать вредоносные текстовые и визуальные запросы, противодействовать небезопасному коду и плагинам, выявлять взломы и скрытые команды, оценивать центры кибербезопасности и самостоятельно устранять уязвимости в коде.

2.2. ПРИМЕНЕНИЕ ИСКУСТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ ДИНАМИЧЕСКОЙ АУТЕНТИФИКАЦИИ

Применение технологий ИИ позволяет использовать динамическую (непрерывную) аутентификацию пользователей, которая осуществляется не на одноразовой проверке во время входа в систему, а проводится в течение всей сессии. Этот метод может основываться на динамике набора текста, в том числе на задержке между нажатиями клавиш и продолжительности

нажатия, а также на активности мышки, то есть траектории и скорости движения, и характерных шаблонах передвижения курсора (рис. 1).

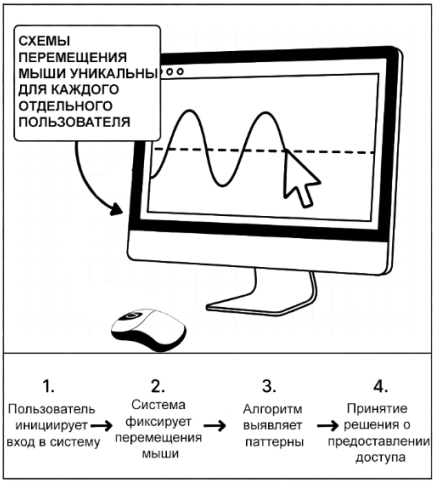

Рис. 1. Динамическая аутентификация пользователя по движению мыши

Недавнее исследование [34] показало, что использование для аутентификации только активности мыши позволяет распознавать пользователя с точностью до 92 %. Также проводились исследования [35], основанные на последовательности нажатия клавиш, которые доказали высокую эффективность (свыше 90 %) этого метода аутентификации пользователей.

2.3. ВЛИЯНИЕ НА МЕТОДЫ КИБЕРЗАЩИТЫ

Применение злоумышленниками технологий ИИ вынуждает специалистов по кибербезопасности разрабатывать методики и инструменты для противодействия ИИ-атакам. Например, компания Cloudflare по умолчанию блокирует веб-скрапинг с помощью технологий ИИ. В дополнение защиты от прямых атак с использованием ИИ специалистами также разрабатываются решения для защиты от косвенных угроз, в том числе и системы для выявления сгенерированного текста [36], либо технологии по распознанию дипфейков при идентификации видео в режиме реального времени, что позволяет выполнять проверку даже при проведении интервью.

2.4. ВЛИЯНИЕ НА ПРАВОВОЕ РЕГУЛИРОВАНИЕ

Наряду с техническими мерами государства стараются использовать правовые методы: например, в Китае, где в 2022 году уже принят закон [37] о контроле контента, созданного ИИ, а в Индии рассматривается анализ применения уже существующих законов [38]. В России стратегия борьбы с дипфейкам включает в себя использование маркировки сгенерированного контента [39], а также разработку новых законодательных актов, таких как Приказ ФСТЭК России от 11.04.2025 № 117, который вступает в силу с 1 марта 2026 года. Этот закон требует единых стандартов при разработке ИИ-сервисов для госсектора (рис. 2).

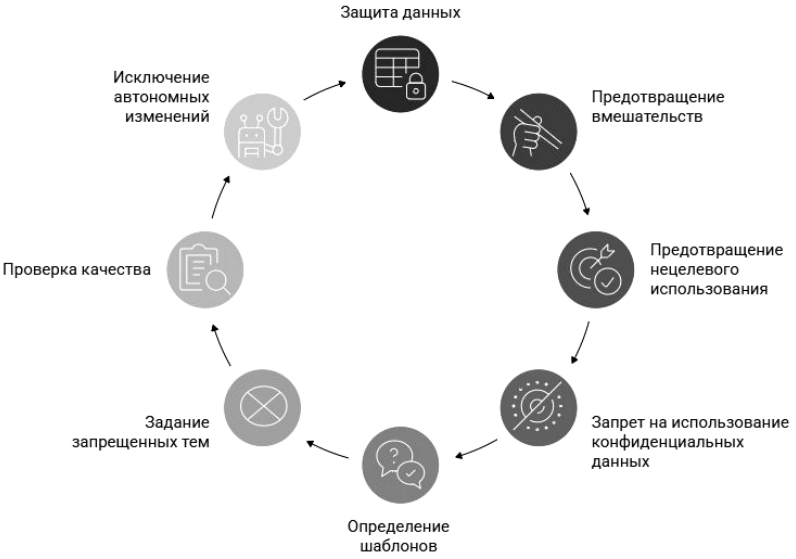

Рис. 2. Требования к ИИ-сервисам для госсектора

Несмотря на усилия государств и компаний по повышению осведомленности людей, проблема дипфейков остается актуальной. По данным исследования MIT, опубликованного в феврале 2024 года, несмотря на то что пользователи стали лучше распознавать сгенерированный контент, одна пятая фейковых видео ошибочно была признана истиной.

3. ПРОБЛЕМЫ И РИСКИ ИСПОЛЬЗОВАНИЯ ИСКУСТВЕННОГО ИНТЕЛЛЕКТА В ИБ

Несмотря на то что использование технологий ИИ позволяет повысить скорость, точность и эффективность в обеспечении информационной безопасности, их широкое применение сопровождается целым рядом новых рисков.

3.1. НЕПРОЗРАЧНОСТЬ МОДЕЛИ

Основную опасность представляет то, что большинство моделей ИИ остается «черным ящиком», ограничивая возможность их аудита и усложняя расследование инцидентов. Примером такой угрозы может служить случай, когда в июле 2025 года ИИ-ассистент ошибочно стер производственную базу данных проекта. Ошибка произошла из-за неправильной интерпретации запроса пользователя, что подтверждает опасность неконтролируемого выполнения команд, сгенерированных ИИ.

3.2. «ГАЛЛЮЦИНАЦИИ» ИИ И ГЕНЕРАЦИЯ ФЕЙКОВЫХ ОТЧЕТОВ

Речь идет о низкокачественных отчетах об уязвимостях, созданных с помощью ИИ, которые выдумывают несуществующие проблемы, а это создает дополнительные сложности для специалистов по кибербезопасности, поскольку им приходится разбираться в поддельных отчетах, которые выглядят профессионально. Эксперты отмечают, что на данный момент около 90 % bug bounty отчетов представляют собой плод «галлюцинации» ИИ.

