DOI: 10.55648/1998-6920-2026-20-77-92 УДК 004.056

Ключевые принципы построения подсистемы информационной безопасности MLOps

Д. В. Нагибин[1], А. Н. Кокоулин[1,2], А. А. Южаков [1]

[1] Пермский национальный исследовательский политехнический университет (ПНИПУ)
[2] ФБУН «Федеральный научный центр медико-профилактических технологий управления рисками здоровью населения»

Аннотация: В статье рассматриваются ключевые принципы построения подсистемы информационной безопасности MLOps (Machine Learning Operations), учитывая растущую актуальность и уязвимость систем машинного обучения в современных условиях. В работе рассматриваются основные регуляторные требования к информационной безопасности систем, использующих машинное обучение, включая анализ требований ФСТЭК России. Обосновывается необходимость адаптации современных решений MLSecOps к условиям российского законодательства. Проведён анализ ключевых рисков кибербезопасности, характерных для каждого этапа жизненного цикла MLOps, а также представлены существующие методы и подходы их митигации. В работе приведена математическая формализация для оценки защищенности системы, учитывающей технические и регуляторные требования, а также требования к импортозамещению. Представлен концептуальный состав модулей подсистемы ИБ. Результаты исследования направлены на создание эффективной и надежной подсистемы защиты конвейеров MLOps и снижения рисков, связанных с использованием технологий машинного обучения.

Ключевые слова: MLOps, MLSecOps, информационная безопасность, безопасность машинного обучения, риски кибербезопасности в MLOps, подсистема ИБ в MLOps, регуляторные документы РФ по безопасности ИИ.

Для цитирования: Нагибин Д. В., Кокоулин А. Н, Южаков А. А. Ключевые принципы построения подсистемы информационной безопасности MLOps // Вестник СибГУТИ. 2026. Т. 20, № 2. С.77–92. https://doi.org/10.55648/1998-6920-2026-20-2-77-92.

Статья поступила в редакцию 18.03.2026;
принята к публикации 22.04.2026.

1. Введение

В современной практике разработки и внедрения решений машинного обучения наблюдается растущая потребность в оптимизации и автоматизации всего жизненного цикла моделей машинного обучения (МО, от англ. Machine Learning − ML). Данный процесс, известный как MLOps (Machine Learning Operations − операции машинного обучения), представляет собой комплексный подход к управлению жизненным циклом моделей машинного обучения, охватывающий все этапы – от подготовки данных и разработки моделей до их развёртывания и последующего мониторинга. Основополагающими принципами MLOps является обеспечение версионности как данных, так и моделей, что гарантирует воспроизводимость экспериментов и возможность отслеживания изменений, а также автоматизация процессов обучения и валидации, позволяющая эффективно выбирать оптимальные конфигурации моделей [1]. Базовая схема этапов жизненного цикла моделей представлена на рис. 1.

