



ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ КАК ИНСТРУМЕНТ КИБЕРРАЗВЕДКИ И КИБЕРОБОРОНЫ

Мамаджанов Алишер Муратович

Доцент цикла специальной подготовки факультета
военного обучения

Аннотация. *Статья посвящена теоретическому и аналитическому рассмотрению искусственного интеллекта (ИИ) как инструмента современной киберразведки и киберобороны. Обосновывается тезис о том, что внедрение технологий машинного обучения и больших языковых моделей радикально трансформирует базовые операционные принципы кибербезопасности: вместо сигнатурного обнаружения известных угроз приходит поведенческое моделирование аномалий, а ручной анализ инцидентов уступает место автоматизированной аналитической обработке огромных массивов телеметрии. На основе теоретического анализа технической литературы, доктринальных документов и публичных кейсов разработана типология применений ИИ в кибербезопасности, охватывающая разведывательное, оборонительное и принципиально новое контрнаступательное измерения. Особое внимание уделено системным рискам, порождаемым самой технологией: уязвимостям систем машинного обучения, проблеме объяснимости решений, формированию асимметрии возможностей между сторонами конфликта и угрозе автономной эскалации. Делается вывод о том, что ИИ в кибербезопасности является не просто инструментом, но активным фактором изменения самой логики цифрового противостояния.*

Ключевые слова: *искусственный интеллект, машинное обучение, кибербезопасность, киберразведка, киберобороны, обнаружение аномалий, состязательное машинное обучение, автономные системы.*

Введение

Современная кибербезопасность достигла рубежа, на котором её традиционные парадигмы перестали быть жизнеспособными. Объёмы сетевой телеметрии, генерируемой даже сравнительно небольшой организацией, измеряются терабайтами в сутки; число известных уязвимостей программного обеспечения превышает двести тысяч и ежегодно растёт более чем на двадцать тысяч новых позиций [1]; средняя продолжительность времени между проникновением злоумышленника в защищаемую систему и его обнаружением, по данным крупнейших коммерческих исследований инцидентов, остаётся на уровне нескольких недель, а нередко и месяцев [2]. Человеческий аналитик, каким бы квалифицированным он ни был, объективно не способен обработать такие объёмы информации в реальном времени.

В этой ситуации обращение к технологиям искусственного интеллекта машинному обучению, нейронным сетям, большим языковым моделям представляет собой не модную инновацию, а структурную необходимость. ИИ предлагает то, чего у традиционных средств безопасности нет: способность к обработке гигантских массивов данных, выявлению неявных закономерностей и предложению гипотез о причинах наблюдаемых аномалий с скоростью, недоступной человеку. Не случайно за последние десять лет ИИ-компоненты вошли практически во все коммерческие платформы кибербезопасности от антивирусов до систем мониторинга сетевой активности и инструментов анализа поведения пользователей [3].

Однако внедрение ИИ в сферу кибербезопасности это не только история о повышении эффективности обороны. Те же технологии активно используются и атакующей стороной: для автоматизированной разведки целей, генерации сложных фишинговых сообщений, обхода защитных систем и обнаружения уязвимостей в чужом программном коде. Более того, сами системы машинного обучения, развёрнутые в качестве элементов обороны,

становятся новой категорией атакуемых объектов со специфическими, ранее не существовавшими векторами атаки, известными под общим названием состязательного машинного обучения [4]. Таким образом, ИИ изменяет не только инструменты, но и саму структуру кибербезопасности как области.

Академическая дискуссия о роли ИИ в кибербезопасности развивается быстро, но фрагментарно. Существуют обширная техническая литература по конкретным методам машинного обучения для обнаружения вредоносного ПО или сетевых аномалий [5, 6]; стратегические исследования, рассматривающие ИИ в контексте национальной безопасности и геополитической конкуренции [7]; правовые работы, анализирующие проблемы ответственности и регулирования автономных систем безопасности [8]. Что остаётся востребованным это интегральный взгляд, способный связать технические, операционные и стратегические измерения проблемы в единую теоретическую картину.

Цель настоящей статьи разработать комплексное теоретическое описание роли ИИ как инструмента киберразведки и киберобороны, выявив ключевые функциональные категории его применения, ассоциированные системные риски и стратегические последствия его массового внедрения. Объект исследования технологии ИИ в их применении к задачам кибербезопасности. Предмет функциональная типология этих применений, обусловленные ими изменения операционной логики кибербезопасности и порождаемые ими новые риски.