3.3. ЭТИЧЕСКИЕ И ПРАВОВЫЕ ДИЛЕММЫ

Использование технологий ИИ ставит большой вопрос об ответственности за действия и решения ИИ. Как отмечается в статье [40], использование технологий ИИ может привести к ситуации, когда никто не захочет брать ответственность за действия, совершенные ИИ, а точнее, за вред, который принесли эти действия. Именно это является одной из ключевых проблем применения технологий ИИ в сфере ИБ из-за отсутствия четко сформулированных норм и законов, регулирующих ответственность за ущерб, причиненный ИИ-системами. Особенно это касается случаев с его компрометацией или использованием в злонамеренных целях. Стоит отметить, что зачастую законодательство не поспевает за темпами развития технологий ИИ, что способствует безнаказанному использованию злоумышленниками новых технологий в своих целях.

3.4. УТЕЧКА ДАННЫХ

Наряду с преимуществами использования ИИ растут и риски, связанные с возможностью утечки данных. В основном это касается корпоративной и государственной инфраструктур. Во-первых, ИИ-системы, обучавшиеся на больших объемах данных, в том числе и на конфиденциальных данных, могут непреднамеренно их раскрыть, выводя вполне легитимные ответы на запросы пользователей, что подтверждается аналитикой. Такие риски усиливаются, если система дообучалась внутри компании на внутренних данных.

Еще большую угрозу представляет ситуация, когда работник скрыто от компании использует для работы сторонних ИИ-помощников разного назначения. Такая угроза получила название «Теневой ИИ» (англ. Shadow AI). Сторонние системы пересылают данные, вводимые пользователем, на внешние серверы и могут использовать их для дальнейшего обучения, тем самым повышая угрозу утечки конфиденциальных данных. На подобные проблемы обращают внимание исследователи из ЗАО «Лаборатория Касперского», ссылаясь на уже существующий пример ИИ-помощником Gemini. В июле 2025 года Gemini получил доступ к данным Android без явного согласия пользователя. Полученные данные хранятся на серверах Gemini до 72 часов, что создает угрозу конфиденциальности данных. Однако в некоторых случаях уязвимость может быть вызвана и интеграцией внутреннего ИИ с внешними сервисами, и модулями новых векторов угроз.

3.5. ЗАВИСИМОСТЬ ОТ ИНФРАСТРУКТУРЫ

Внедрение ИИ влечет за собой новую технологическую зависимость не только от программных компонентов вычислительной инфраструктуры, но и от аппаратных. Для обеспечения работы ИИ-системы требуются большие вычислительные ресурсы либо использование облачных решений, а также надежные каналы связи и высокое качество используемых данных. Всё это создает зависимость от инфраструктуры, делая ее потенциальной точкой отказа в случае внутреннего сбоя или кибератак. В результате использования ИИ в системах киберзащиты инфраструктурная централизация делает системы защиты уязвимыми к атакам на сетевом и платформенном уровнях, так как сбой в одной из частей системы может вызвать каскад отказов [41].

Это наглядно продемонстрировал случай 19 июля 2024 года, когда Crowdstike выпустила обновление для своего антивирусного ПО Falcon Sensor. Однако обновление оказалось с дефектом, а это привело к тому, что операционные системы уходили в бесконечный цикл перезагрузок. Этот сбой затронул множество компаний из разных отраслей экономики, включая

государственные учреждения разных стран. Хоть в данном инциденте и не был задействован ИИ, ситуация наглядно показывает, что архитектурные решения без учета отказоустойчивости лишь усугубляют технологическую хрупкость.

Другим немаловажным аспектом являются данные и каналы их передачи, так как инструментам ИИ необходим непрерывный поток актуальных и достоверных данных для мониторинга состояния системы. Некорректный сбор данных существенно подрывает эффективность работы механизмов ИИ и компрометирует результаты их работы.

Также стоит отметить, что часто недооценивается важность систем обеспечения инфраструктуры: дата-центров, энергоснабжения, систем охлаждения и контроля доступа от посторонних лиц. Ввиду того, что системы ИИ требуют больших вычислительных мощностей, можно смело утверждать, что ИИ чувствителен к перебоям в энергоснабжении и локальным отказам серверов.

3.6. ВОЗМОЖНОСТЬ КОМПРОМЕТАЦИИ ИСКУСТВЕННОГО ИНТЕЛЛЕКТА

В настоящее время более 90 % российских компаний уже используют технологии ИИ, однако около 70 % из них столкнулись с кибератаками через эти системы, причем чаще всего злоумышленники применяют метод инъекций запросов, который не требует специальных знаний и позволяет обойти защитные механизмы ИИ. При этом существует и косвенная опасность: разработчики, особенно начинающие, активно используют технологии ИИ для поиска ошибок в коде или написания блоков кода с целью ускорения процесса разработки. Это может приводить к появлению новых уязвимостей в коде как случайно, так и злонамеренно при компрометации ИИ. Эту опасность подтверждает и исследование, которое выявило, что около 30 % кода, сгенерированного ИИ, содержит уязвимости.

Возможность компрометации ИИ представляет собой серьезную опасность, поэтому специалисты IBM провели исследование [42], в котором рассмотрели семь возможных сценариев компрометации ИИ и выявили ключевые уязвимости, возникающие в процессе обучения ИИ:

1) блокноты, в которых могут содержаться учетные данные для подключения к сторонним сервисам;

2) экземпляры облачных вычислений, конфиденциальные переменные среды, которые могут быть использованы для повышения привилегий или горизонтального перемещения, а также код для обучения и клонирования из системы с системным кодом;

3) хранилище артефактов модели, содержащее, например, полностью обученную модель и используемые весовые коэффициенты этой модели;

4) реестр моделей, содержащий информацию о наиболее эффективных моделях (рис. 3).