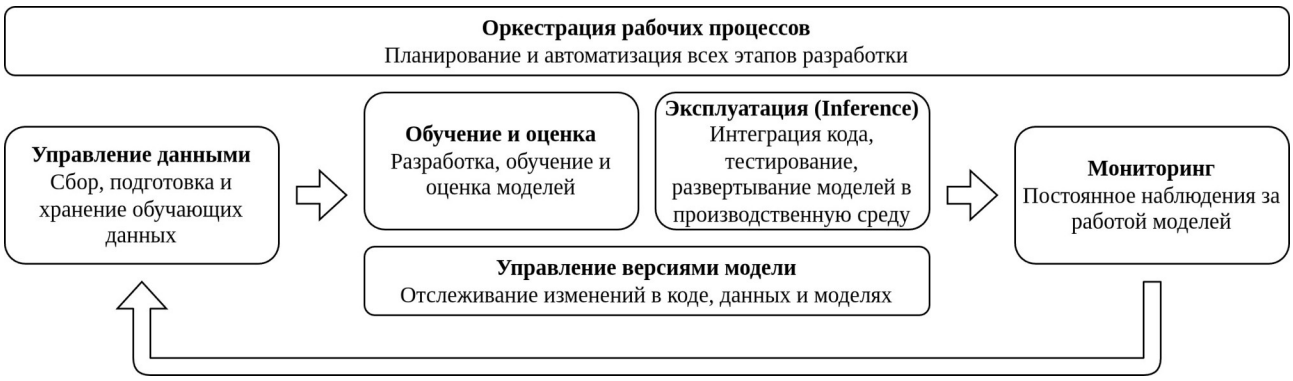

Рис. 1. Базовая схема этапов жизненного цикла MLOps

Широкое внедрение MLOps позволяет ускорить вывод ML-систем в производственную среду (production) и увеличивает их надёжность за счёт автоматизации, но при этом значительно расширяется поверхность атаки, поскольку на каждом этапе конвейера МО существуют свои уязвимости [2]. Защита MLOps-систем приобретает особую важность, поскольку компрометация данных, моделей или инфраструктуры может привести к серьезным последствиям, включая финансовые потери, репутационный ущерб и угрозу физической безопасности. Поэтому защита систем машинного обучения требует принципиально новых подходов и комплексного охвата всех стадий MLOps − так называемого MLSecOps. Интеграция практик безопасности в MLOps позволяет выявлять и смягчать риски на ранних этапах разработки, обеспечивая защиту данных, соответствие регуляторным требованиям и устойчивость к злонамеренным воздействиям [3].

Целью данного исследования является выявление ключевых архитектурных решений для создания комплексной подсистемы информационной безопасности (ИБ) систем машинного обучения (MLOps).

Для достижения поставленной цели необходимо решить следующие задачи:

1. Рассмотреть основные регуляторные требования РФ к информационной безопасности систем, использующих машинное обучение.
2. Идентифицировать ключевые риски кибербезопасности, характерные для каждого этапа жизненного цикла MLOps, и проанализировать существующие методы и подходы их митигации.
3. Выделить ключевые принципы при проектировании подсистемы ИБ в MLOps с учётом имеющихся решений.

2. Нормативные требования

2.1. Регуляторные документы РФ по безопасности ИИ

В настоящее время в Российской Федерации активно формируется комплексная нормативно-правовая база, регулирующая вопросы информационной безопасности в сфере искусственного интеллекта (ИИ). Данный процесс характеризуется переходом от общих принципов к конкретным техническим требованиям и стандартам безопасности, обязательным для исполнения организациями государственного сектора и операторами критической информационной инфраструктуры (КИИ). Основополагающим документом в данной области выступает Национальная стратегия развития искусственного интеллекта на период до 2030 года, в которой подчёркивается важность защиты прав граждан и обеспечение технологического суверенитета в сфере высоких технологий [4].

Ключевую роль в регулировании рисков ИИ играет Федеральная служба по техническому и экспортному контролю (ФСТЭК России). С декабря 2025 года риски, связанные с машинным обучением, включены в Банк данных угроз, что обязывает организации учитывать специфические уязвимости нейронных сетей, такие как отравление данных, внедрение

вредоносных инструкций и кража весов, при разработке систем защиты информационных систем (ИС). ФСТЭК России также инициировала переход к непрерывному мониторингу кибербезопасности государственных органов, что подчёркивает важность интеграции средств защиты в рабочие процессы MLOps [5].

В рамках данного процесса важным этапом стало принятие Приказа ФСТЭК России №117 от 11 апреля 2025 года (вступил в силу с 1 марта 2026 года), который вводит циклическую модель управления защитой информации. Данный приказ требует формирования полноценной политики информационной безопасности, охватывающей технологии контейнеризации, оркестрации и программные интерфейсы взаимодействия приложений [6]. Особое внимание уделяется защите данных, моделей, параметров, процессов и сервисов обработки данных, включая запрет передачи информации ограниченного доступа разработчикам для улучшения функционирования модели. Ключевые мероприятия включают в себя защиту информации при использовании ИИ, контроль взаимодействия «запрос-ответ» с учётом различных типов запросов, разработку статистических критериев для выявления недостоверных ответов, предотвращение несанкционированного изменения параметров модели и функционирования информационных систем. Реализация приказа предполагает детальное описание границ AI-контура, перечня защищаемых активов, а также разработку архитектурных схем и политик безопасности [7].