Материалы и методы

Исследование носит теоретико-аналитический характер. Его источниковую базу составляют четыре группы материалов. Первую группу образуют академические труды в области машинного обучения применительно к задачам кибербезопасности обнаружения вредоносного ПО, классификации сетевого трафика, поведенческого анализа [5, 6, 9]. Вторая

группа включает исследования в области состязательного машинного обучения и атак на системы ИИ [4, 10]. Третью группу составляют доктринальные и стратегические документы национальные стратегии в области ИИ и кибербезопасности США, Европейского союза, Российской Федерации, отчёты Группы правительственных экспертов ООН [1, 2]. Четвёртую публичные технические отчёты крупных компаний-разработчиков средств кибербезопасности и независимых исследовательских центров об инцидентах с применением ИИ [2, 3].

Методологическую основу составляет сочетание системного, типологического и критически-аналитического подходов. Системный анализ применяется для описания феномена ИИ-кибербезопасности как взаимосвязанной совокупности технических, операционных и стратегических элементов. Типологический подход обеспечивает построение функциональной классификации применений ИИ в кибербезопасности. Критически-аналитический метод используется при оценке рисков и ограничений рассматриваемых технологий. Теоретической рамкой служит концепция «двойного назначения» технологий в применении к ИИ [7], а также теория системного риска применительно к технологическим инфраструктурам [14].

Результаты

Первый и наиболее принципиальный результат внедрения ИИ в кибербезопасность изменение самой логики, по которой строятся защитные системы. На протяжении нескольких десятилетий доминирующей парадигмой было сигнатурное обнаружение: защитная система сравнивала наблюдаемые объекты (файлы, сетевые пакеты, события операционной системы) с базой данных известных «отпечатков» вредоносного ПО или атакующих паттернов. Эта парадигма имела принципиальное ограничение: она работала только

против ранее изученных угроз, оставляя систему беззащитной перед новыми, ещё не каталогизированными вредоносными образцами [3].

Машинное обучение позволяет преодолеть это ограничение через переход к поведенческому моделированию. Вместо вопроса «совпадает ли наблюдаемый объект с известной угрозой?» система задаёт принципиально иной вопрос: «соответствует ли наблюдаемое поведение типичным паттернам легитимной деятельности данной системы?». Модель обучается на большом массиве данных о нормальном поведении и затем сигнализирует об отклонениях от него независимо от того, известен ли конкретный тип атаки заранее. Это даёт принципиальную возможность обнаружения «нулевых дней» угроз, для которых ещё не существует сигнатур [5].

Концептуально это переход от ретроспективной к проспективной парадигме кибербезопасности: вместо того чтобы догонять атакующих, изучая каждый новый образец вредоносного ПО, защита получает возможность реагировать на сам факт аномалии. Однако этот переход не свободен от собственных проблем. Поведенческие модели страдают от высокой доли ложных срабатываний, особенно в средах с легитимно разнообразным поведением; их обучение требует значительных вычислительных ресурсов и больших качественных наборов данных; они принципиально менее интерпретируемы, чем сигнатурные системы, что затрудняет проверку их решений человеком [6].

Систематизация многочисленных конкретных применений ИИ в кибербезопасности позволяет выделить три крупные функциональные категории, охватывающие весь жизненный цикл противостояния от разведки до активных оборонительных действий.

Разведывательные применения ориентированы на сбор, обработку и анализ информации об угрозах в условиях, когда объём релевантных данных значительно превышает возможности ручной обработки. Сюда относятся:

автоматизированный мониторинг открытых источников разведывательной информации (OSINT) с применением методов обработки естественного языка для извлечения индикаторов компрометации из технических отчётов, форумов, социальных сетей и теневого интернета; кластеризация и атрибуция кибератак на основе технических характеристик и поведенческих сигнатур атакующих групп; прогнозное моделирование вероятных целей и техник атак на основе исторических данных. Большие языковые модели последнего поколения существенно расширили возможности этой категории, позволив обрабатывать неструктурированные текстовые источники с глубиной, ранее недостижимой [3].