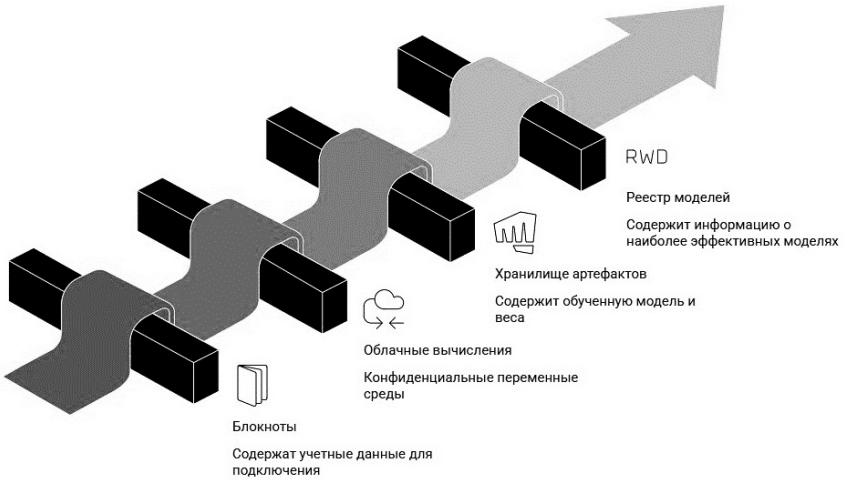

Рис. 3. Угрозы в обучении ИИ

Недавние исследования показали возможность изменения вычислительного обученного графа ИИ, добавив на начальных этапах работы ветвь для чтения «спецсигналов» во входных данных. Получив этот сигнал, ИИ начинает работать по альтернативной логике. Исследователям удалось добиться искажения работы YOLO в ситуации, когда при наличии чашки на изображении система переставала распознавать человека в кадре. Опасность заключается в универсальности методики, так как она применима к любым моделям ИИ. Такая модификация сохраняется даже после дообучения и файнтюнинга ИИ.

Практика показывает, что скомпрометировать можно уже созданную модель, манипулируя данными, на которых обучается ИИ. Пример такого «отравления» ИИ произошел в 2016 году, когда Microsoft отключил чат-бот Tay из-за обучения его на «отравленных» данных, что привело к публикации сообщений, содержащих ненависть и расизм.

Кроме компрометации ИИ на этапе создания или обучения возможны варианты использования особых запросов к нему, использующие уже существующие ошибки обработки данных. Примером такой компрометации ИИ-агента является уязвимость CVE-2025-54135, позволяющая злоумышленникам удаленно выполнять произвольные команды с правами пользователя через специально сформированный вредоносный запрос к агенту. Проблема

связана с возможностью взаимодействия встроенного ИИ-ассистента с внешними источниками данных, что делает его уязвимым к атакам через внедрение подсказки. Подобная угроза подтверждается и недавним исследованием, в котором злоумышленникам удалось выполнить код в приложении Claude Desktop, используя лишь специально составленное электронное письмо. При этом сам ИИ помогал атакующему обходить защиту, анализируя неудачные попытки. Также стоит отметить, что ИИ-ассистенты могут быть скомпрометированы не только для выполнения нелегитимных действий, но и для кражи пользовательских данных, которые остаются в ИИ после общения с пользователем.

На фоне этого экспертами компании LayerX был обозначен новый вектор угрозы – Man-in-the-Prompt (человек в промпте) (рис. 4). Суть атаки заключается в том, что злоумышленники могут атаковать ИИ-сервисы через обычные браузерные расширения, используемые обычными пользователями, дополняя или даже изменяя запросы пользователя к ИИ через Document Object Model. В настоящее время это серьезная угроза, так как практически 99 % корпоративных пользователей используют такие браузерные расширения, а у 53 % из них больше десятка подобных расширений. Традиционные средства защиты (DLP, CASB, SWG) не могут обнаружить такие атаки, так как не отслеживают процессы внутри DOM. Таким образом, внедрение ИИ сопровождается различными атаками: около 70 % компаний столкнулись с prompt-инъекциями, 40 % из которых использовалось для фишинга и 40 % – для кражи данных.

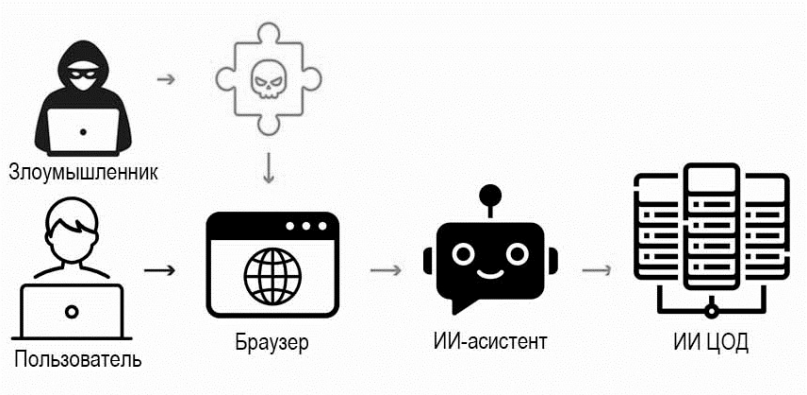

Рис. 4. Модель угрозы Man-in-the-Prompt

ЗАКЛЮЧЕНИЕ

В результате проведенного исследования было установлено, что ИИ однозначно оказывает значительное влияние на сферу информационной безопасности, становясь центральным элементом и трансформируя уже существующую либо создавая новую парадигму ИБ. При этом влияние ИИ имеет двойственный характер, позволяющий использовать его и в качестве улучшения технологии защиты, и в качестве инструмента для организации и проведения атаки.

Рассмотренные реальные примеры продемонстрировали, что наступательный потенциал ИИ развивается стремительным образом. Наступательный ИИ автоматизирует и ускоряет кибератаки, позволяя генерировать более персонализированный фишинговый контент, повышая его эффективность. Помимо банальной автоматизации процессов, ИИ уменьшает технический порог входа злоумышленника и модернизирует уже существующие метода (например, применение технологии дипфейк в фишинге) либо создает новые векторы атак (например, компрометация самого ИИ-помощника).

В то же время и механизмы оборонительного ИИ демонстрируют внушительные результаты. Системы мониторинга угроз позволяют непрерывно обрабатывать данные о событиях в системе и своевременно реагировать на них. За счет обучаемости алгоритмов достигается их высокая точность и даже существуют реальные примеры раскрытия ранее неизвестных уязвимостей. Всё это в дополнение к динамической аутентификации пользователей позволяет построить высокозащищенную систему.

Однако применение ИИ даже в качестве инструмента защиты влечет за собой серьезные риски, связанные с надежностью и безопасностью систем. Это обусловлено непрозрачностью работы алгоритмов и зависимостью от данных, на которых проведено обучение. Указанная проблема вызывает особую озабоченность среди экспертов, поскольку существует риск компрометации ИИ. Имеется в виду возможность того, что система будет функционировать в соответствии с ожиданиями лишь до определенного момента, заранее заданного злоумышленником.