Методика оценки защищенности ФСТЭК (от 25.11.2025) дополняет Приказ №117, формализуя процесс проверки и задавая состав работ при анализе уязвимостей в рамках испытаний и аттестации. Методика описывает анализ уязвимостей как последовательность этапов:

1) сбор исходной информации: состав ПО, места запуска моделей (inference), наличие RAG (Retrieval-Augmented Generation) и каналы формирования запросов и ответов;
2) внешний и внутренний анализ уязвимостей;
3) оценка выявленных уязвимостей.

Внешний анализ предусматривает моделирование атак внешнего нарушителя (AI Red Team) в режиме «черного ящика», а внутренний анализ – моделирование атак со стороны внутреннего нарушителя, включая подмену конфигураций, отравление данных обучения и модификацию весов моделей. Результатом анализа является формирование отчётов с конкретными требованиями по устранению выявленных уязвимостей [8].

В дополнение к техническому регулированию, развивается этическое регулирование сферы ИИ, направленное на ответственное создание систем и минимизацию гуманитарных рисков и предвзятости. Однако для объектов КИИ этические нормы дополняются строгими требованиями по импортозамещению и использованию только сертифицированных библиотек и моделей [5].

2.2. Возможности адаптации решений к российским нормативным требованиям

Внедрение и адаптация современных решений машинного обучения (ML) и связанных с ними практик безопасности машинного обучения (MLSecOps) в условиях российской нормативно-правовой базы требуют учета ряда специфических требований, предъявляемых ФСТЭК России. Особую значимость приобретают аспекты, связанные с сертификацией, импортозамещением и соблюдением требований защиты персональных данных. Ключевым фактором является обеспечение локализации всех процессов обработки данных и эксплуатации моделей в пределах защищённого контура организации или в доверенных облачных средах [7]. Этот принцип подкрепляется положениями Приказа ФСТЭК № 117, который ограничивает передачу конфиденциальной информации внешним разработчикам с целью оптимизации моделей, что, в свою очередь, затрудняет интеграцию и использование многих зарубежных решений, основанных на модели SaaS (Software As a Service) [6].

Для соответствия требованиям российского законодательства, адаптация MLSecOps-решений должна осуществляться по нескольким ключевым направлениям. Прежде всего, требуется обеспечение совместимости программного обеспечения с отечественными опера-

ционными системами, такими как Astra Linux, РЕД ОС, ALT Linux, Rosa Linux и др, что является необходимым условием для внесения продуктов в Единый реестр российского ПО. Кроме того, необходима поддержка сертифицированных криптографических алгоритмов (ГОСТ) при передаче данных между компонентами MLOps-конвейера, что обеспечивает защиту информации на всех этапах жизненного цикла модели. Наконец, методики тестирования робастности моделей должны быть согласованы с требованиями ГОСТ Р 70462.1-2022/ ISO/IEC TR 24029-1-2021, регламентирующего применение статистических и формальных методов оценки устойчивости нейронных сетей к внешним воздействиям [9].

В соответствии с Методикой анализа защищенности ФСТЭК от ноября 2025 года, организациям необходимо проводить регулярный контроль уровня безопасности конфиденциальной информации не реже одного раза в три года, либо после внесения существенных изменений в системы обработки данных [8]. В контексте MLOps это означает необходимость разработки и внедрения автоматизированных процедур повторного тестирования моделей после каждого цикла переобучения или обновления весов, что является важным элементом поддержания требуемого уровня информационной безопасности.

3. Анализ рисков кибербезопасности в MLOps и возможные способы их митигации

3.1. Анализ рисков

Уязвимости присутствуют на всех этапах жизненного цикла MLOps. Последствия успешных атак на MLOps-системы могут быть весьма серьёзными, включая финансовые и репутационные потери, нарушение конфиденциальности данных и потерю интеллектуальной собственности. Например, кража обученной модели может привести к потере конкурентного преимущества, а утечка данных – к юридическим последствиям и, например, штрафам от регулятора [10].