Оборонительные применения охватывают весь спектр защитных функций от классификации вредоносного ПО до обнаружения вторжений и анализа поведения пользователей. Машинное обучение применяется в системах обнаружения сетевых атак (NIDS), системах обнаружения целевых атак (EDR), системах управления событиями информационной безопасности (SIEM). Принципиальным достижением последних лет стало развитие платформ автоматизированного реагирования на инциденты (SOAR), способных не только обнаруживать угрозы, но и инициировать ответные действия от изоляции скомпрометированных узлов до блокировки подозрительного трафика без участия человека. Это сдвигает традиционную границу между обнаружением и реагированием, ставя принципиально новые вопросы о допустимой степени автоматизации защитных решений [9].

Контрнаступательные применения наиболее новая и наиболее спорная категория, охватывающая использование ИИ для активных действий в отношении атакующих или их инфраструктуры. Сюда относятся: автоматизированное проникновение в системы атакующих с целью извлечения разведывательных данных; развёртывание динамически генерируемых ложных сетей (honeynets) и обманных артефактов, специально

создаваемых ИИ для введения противника в заблуждение; автоматизированное прерывание вредоносной активности на стороне источника. В отличие от первых двух категорий, контрнаступательные применения находятся в значительной мере в правовой и этической серой зоне: даже когда они технически реализуемы, их легитимность с точки зрения международного права и национальных юрисдикций остаётся предметом серьёзных дискуссий [8].

Анализ практики внедрения ИИ в кибербезопасность позволяет выделить четыре крупные категории системных рисков, специфических именно для этой технологии. Эти риски не сводятся к традиционным недостаткам информационных систем: они имеют принципиально новую природу и требуют специального теоретического осмысления.

Первый риск – состязательное машинное обучение (adversarial machine learning). Это совокупность атак, направленных не против защищаемых системой объектов, а против самой обучающейся модели, лежащей в основе защиты. Атаки этого класса включают: «отравление» обучающего набора данных (poisoning attacks), при котором атакующий вводит в данные специально подготовленные образцы, искажающие итоговую модель; «уклоняющие» атаки (evasion attacks), создающие образцы, по сути являющиеся вредоносными, но классифицируемые моделью как безопасные за счёт минимальных, незаметных для человека изменений; атаки на конфиденциальность модели (model inversion), позволяющие извлечь из обученной модели информацию об использованных для её обучения данных [4, 10]. Принципиально то, что эти атаки не используют традиционных уязвимостей программного обеспечения: они эксплуатируют статистическую природу самих алгоритмов машинного обучения, что делает их сложно предотвратимыми классическими средствами безопасности.

Второй риск проблема объяснимости (explainability). Современные высокоэффективные модели машинного обучения, особенно глубокие нейронные сети, являются «чёрными ящиками»: они выдают решения, не предоставляя интерпретируемого объяснения причин этих решений. В кибербезопасности это создаёт серьёзные операционные проблемы: аналитик, получивший от системы сигнал об угрозе, не может проверить обоснованность этого сигнала, не может объяснить решение руководству или регулятору, не может извлечь из инцидента уроки, переносимые на будущее. Это противоречит и базовому принципу управления рисками: невозможно эффективно управлять рисками, природа которых остаётся непонятной [6].

Третий риск асимметрия возможностей между сторонами конфликта. ИИ-технологии в кибербезопасности являются ресурсоёмкими: их разработка требует крупных наборов данных, значительных вычислительных мощностей и высокой квалификации специалистов. Это создаёт ситуацию, при которой передовые защитные возможности концентрируются у крупных корпораций и развитых государств, тогда как малые организации и государства с ограниченными ресурсами оказываются в положении растущего технологического отставания. Парадокс заключается в том, что атакующие возможности на основе ИИ значительно проще тиражируются: готовые наступательные инструменты могут быть переданы или украдены, тогда как защитные системы требуют постоянного дообучения и сопровождения [7].