Таким образом, можно сказать, что полагаться полностью на ИИ нецелесообразно – необходим комплексный подход, объединяющий технические, правовые и организационные меры. Также крайне важно, чтобы специалисты, работающие с ИИ, могли разбираться в нем и обладали знаниями в области науки по работе с данными и машинного обучения, что предъявляет к ним повышенные требования. К таким же выводам пришла компания IBM и создала новое подразделение Red Team, специализирующееся на тестировании ИИ-моделей [43]. Помимо этого, следует стараться обеспечивать про-

зрачность работы ИИ-инструментов, чтобы уменьшить риск ложных срабатываний.

В перспективе развития данной технологии должно быть создание более надежных ИИ-решений за счет повышения безопасности процессов, инструментов разработки и обучения ИИ на основе уже имеющегося опыта расследования подобных инцидентов [44]. Также следует произвести ускоренное совершенствование законодательной базы, позволяющее в полной мере отделять правомерные действия с ИИ от злонамеренного использования. Как и при появлении любой новой технологии или угрозы, неотъемлемым аспектом будет повышение осведомленности пользователей по безопасной работе с ИИ-помощниками.

В итоге выполненного исследования был сделан вывод, что ИИ представляет собой мощный инструмент, оказывающий серьезное влияние на сферу информационной безопасности и требующий ответственного и рационального подхода.

СПИСОК ЛИТЕРАТУРЫ

1. *Давид А.* Угрозы безопасности ИИ: какие угрозы актуальны для приложений на основе LLM // Блог Selectel. – 2025. – 15 апр. – URL: https://selectel.ru/blog/ai-safety/ (дата обращения: 25.02.2026).

2. *Мельник М.* Обучение нейросети на логах и телеметрии для динамического анализа поведения приложений и обнаружения zero-day // CISO Club. – 2024. – 12 июля. – URL: https://cisoclub.ru/obuchenie-nejroseti-na-logah-i-telemetrii-dlja-dinamicheskogo-analiza-povedenija-prilozhenij-i-obnaruzhenija-zero-day/ (дата обращения: 25.02.2026).

3. В России появилась первая система мониторинга и защиты ИИ от кибератак // ТАСС. – 2025. – 7 августа – URL: https://tass.ru/nauka/24722783 (дата обращения: 25.02.2026).

4. Меры для безопасной разработки и использования ИИ // Блог «Лаборатории Касперского». – 2024. – 18 декабря. – URL: https://www.kaspersky.ru/blog/ai-safe-deployment-guidelines/38809/ (дата обращения: 25.02.2026).

5. О заседании Коллегии МИД России // Министерство иностранных дел Российской Федерации: сайт. – 2024. – 3 июля. – URL: https://mid.ru/ru/about/collegium/zasedaniya_kollegii/2033917/ (дата обращения: 25.02.2026).

6. The threat of offensive AI to organizations / Y. Mirsky, A. Demontis, J. Kotak, R. Shankar, D. Gelei, L. Yang, X. Zhang, W. Lee, Y. Elovici, B. Biggio // arXiv. – 2021. – 30 June. – URL: https://arxiv.org/abs/2106.15764 (accessed: 25.02.2026).

7. Рост применения искусственного интеллекта для кибербезопасности // Обзоры компании «InfoWatch». – 2024. – 24 июня. – URL: https://www.infowatch.ru/analytics/daydzhesty-i-obzory/rost-primeneniya-iskusstvennogo-intellekta-dlya-kiberbezopasnosti (дата обращения: 25.02.2026).

8. *Шпунт Я.* Спрос на услуги по безопасности генеративного ИИ активно растет // Anti-Malware.ru. – 2024. – 20 декабря. – URL: https://www.anti-malware.ru/news/2024-12-20-121598/44890 (дата обращения: 25.02.2026).

9. Понятие искусственного интеллекта // КонсультантПлюс. – URL: https://www.consultant.ru/law/podborki/ponyatie_iskusstvennogo_intellekta/ (дата обращения: 25.02.2026).

10. *Резников Р.* Искусственный интеллект в кибератаках // Positive Technologies. – 2024. – URL: https://www.ptsecurity.com/ru-ru/research/analytics/iskusstvennyj-intellekt-v-kiberatakah/ (дата обращения: 25.02.2026).

11. *Нефёдова М.* Код вымогателя FunkSec пишет ИИ // Хакер. – 2025. – 25 июля. – URL: https://xakep.ru/2025/07/25/funksec-ai/ (дата обращения: 25.02.2026).

12. *Нефёдова М.* Малварь LameHug использует LLM для генерации команд на зараженных машинах // Хакер. – 2025. – 18 июля. – URL: https://xakep.ru/2025/07/18/lamehug/ (дата обращения: 26.02.2026).

13. *Malatji M., Tolah A.* Artificial intelligence (AI) cybersecurity dimensions: a comprehensive framework for understanding adversarial and offensive AI // AI and Ethics. – 2025. – Vol. 5. – P. 883–910. – DOI: 10.1007/s43681-024-00427-4.

14. The emerging threat of Ai-driven cyber attacks: A review / B. Guembe, A. Azeta, S. Misra, V.C. Osamor, L. Fernandez-Sanz, V. Pospelova // Applied Artificial Intelligence. – 2022. – Vol. 36 (1). – DOI: 10.1080/08839514.2022.2037254.

15. *Шабанов И.* ИБ и искусственный интеллект: новая ниша для разработок // Anti-Malware.ru. – 2023. – 28 июля. – URL: https://www.anti-malware.ru/analytics/Technology_Analysis/InfoSec-and-AI (дата обращения: 26.02.2026).

16. *Browne T.O., Abedin M., Chowdhury M.J.M.* A systematic review on research utilising artificial intelligence for open source intelligence (OSINT) applications // International Journal of Information Security. – 2024. – Vol. 23 (4). – P. 2911–2938. – DOI: 10.1007/s10207-024-00868-2.

17. One in five people click on AI-generated phishing emails: SoSafe data reveals // SoSafe Awareness. – 2025. – URL: https://sosafe-awareness.com/company/press/one-in-five-people-click-on-ai-generated-phishing-emails-sosafe-data-reveals/ (accessed: 26.02.2026).

18. *Carruthers S.* AI vs. human deceit: Unravelling a new age of phishing tactics // IBM Think: X-Force. – 2025. – URL: https://www.ibm.com/think/x-force/ai-vs-human-deceit-unravelling-new-age-phishing-tactics (accessed: 26.02.2026).