В РФ риски усугубляются использованием открытого программного обеспечения и предобученных моделей из зарубежных репозиториев, что создаёт угрозы в цепочке поставок. Основные риски могут быть разделены на три категории: угрозы данным, угрозы самой модели и инфраструктурные угрозы. Рассмотрим конкретные примеры атак на каждом этапе жизненного цикла MLOps.

На этапе подготовки данных наиболее критичным риском является отравление данных (Data Poisoning). Злоумышленник может преднамеренно внести вредоносные образцы в набор данных, что приведёт к искажению поведения модели в будущем. Одним из механизмов реализации такой атаки является внедрение скрытых триггеров, при которых модель учится ассоциировать определённый паттерн (шаблон) с неверным результатом. При отсутствии триггера модель сохраняет высокую точность, что делает атаку крайне сложной для обнаружения [10]. Также актуальна манипуляция метками классов (Label Flipping), способная вызвать предсказуемые сбои в антифрод-системах или медицинских диагностических комплексах [11]. Отдельную проблему представляет риск восстановления персональных данных из обучающей выборки через атаки инверсии модели или утечки из хранилищ признаков, что прямо нарушает требования ФЗ-152 [12].

Этап обучения и разработки модели сопряжён с угрозами безопасности инфраструктуры и возможностью исполнения произвольного кода через артефакты моделей. Использование форматов сериализации, таких как pickle, позволяет внедрять вредоносный код непосредственно в файлы весов (атаки типа Model Tampering) [13]. Загрузка подобной модели из публичного репозитория (например, Hugging Face) может привести к выполнению вредоносного сценария на сервере обучения и, как следствие, к компрометации инфраструктуры организации. Кроме того, уязвимости в MLOps-фреймворках, таких как Ray или Kubeflow, могут быть использованы для перехвата контроля над вычислительными ресурсами и изменения параметров модели в процессе обучения. Специфическим риском является компрометация LoRA-адаптеров, используемых для дообучения больших моделей, поскольку они представ-

ляют собой скрытый канал внедрения вредоносной логики, недоступный для традиционных средств статического анализа кода [13]. Атаки с определением членства (Membership Inference) позволяют злоумышленникам определить, использовалась ли конкретная запись данных при обучении модели [10]. Атаки на целостность модели, такие как атаки на извлечение модели (Model Inversion Attacks), направлены на раскрытие информации о структуре и параметрах модели [12].

Этап развёртывания и эксплуатации модели в продуктивной среде (Inference) делает её мишенью для атак, направленных на обход логики принятия решений. Состязательные атаки (Adversarial Examples) заключаются в подаче на вход модели данных с минимальными, незаметными для человеческого глаза изменениями, которые заставляют нейронную сеть совершать грубые ошибки классификации [13]. Для современных генеративных систем на базе больших языковых моделей критическую опасность представляет внедрение вредоносных инструкций (Prompt Injection − промпт-инъекции**)**. Злоумышленник может внедрить скрытые команды в запрос, заставляя модель игнорировать системные ограничения, раскрывать инструкции разработчика или выполнять несанкционированные действия в интегрированных сервисах [14]. Примером является атака EchoLeak, позволяющая извлекать конфиденциальные данные из корпоративных ассистентов без прямого взаимодействия с пользователем [15].

Этап мониторинга и переобучения также не застрахован от атак. Атаки на системы мониторинга могут позволить злоумышленникам скрыть вредоносную активность, а манипуляции с метриками могут исказить представление о производительности модели. Атаки на переобучаемые модели могут привести к тому, что модели будет адаптироваться к вредоносным данным, что приведёт к ухудшению качества.

3.2. Возможные способы митигации

Понимание рисков кибербезопасности в MLOps и разработка эффективных стратегий митигации являются критически важными задачами для обеспечения надёжности и безопасности систем машинного обучения. В данной работе сделан фокус на некоторых важных аспектах в рамках разработки подсистемы ИБ для систем машинного обучения, поэтому выбор методов митигации определяется спецификой этой задачи и необходимостью интеграции с существующей инфраструктурой MLOps.

Прежде чем перейти к конкретным аспектам, необходимо провести обзор существующих подходов к защите систем машинного обучения, которые можно разделить на несколько категорий: защита данных, защита моделей, защита инфраструктуры и защита конвейеров MLOps.