Четвёртый риск угроза автономной эскалации. По мере того как ИИ-системы получают всё большую автономию в принятии оборонительных решений от блокировки трафика до изоляции узлов и инициирования ответных действий, возникает принципиально новая возможность: эскалация инцидента, не санкционированная человеком. Гипотетический сценарий, в котором две автономные системы кибербезопасности взаимно интерпретируют действия друг друга как враждебные и инициируют

последовательность ответных шагов, ускоряющихся быстрее, чем человек может вмешаться, перестаёт быть фантастикой и переходит в категорию обоснованных озабоченностей. Существующие исследования по аналогичным проблемам в сфере высокочастотной торговли на финансовых рынках демонстрируют, что подобные сценарии не только возможны, но и периодически реализуются в реальности [4].

Массовое внедрение ИИ в кибербезопасность имеет последствия, выходящие далеко за пределы конкретных организаций и достигающие уровня глобального стратегического равновесия. Эти последствия пока в полной мере не осмыслены ни в академической литературе, ни в государственной политике, однако их контуры уже различимы.

Принципиально меняется временной ритм киберконфликта. Если раньше инциденты разворачивались в темпе, доступном человеческому восприятию часы, дни, недели, то ИИ-системы способны на действия в миллисекундных масштабах. Это означает, что традиционные процедуры принятия решений, основанные на человеческой экспертизе и иерархической ответственности, всё чаще будут запаздывать по сравнению с темпом самих событий. Возникает структурное давление в пользу делегирования всё большего объёма решений автономным системам, которое в долгосрочной перспективе может привести к ситуации, в которой человек выходит из процесса принятия оперативных решений в кибербезопасности почти полностью [1].

Одновременно меняется и характер атрибуции. ИИ-инструменты разведки позволяют значительно глубже анализировать технические артефакты атак и связывать их с конкретными атакующими группами, что в принципе делает атрибуцию более точной. С другой стороны, те же ИИ-инструменты дают атакующим возможность создавать поддельные «отпечатки» намеренно имитировать стиль и инструменты других известных групп с целью введения аналитиков в заблуждение. Это создаёт новый уровень

информационной войны вокруг самой атрибуции и ставит под вопрос доказательную ценность технических индикаторов, которые традиционно были одним из основных аргументов в публичных обвинениях [13].

Обсуждение

Феномен ИИ в кибербезопасности наглядно демонстрирует то, что в теории технологической политики называется проблемой технологий двойного назначения: одна и та же базовая технология может использоваться как для защитных, так и для наступательных целей, и эти применения принципиально неотделимы друг от друга. Алгоритм, обучающий обнаруживать вредоносное ПО, является зеркальным отражением алгоритма, обучающего создавать вредоносное ПО, способное уклоняться от обнаружения. Большие языковые модели, помогающие аналитику в работе с разведывательными данными, одновременно являются инструментом, позволяющим атакующему создавать высококачественные фишинговые сообщения, индивидуализированные под конкретного получателя [7].

Это принципиально отличает ИИ от ряда предыдущих военно-технологических инноваций, в которых атакующие и защитные применения хотя бы технически разводились. В случае ИИ разделение является, по сути, риторическим: оно зависит от намерений пользователя, а не от свойств самой технологии. Это серьёзно осложняет любые регуляторные подходы, основанные на различении «хорошего» и «плохого» ИИ: запрет на разработку «вредоносного» ИИ при сохранении свободы разработки «полезного» оказывается практически невыполнимым в условиях, когда речь идёт об одних и тех же технических компонентах [8].

Один из принципиальных открытых вопросов каков должен быть уровень человеческого контроля над ИИ-системами в кибербезопасности? С одной стороны, скорость современных киберконфликтов делает полный человеческий контроль невозможным: аналитик не успевает рассмотреть

каждое решение системы в реальном времени. С другой стороны, полностью автономные системы кибербезопасности порождают перечисленные выше риски автономной эскалации и непредсказуемого поведения.

Компромиссное решение, обозначаемое как «человек в контуре» (human-in-the-loop) или «человек над контуром» (human-on-the-loop), предполагает, что ИИ-системы могут самостоятельно выполнять рутинные действия, но критические решения проходят через человеческое подтверждение. Однако в реальной операционной практике граница между «рутинным» и «критическим» оказывается значительно более размытой, чем предполагают академические схемы. Принципиальный вопрос состоит не только в том, какие решения должен принимать человек, но и в том, какие возможности у него остаются для осмысленной оценки этих решений в условиях информационной перегрузки и временного давления.