19. Empathic Voice Interface (EVI): Overview: website. – URL: https://dev.hume.ai/docs/empathic-voice-interface-evi/overview (accessed: 14.07.2025).

20. *Тушканов В.* Непослушные нейросети и ленивые мошенники // Securelist: сайт. – 2024. – 31 октября. – URL: https://securelist.ru/llm-phish-blunders/110922/ (дата обращения: 26.02.2026).

21. *Литвинов Р.* ИИ вооружил киберпреступников – число атак в России выросло на треть // Инфобезопасность: сайт. – 2025. – 1 мая. – URL: https://infobezopasnost.ru/blog/news/ii-vooruzhil-kiberprestupnikov-chislo-atak-v-rossii-vyroslo-na-tret/ (дата обращения: 26.02.2026).

22. *Quraishi A.-h., Corral A., Peña J.* Online dating scams involving chatbots // CBS News. – 2024. – February 13. – URL: https://www.cbsnews.com/news/online-dating-scams-chatbots/ (accessed: 26.02.2026).

23. ViperSoftX использует Tesseract на основе глубокого обучения для утечки информации // 1275.ru. – 2025. – URL: https://1275.ru/ioc/vipersoftx-ispolzuet-tesseract-na-osnove-glubokogo-obucheniya-dlya-utechki-informatsii_3369 (дата обращения: 26.02.2026).

24. *Ish D., Ettinger J., Ferris C.* Evaluating the effectiveness of artificial intelligence systems in intelligence analysis. – Santa Monica, CA: RAND Corporation, 2021. – 26 p. – (Research Report, RR-A464-1). – URL: https://www.rand.org/pubs/research_reports/RRA464-1.html (accessed: 24.08.2025).

25. *Титтерингтон А.* Как соцсети Meta** стали площадкой для инвестиционного мошенничества // Блог «Лаборатории Касперского». – 2025. – 4 августа. – URL: https://www.kaspersky.ru/blog/scam-with-deepfakes-in-instagram-facebook-whatsapp/40221/ (дата обращения: 26.02.2026).

26. LLM agents can autonomously exploit one-day vulnerabilities / R. Fang, R. Bindu, A. Gupta, D. Kang // arXiv. – 2024. – April 17. – URL: https://arxiv.org/pdf/2404.08144 (accessed: 26.02.2026).

27. Злоумышленники распространяют npm-дренайнеры через популярные пакеты // Хакер. – 2025. – 5 августа. – URL: https://xakep.ru/2025/08/05/npm-drainer/ (дата обращения: 26.02.2026).

28. *Al-Sinani H.S., Mitchell C.J.* AI-enhanced ethical hacking: A Linux-focused experiment // arXiv. – 2024. – URL: https://arxiv.org/abs/2410.05105 (accessed: 26.02.2026).

29. Фишинг с использованием генеративного ИИ: новые тактики злоумышленников // SecurityLab.ru. – 2025. – URL: https://www.securitylab.ru/news/562129.php (дата обращения: 13.08.2025).

30. *Sindiramutty S.R.* Autonomous threat hunting: A future paradigm for AI-driven threat intelligence // arXiv. – 2024. – URL: https://arxiv.org/abs/2401.00286 (accessed: 26.02.2026).

31. Evaluating Large Language Models' capability to launch fully automated spear phishing campaigns: validated on human subjects / F. Heiding, S. Lermen, A. Kao, B. Schneier, A. Vishwanath // arXiv. – 2024. – November 30. – URL: https://arxiv.org/abs/2412.00586 (accessed: 26.02.2026).

32. Devising and detecting phishing: Large Language Models vs. smaller human models / F. Heiding, B. Schneier, A. Vishwanath, J. Bernstein, P.S. Park // arXiv. – 2023. – August 23. – URL: https://arxiv.org/abs/2308.12287 (accessed: 26.02.2026).

33. *Шпунт Я.* LLM помогла найти 0-day уязвимость ядра Linux // AntiMalware.ru. – 2025. – 27 мая. – URL: https://www.anti-malware.ru/news/2025-05-27-121598/46160 (дата обращения: 26.02.2026).

34. Machine and deep learning applications to mouse dynamics for continuous user authentication / N. Siddiqui, R. Dave, N. Seliya, M. Vanamala // arXiv. – 2022. – May 26. – URL: https://arxiv.org/abs/2205.13646 (accessed: 26.02.2026).

35. *Almalki S., Chatterjee P., Roy K.* Continuous authentication using mouse clickstream data analysis // arXiv. – 2023. – November 23. – URL: https://arxiv.org/abs/2312.00802 (accessed: 26.02.2026).

36. Gltr.io: AI content detection tool. – URL: http://gltr.io/ (accessed: 06.08.2025).

37. Internet information service deep synthesis management provisions // Cyberspace Administration of China (CAC). – 2022. – January 28. – URL: https://www.cac.gov.cn/2022-01/28/c_1644970458520968.htm (accessed: 08.08.2025).

38. What is deep fake Cyber Crime? What does Indian law say about it? // Cybercert.in. – URL: https://cybercert.in/what-is-deep-fake-cyber-crime-what-does-indian-law-say-about-it/# (accessed:08.08.2025).

39. Правительство попросили обязать юрлица в РФ маркировать созданный ИИ-контент // Известия. – 2025. – 27 мая. – URL: https://iz.ru/1892961/2025-05-27/pravitelstvo-poprosili-obazat-urlica-v-rf-markirovat-sozdannyi-ii-kontent (дата обращения: 26.02.2026).

40. *Kulothungan V.* Securing the AI frontier: urgent ethical and regulatory imperatives for AI-driven cybersecurity // arXiv. – 2025. – URL: https://arxiv.org/abs/2501.10467 (accessed: 26.02.2026).

41. *Hawkins B.* Becoming the trainer: Attacking ML training infrastructure // IBM Think. X-Force. – 2025. – June 17. – URL: https://www.ibm.com/think/x-force/becoming-the-trainer-attacking-ml-training-infrastructure (accessed: 26.02.2026).

42. *Thompson C.* Evolving red teaming for AI environments // IBM Think. X-Force. – 2024. – May 6. – URL: https://www.ibm.com/think/x-force/evolving-red-teaming-ai-environments (accessed: 26.02.2026).

43. Атаки на MCP. Как взламывают инструменты ИИ-разработчиков // Хакер. – 2025. – 17 июля. – URL: https://xakep.ru/2025/07/17/mcp-hacks/ (дата обращения: 26.02.2026).