Защита данных является фундаментальным аспектом безопасности. Для обеспечения конфиденциальности и целостности данных целесообразно применение шифрования данных как в состоянии покоя, так и при передаче, с акцентом на надёжное управление ключами и их регулярную ротацию. Строгий контроль доступа, основанный на принципе наименьших привилегий и ролевой модели доступа, ограничивает доступ пользователей и процессов только теми правами, которые необходимы для выполнения их конкретных задач. Регулярный мониторинг данных для выявления аномалий и потенциальных угроз, а также использование инструментов профилирования данных, способствуют своевременному обнаружению отклонений.

Для повышения устойчивости моделей к атакам на этапе эксплуатации, перспективным подходом является состязательное обучение (Adversarial Training), при котором модель намеренно обучается на данных, содержащих возмущения [3]. Также может оказаться полезным использование робастных функций потерь (Robust Loss Functions), которые менее чувствительны к выбросам и шуму в данных, часто являющимися признаками попыток отравления датасета [13].

Для защиты интеллектуальной собственности и предотвращения кражи весов моделей целесообразно применение методов маркировки (Model Watermarking). Внедрение скрытых

«водяных знаков» в веса или ответы модели позволяет юридически доказать факт незаконного копирования при обнаружении клона системы во внешней среде [16]. Другим эффективным методом является метод Staining and Locking, предполагающий встраивание маркеров для проверки подлинности и выявления фактов подмены модели в конвейере [17].

Отслеживание «родословной» модели (Model Lineage), включая данные, использованные для обучения, параметры обучения и версии зависимостей, позволяет отследить происхождение модели, обеспечить воспроизводимость и облегчить аудит и управление данными [13]. Создание «карточек модели» (Model Cards), содержащих документацию о характеристиках модели, её ограничениях, предполагаемых областях применения и потенциальных рисках, обеспечивает прозрачность и способствует ответственному использованию модели [3].

Безопасность инфраструктуры, на которой запущена модель МО, также имеет критическое значение. Использование безопасных образов контейнеров, сканирование на наличие уязвимостей и регулярное обновление ПО являются необходимыми мерами. Мониторинг активности в инфраструктуре, выявление аномалий и потенциальных угроз с использованием систем обнаружения вторжений (IDS) и систем предотвращения вторжений (IPS), позволяет своевременно реагировать на инциденты. Регулярное управление уязвимостями и своевременное их устранение, а также сегментация сети для ограничения распространения угроз, повышают общую устойчивость системы.

Важным аспектом является переход к концепции «безопасность как код» (Security-as-Code), при которой политики и тесты робастности описываются в виде программного кода и автоматически применяются, что способствует автоматизации и повышению эффективности защиты. Интеграция практик безопасности в конвейеры (pipelines) MLOps на всех этапах жизненного цикла модели является ключевым фактором обеспечения безопасности. Расширение практик DevSecOps на MLOps (MLSecOps), включая автоматизированное тестирование безопасности в конвейерах CI/CD и сканирование кода, позволяет выявлять уязвимости на ранних этапах разработки. Также данная интеграция позволяет формировать спецификацию (состав) программных компонентов (Software Bill of Materials − SBOM) для каждого этапа MLOps, что критично для выявления уязвимостей в сторонних библиотеках и обеспечения соответствия требованиям ФСТЭК России по контролю состава программного обеспечения [3].

Помимо прочего в условиях импортозамещения особое внимание должно уделяться использованию сертифицированных образов контейнеров и регулярному сканированию инфраструктуры на наличие уязвимостей с применением отечественного ПО.

4. Существующие решения в сфере MLSecOps

В настоящее время мировой рынок решений для обеспечения безопасности машинного обучения (MLSecOps) демонстрирует значительный рост и характеризуется широким спектром инструментов, предназначенных для защиты различных этапов жизненного цикла моделей ИИ [18]. Среди наиболее заметных решений, предлагающих специализированные решения для выявления и устранения уязвимостей, можно выделить Giskard и ModelScan. Giskard предоставляет платформу для динамического стресс-тестирования LLM-агентов, используя более 50 зондов для обнаружения галлюцинаций, инъекций и утечек данных, а также предлагая конкретные рекомендации по устранению выявленных проблем [19]. В свою очередь, ModelScan от Protect AI фокусируется на защите от атак, связанных с сериализацией, осуществляя проверку файлов моделей в форматах H5, Pickle и SavedModel на наличие вредоносного кода [20].