Теоретический характер настоящего исследования определяет основные его ограничения. Анализ опирается на открытые источники, тогда как значительная часть реальной практики применения ИИ в кибербезопасности остаётся закрытой прежде всего в государственном секторе. Это создаёт неизбежный разрыв между академической картиной и реальным состоянием дел, который частично компенсируется анализом утечек и публичных инцидентов, но не устраняется полностью.

Перспективными направлениями дальнейших исследований представляются: сравнительный анализ национальных стратегий применения ИИ в кибербезопасности с акцентом на их операционные различия; разработка методологии независимого тестирования ИИ-систем безопасности на устойчивость к состязательным атакам; изучение специфических потребностей и возможностей государств с развивающимися национальными кибератакам прежде всего стран Центральной Азии и других регионов, где

формирование собственных ИИ-возможностей в сфере кибербезопасности находится в начальной стадии.

Заключение

Внедрение искусственного интеллекта в кибербезопасность представляет собой не просто очередное технологическое усовершенствование, а фундаментальную трансформацию операционных и стратегических основ этой области. Переход от сигнатурного к поведенческому моделированию угроз, от ручного анализа к автоматизированной аналитике, от человеческого реагирования к автономным защитным действиям всё это сдвиги, затрагивающие не отдельные инструменты, но саму архитектуру кибербезопасности как практики.

Предложенная типология применений ИИ в кибербезопасности разведывательное, оборонительное и контрнаступательное измерения позволяет упорядочить разнообразный технический материал и связать его со стратегическими и правовыми измерениями проблемы. Каждая категория имеет свою специфику, свой набор практик и свой профиль рисков, и их совместное рассмотрение даёт более полную картину, чем традиционные подходы, концентрирующиеся преимущественно на оборонительной стороне.

Анализ системных рисков состязательного машинного обучения, проблемы объяснимости, асимметрии возможностей и угрозы автономной эскалации показывает, что ИИ в кибербезопасности не является технологией, не имеющей издержек. Эти издержки качественно иной природы, чем риски традиционных средств безопасности, и их адекватное теоретическое и практическое осмысление является одной из главных задач исследовательского и регуляторного сообщества на ближайшие годы.

В более широкой перспективе феномен ИИ в кибербезопасности обнажает структурное противоречие современной цифровой эпохи: технологические возможности развиваются быстрее, чем способность

общества и государств осмысленно их регулировать. Преодоление этого разрыва является не техническим, а политическим и интеллектуальным вызовом, и его решение определит во многом тот цифровой мир, в котором будут жить ближайшие поколения.

Список литературы

1. CVE Details: National Vulnerability Database Statistics. – NIST, 2024. – URL: <https://nvd.nist.gov/general/visualizations/vulnerability-visualizations> (дата обращения: 20.03.2025).
2. Mandiant. M-Trends 2023 Report. – Reston : Mandiant, 2023. – 76 p.
3. Buczak A. L., Guven E. A survey of data mining and machine learning methods for cyber security intrusion detection // IEEE Communications Surveys & Tutorials. – 2016. – Vol. 18, № 2. – P. 1153–1176.
4. Goodfellow I. J., Shlens J., Szegedy C. Explaining and harnessing adversarial examples // International Conference on Learning Representations. – San Diego, 2015. – 11 p.
5. Sommer R., Paxson V. Outside the closed world: On using machine learning for network intrusion detection // IEEE Symposium on Security and Privacy. – 2010. – P. 305–316.
6. Apruzzese G., Colajanni M., Ferretti L., Marchetti M. On the effectiveness of machine and deep learning for cyber security // Tenth International Conference on Cyber Conflict (CyCon). – Tallinn : NATO CCDCOE, 2018. – P. 371–390.
7. Brundage M., Avin S., Clark J. et al. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. – Oxford : Future of Humanity Institute, 2018. – 101 p.
8. Schmitt M. N. (ed.) Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations. – 2nd ed. – Cambridge : Cambridge University Press, 2017. – 642 p.

9. Chio C., Freeman D. Machine Learning and Security: Protecting Systems with Data and Algorithms. – Sebastopol : O'Reilly Media, 2018. – 386 p.
10. Biggio B., Roli F. Wild patterns: Ten years after the rise of adversarial machine learning // Pattern Recognition. – 2018. – Vol. 84. – P. 317–331.