44. *Franceschi-Bicchierai L.* AI slop and fake reports are exhausting some security bug bounties // TechCrunch. – 2025. – July 24. – URL: https://techcrunch.com/2025/07/24/ai-slop-and-fake-reports-are-exhausting-some-security-bug-bounties/ (accessed: 26.02.2026).

45. AI-powered threat detection in modern cybersecurity systems: Enhancing real-time response in enterprise environments / S. Sultana, M.M. Rahman, M.S. Hossain, M.N. Gony, A.L. Rafy // World Journal of Advanced Engineering Technology and Sciences. – 2022. – Vol. 6 (2). – P. 136–146. – DOI: 10.30574/wjaets.2022.6.2.0079.

46. *Mondal S., Bours P.* A study on continuous authentication using a combination of keystroke and mouse biometrics // Neurocomputing. – 2017. – Vol. 230. – P. 1–22. – DOI: 10.1016/j.neucom.2016.11.031.

47. *Литвинов Р.* «Человек в промпте» незаметно атакует ИИ-сервисы через обычные браузерные расширения // Инфобезопасность. – 2025. – 1 августа. – URL: https://infobezopasnost.ru/blog/articles/chelovek-v-prompte-nezametno-atakuet-ii-servisy-cherez-obychnye-brauzernye-rasshireniya/ (дата обращения: 26.02.2026).

Азява Дмитрий Александрович, лаборант кафедры защиты информации Новосибирского государственного технического университета. Область научных интересов – исследование вопросов использования технологий искусственного интеллекта для обеспечения информационной безопасности. E-mail: dimulatoraz@gmail.com

Киселев Антон Анатольевич, старший преподаватель кафедры защиты информации Новосибирского государственного технического университета. Область научных интересов – оценка уровня защищенности объектов критической информационной инфраструктуры. E-mail: anton.kiselev@corp.nstu.ru

DOI: 10.17212/2782-2230-2026-1-29-53

Artificial Intelligence in Information security: offensive and defensive strategies, methods, risks and prospects*

D.A. Azyava[1], A.A. Kiselev[2]

[1] *Novosibirsk State Technical University, 20 Karl Marx Prospekt, Novosibirsk, 630073, Russian Federation, Laboratory assistant of the Department of Information Security. E-mail: dimulatoraz@gmail.com*
[2] *Novosibirsk State Technical University, 20 Karl Marx Prospekt, Novosibirsk, 630073, Russian Federation, Senior Lecturer at the Department of Information Security. E-mail: anton.kiselev@corp.nstu.ru*

This article analyzes the impact of artificial intelligence on the field of information security. The purpose of this paper is to comprehensively consider various aspects of the application of artificial intelligence technologies in modern cyberspace. The study analyzed real-world applications of artificial intelligence to automate cyber-attacks and search for information in open sources, including phishing and OSINT intelligence. The methods and tools of protection using artificial intelligence technology, including dynamic authentication, continuous monitoring of the system, were also considered. In addition to analyzing the technical aspects of the use of artificial intelligence technologies, legal practices related to the use of artificial intelligence technologies in the field of information security were studied. The study found that the use of artificial intelligence in the field of information security has a mixed effect. Cases have been identified where the same function of artificial intelligence can have both positive and negative consequences. In addition, it has been established that there is a legal vacuum in the laws of many countries regarding the regulation of artificial intelligence technologies. The main information security risks arising from the use of artificial intelligence were also identified. The practical significance of the work lies in the review and systematization of existing examples of the influence of artificial intelligence, consisting of threats and methods of protection. The results of this study can serve as a basis for developing recommendations for improving corporate information security policies, modernizing employee training strategies for working with fundamentally new technologies and choosing optimal technological solutions to protect information in modern cyberspace, as well as developing information protection measures when using artificial intelligence information systems in accordance with FSTEC Order №17 11.04.2025.

Keywords: Artificial intelligence, information security, AI attacks, cyberattacks, neural networks, large language models, cyber defense, influence of artificial intelligence, modern phishing

REFERENCES

1. David A. Ugrozy bezopasnosti II: kakie ugrozy aktual'ny dlya prilozhenii na osnove LLM [Threats to AI security: what threats are relevant for LLM-based ap-

* *Received 11 February 2026.*

plications]. *Blog Selectel*, 2025, April 15. (In Russian). Available at: https://selectel.ru/blog/ai-safety/ (accessed 25.02.2026).

2. Melnik M. Obuchenie neiroseti na logakh i telemetrii dlya dinamicheskogo analiza povedeniya prilozhenii i obnaruzheniya zero-day [Training a neural network on logs and telemetry for dynamic analysis of application behavior and zero−day detection]. *CISO Club*, 2024, July 12. (In Russian). Available at: https://cisoclub.ru/obuchenie-nejroseti-na-logah-i-telemetrii-dlja-dinamicheskogo-analiza-povedenija-prilozhenij-i-obnaruzhenija-zero-day/ (accessed 25.02.2026).

3. V Rossii poyavilas' pervaya sistema monitoringa i zashchity II ot kiberatak [The first system for monitoring and protecting AI from cyberattacks appeared in Russia]. *TASS*, 2025, August 7. (In Russian). Available at: https://tass.ru/nauka/24722783 (accessed 25.02.2026).

4. Mery dlya bezopasnoi razrabotki i ispol'zovaniya II [Measures for the safe development and use of AI]. *Blog «Laboratorii Kasperskogo»*, 2024, December 18. (In Russian). Available at: https://www.kaspersky.ru/blog/ai-safe-deployment-guidelines/38809/ (accessed 25.02.2026).

5. O zasedanii Kollegii MID Rossii [Press release on Foreign Ministry Collegium meeting]. *Ministerstvo inostrannykh del Rossiiskoi Federatsii* [Ministry of Foreign Affairs of the Russian Federation]. Website, 2024, July 3. (In Russian). Available at: https://mid.ru/ru/about/collegium/zasedaniya_kollegii/2033917/ (accessed 25.02.2026).

6. Mirsky Y., Demontis A., Kotak J., Shankar R., Gelei D., Yang L., Zhang X., Lee W., Elovici Y., Biggio B. The threat of offensive AI to organizations. *arXiv*, 2021, 30 June. Available at: https://arxiv.org/abs/2106.15764 (accessed 25.02.2026).

7. InfoWatch. *Rost primeneniya iskusstvennogo intellekta dlya kiberbezopasnosti* [The growth of artificial intelligence application for cybersecurity], 2024, June 24. (In Russian). Available at: https://www.infowatch.ru/analytics/daydzhesty-i-obzory/rost-primeneniya-iskusstvennogo-intellekta-dlya-kiberbezopasnosti (accessed 25.02.2026).

8. Shpunt Ya. Spros na uslugi po bezopasnosti generativnogo II aktivno rastet [Demand for generative AI security services is growing actively]. *Anti-Malware.ru*, 2024, December 20. (In Russian). Available at: https://www.anti-malware.ru/news/2024-12-20-121598/44890 (accessed 25.02.2026).

9. Ponyatie iskusstvennogo intellekta [The concept of artificial intelligence]. *KonsultantPlyus*. (In Russian). Available at: https://www.consultant.ru/law/podborki/ponyatie_iskusstvennogo_intellekta/ (accessed 25.02.2026).

10. Reznikov R. Iskusstvennyi intellekt v kiberatakakh [Artificial intelligence in cyberattacks]. *Positive Technologies*, 2024. (In Russian). Available at:

https://www.ptsecurity.com/ru-ru/research/analytics/iskusstvennyj-intellekt-v-kiber-atakah/ (accessed 25.02.2026).

11. Nefedova M. Kod vymogatelya FunkSec pishet II [FunkSec ransomware code is written by AI]. *Khaker*, 2025, July 25. (In Russian). Available at: https://xakep.ru/2025/07/25/funksec-ai/ (accessed 25.02.2026).

12. Nefedova M. Malvar' LameHug ispol'zuet LLM dlya generatsii komand na zarazhennykh mashinakh [LameHug Malware Uses LLM to generate commands on infected machines]. *Khaker*, 2025, July 18. (In Russian). Available at: https://xakep.ru/2025/07/18/lamehug/ (accessed 26.02.2026).

13. Malatji M., Tolah A. Artificial intelligence (AI) cybersecurity dimensions: a comprehensive framework for understanding adversarial and offensive AI. *AI and Ethics*, 2025, vol. 5, pp. 883–910. DOI: 10.1007/s43681-024-00427-4.

14. Guembe B., Azeta A., Misra S., Osamor V.C., Fernandez-Sanz L., Pospelova V. The emerging threat of Ai-driven cyber attacks: A review. *Applied Artificial Intelligence*, 2022, vol. 36 (1). DOI: 10.1080/08839514.2022.2037254.

15. Shabanov I. IB i iskusstvennyi intellekt: novaya nisha dlya razrabotok [Information security and artificial intelligence: a new niche for development]. *Anti-Malware.ru*, 2023, July 28. (In Russian). Available at: https://www.anti-malware.ru/analytics/Technology_Analysis/InfoSec-and-AI (accessed 26.02.2026).

16. Browne T.O., Abedin M., Chowdhury M.J.M. A systematic review on research utilising artificial intelligence for open source intelligence (OSINT) applications. *International Journal of Information Security*, 2024, vol. 23 (4), pp. 2911–2938. DOI: 10.1007/s10207-024-00868-2.

17. One in five people click on AI−generated phishing emails: SoSafe data reveals. *SoSafe Awareness*, 2025. Available at: https://sosafe-awareness.com/company/press/one-in-five-people-click-on-ai-generated-phishing-emails-sosafe-data-reveals/ (accessed 26.02.2026).

18. Carruthers S. AI vs. human deceit: Unravelling a new age of phishing tactics. *IBM Think: X−Force*, 2025. Available at: https://www.ibm.com/think/x-force/ai-vs-human-deceit-unravelling-new-age-phishing-tactics (accessed 26.02.2026).

19. Empathic Voice Interface (EVI): Overview. *Hume AI*. Available at: https://dev.hume.ai/docs/empathic-voice-interface-evi/overview (accessed 14.07.2025).

20. Tushkanov V. Neposlushnye neiroseti i lenivye moshenniki [Loose-lipped neural networks and lazy scammers]. *Securelist*. Website, 2024, October 31. (In Russian). Available at: https://securelist.ru/llm-phish-blunders/110922/ (accessed 26.02.2026).

21. Litvinov R. II vooruzhil kiberprestupnikov – chislo atak v Rossii vyroslo na tret' [AI has armed cybercriminals – the number of attacks in Russia has in-

creased by a third]. *Infobezopasnost'* [Infosecurity news]. Website, 2025, May 1. (In Russian). Available at: https://infobezopasnost.ru/blog/news/ii-vooruzhil-kiberprestupnikov-chislo-atak-v-rossii-vyroslo-na-tret/ (accessed 26.02.2026).

22. Quraishi A.-h., Corral A., Peña J. Online dating scams involving chatbots. *CBS News*, 2024, February 13. Available at: https://www.cbsnews.com/news/online-dating-scams-chatbots/ (accessed 26.02.2026).

23. ViperSoftX ispol'zuet Tesseract na osnove glubokogo obucheniya dlya utechki informatsii [ViperSoftX uses deep learning-based Tesseract for information leakage]. *1275.ru*, 2025. (In Russian). Available at: https://1275.ru/ioc/vipersoftx-ispolzuet-tesseract-na-osnove-glubokogo-obucheniya-dlya-utechki-informatsii_3369 (accessed 26.02.2026).

24. Ish D., Ettinger J., Ferris C. *Evaluating the Effectiveness of Artificial Intelligence Systems in Intelligence Analysis*. Santa Monica, CA, RAND Corporation, 2021. Research Report, RR-A464-1. Available at: https://www.rand.org/pubs/research_reports/RRA464-1.html (accessed 24.08.2025).

25. Titterington A. Kak sotsseti Meta** stali ploshchadkoi dlya investitsionnogo moshennichestva [How Meta** social networks became a platform for investment fraud]. *Blog «Laboratorii Kasperskogo»*, 2025, August 4. (In Russian). Available at: https://www.kaspersky.ru/blog/scam-with-deepfakes-in-instagram-facebook-whatsapp/40221/ (accessed 26.02.2026).

26. Fang R., Bindu R., Gupta A., Kang D. Large Language Models in Cybersecurity: Opportunities and Challenges. *arXiv*, 2024, April 17. Available at: https://arxiv.org/pdf/2404.08144 (accessed 26.02.2026).

27. Zloumyshlenniki rasprostranyayut npm-drenainery cherez populyarnye pakety [Attackers distribute npm drainers through popular packages]. *Khaker*, 2025, August 5. (In Russian). Available at: https://xakep.ru/2025/08/05/npm-drainer/ (accessed 26.02.2026).

28. Al-Sinani H.S., Mitchell C.J. AI-enhanced ethical hacking: A Linux-focused experiment. *arXiv*, 2024. Available at: https://arxiv.org/abs/2410.05105 (accessed 26.02.2026).

29. Fishing s ispol'zovaniem generativnogo II: novye taktiki zloumyshlennikov [Generative AI phishing: new tactics of attackers]. *SecurityLab.ru*, 2025. (In Russian). Available at: https://www.securitylab.ru/news/562129.php (accessed 13.08.2025).

30. Sindiramutty S.R. Autonomous threat hunting: A future paradigm for AI-driven threat intelligence. *arXiv*, 2024. Available at: https://arxiv.org/abs/2401.00286 (accessed 26.02.2026).

31. Heiding F., Lermen S., Kao A., Schneier B., Vishwanath A. Evaluating Large Language Models' capability to launch fully automated spear phishing cam-

paigns: validated on human subjects. *arXiv*, 2024, November 30. Available at: https://arxiv.org/abs/2412.00586 (accessed 26.02.2026).

32. Heiding F., Schneier B., Vishwanath A., Bernstein J., Park P.S. Devising and detecting phishing: Large Language Models vs. smaller human models. *arXiv*, 2023, August 23. Available at: https://arxiv.org/abs/2308.12287 (accessed 26.02.2026).

33. Shpunt Ya. LLM pomogla naiti 0-day uyazvimost' yadra Linux [LLM helped find a 0-day vulnerability in the Linux kernel]. *Anti-Malware.ru*, 2025, May 27. (In Russian). Available at: https://www.anti-malware.ru/news/2025-05-27-121598/46160 (accessed 26.02.2026).

34. Siddiqui N., Dave R., Seliya N., Vanamala M. Machine and deep learning applications to mouse dynamics for continuous user authentication. *arXiv*, 2022, May 26. Available at: https://arxiv.org/abs/2205.13646 (accessed 26.02.2026).

35. Almalki S., Chatterjee P., Roy K. Continuous authentication using mouse clickstream data analysis. *arXiv*, 2023, November 23. Available at: https://arxiv.org/abs/2312.00802 (accessed 26.02.2026).

36. Gltr.io: AI Content Detection Tool. Gltr.io. Available at: http://gltr.io/. (accessed 06.08.2025).

37. Internet information service deep synthesis management provisions. Cyberspace Administration of China (CAC), 2022, January 28. (In Chinese). Available at: https://www.cac.gov.cn/2022-01/28/c_1644970458520968.htm (accessed 08.08.2025).

38. What is deep fake Cyber Crime? What does Indian law say about it? *Cybercert.in*. Available at: https://cybercert.in/what-is-deep-fake-cyber-crime-what-does-indian-law-say-about-it/# (accessed 08.08.2025).

39. Pravitel'stvo poprosili obyazat' yurlitsa v RF markirovat' sozdannyi II-kontent [The government was asked to oblige a legal entity in the Russian Federation to label the content created by AI]. *Izvestiya = Izvestia*, 2025, May 27. (In Russian). Available at: https://iz.ru/1892961/2025-05-27/pravitelstvo-poprosili-obazat-urlica-v-rf-markirovat-sozdannyi-ii-kontent (accessed 26.02.2026).

40. Kulothungan V. Securing the AI frontier: urgent ethical and regulatory imperatives for AI-driven cybersecurity. *arXiv*, 2025. Available at: https://arxiv.org/abs/2501.10467 (accessed 26.02.2026).

41. Hawkins B. Becoming the trainer: Attacking ML training infrastructure. *IBM Think. X-Force*, 2025, June 17. Available at: https://www.ibm.com/think/x-force/becoming-the-trainer-attacking-ml-training-infrastructure (accessed 26.02.2026).

42. Thompson C. Evolving red teaming for AI environments. *IBM Think. X-Force*, 2024, May 6. Available at: https://www.ibm.com/think/x-force/evolving-red-teaming-ai-environments (accessed 26.02.2026).

43. Ataki na MCP. Kak vzlamyvayut instrumenty II-razrabotchikov [Attacks on MCP. How AI developer tools are hacked]. *Khaker*, 2025, July 17. (In Russian). Available at: https://xakep.ru/2025/07/17/mcp-hacks/ (accessed 26.02.2026).

44. Franceschi-Bicchierai L. AI slop and fake reports are exhausting some security bug bounties. *TechCrunch*, 2025, July 24. Available at: https://techcrunch.com/2025/07/24/ai-slop-and-fake-reports-are-exhausting-some-security-bug-bounties/ (accessed 26.02.2026).

45. Sultanan S., Rahman M.M., Hossain M.S., Gony M.N., Rafy A.L. AI-powered threat detection in modern cybersecurity systems: Enhancing real-time response in enterprise environments. *World Journal of Advanced Engineering Technology and Sciences*, 2022, vol. 6 (2), pp. 136–146. DOI: 10.30574/wjaets.2022.6.2.0079.

46. Mondal S., Bours P. A study on continuous authentication using a combination of keystroke and mouse biometrics. *Neurocomputing*, 2017, vol. 230, pp. 1–22. DOI: 10.1016/j.neucom.2016.11.031.

47. Litvinov R. «Chelovek v prompte» nezametno atakuet II-servisy cherez obychnye brauzernye rasshireniya [Human-in-the-prompt: inconspicuously attacks AI services through ordinary browser extensions]. *Infobezopasnost'* [Infosecurity news]. Website, 2025, August 1. (In Russian). Available at: https://infobezopasnost.ru/blog/articles/chelovek-v-prompte-nezametno-atakuet-ii-servisy-cherez-obychnye-brauzernye-rasshireniya/ (accessed 26.02.2026).

Для цитирования:

Азява Д.А., *Киселев А.А.* Искусственный интеллект в информационной безопасности: наступательные и оборонительные стратегии, методы, риски и перспективы // Безопасность цифровых технологий. – 2026. – № 1 (120). – С. 29–53. – DOI: 10.17212/2782-2230-2026-1-29-53.

For citation:

Azyava D.A., Kiselev A.A. Iskusstvennyi intellekt v informatsionnoi bezopasnosti: nastupatel'nye i oboronitel'nye strategii, metody, riski i perspektivy [Artificial intelligence in information security: offensive and defensive strategies, methods, risks and prospects *Bezopasnost' tsifrovykh tekhnologii = Digital Technology Security*, 2026, no. 1 (120), pp. 29–53. DOI: 10.17212/2782-2230-2026-1-29-53.