Российский рынок MLSecOps также активно развивается, предлагая решения, адаптированные к национальным стандартам и требованиям регуляторов. В частности, платформа HiveTrace выделяется как одна из ведущих для мониторинга безопасности генеративного ИИ, обеспечивая непрерывный контроль входящих и исходящих сообщений с целью выявле-

ния вредоносных инструкций (Prompt Injection), попыток взлома и попыток обхода системных ограничений. Интеграция с ролевой моделью доступа и совместимость с системами мониторинга Prometheus и Loki повышают эффективность и гибкость платформы [21]. В таблице 1 представлено краткое сравнение некоторых российских и зарубежных MLSecOps-систем.

Таблица 1. Краткое сравнение имеющихся решений в сфере защиты MLOps

Название	Страна	Ключевая функциональность и специфика	Стадия MLOps
HiveTrace	РФ	Мониторинг безопасности LLM, очистка PII, детекция инъекций	Эксплуатация (Inference)
LLAMATOR	РФ	Open Source фреймворк для тестирования уязвимостей по OWASP	Валидация и тестирование
Garda Analytics	РФ	Анализ аномалий в данных, защита от утечек, мониторинг сети	Данные и инфраструктура
PT Dephaze	РФ	Автоматизированный внутренний пентест ML-систем	Разработка и тестирование
Giskard	ЕС	Динамический Red Teaming для LLM, автоматизация тестов	Валидация и тестирование
Protect AI	США	Сканирование артефактов моделей, защита цепочки поставок	Обучение и поставки
HiddenLayer	США	Runtime-мониторинг, защита от экстракции весов и кражи моделей	Эксплуатация (Inference)
NeMo Guardrails	США	Наборы правил и фильтров для контроля поведения LLM	Эксплуатация (Inference)

Фреймворк LLAMATOR представляет собой отечественный инструмент с открытым исходным кодом для автоматического тестирования безопасности чат-ботов и LLM-пайплайнов. Он позволяет воспроизводить критические сценарии атак, оценивать ответы моделей и фиксировать метрики устойчивости, ориентируясь на выявление угроз, соответствующих списку OWASP Top 10 for LLM, включая утечку системных промптов и неограниченное потребление ресурсов. Благодаря поддержке библиотек LangChain и совместимости с API типа OpenAI, LLAMATOR легко интегрируется в существующие процессы разработки [22].

Наряду с инструментами для тестирования и мониторинга, существуют комплексные решения, направленные на защиту данных и выявление аномалий. «Гарда Аналитика» от группы компаний «Гарда» объединяет инструменты DLP с аналитическими возможностями для обнаружения аномалий в поведении моделей и пользователей, что особенно важно для выявления атак типа «отравление» на этапе эксплуатации [23]. Компания Positive Technologies, в свою очередь, предлагает специализированные инструменты для построения процессов безопасной разработки моделей, включая систему PT Dephaze для автоматизированного пентеста внутренней инфраструктуры [24].

5. Формализация требований к разработке подсистемы информационной безопасности для систем машинного обучения

5.1. Ключевые особенности подсистемы

Подсистема информационной безопасности для систем машинного обучения должна представлять собой комплексный иерархический слой управления, бесшовно интегрируемый в MLOps-конвейер на всех этапах его функционирования. Основной целью проектируемой подсистемы является минимизация специфических ML-рисков при сохранении высокой производительности и доступности сервисов ИИ.

Реализация подсистемы может быть осуществлена на основе архитектуры, аналогичной Scientific Data Services Framework [25], где создаётся надстройка над основной системой для контроля за её наиболее важными аспектами информационной безопасности. Также перспективным является использование специализированного фреймворка для MLOps, предоставляющего готовые инструменты для обнаружения сетевых вторжений, отслеживания происхождения моделей и обеспечения воспроизводимости результатов, что может значительно упростить процесс внедрения и интеграции [26]. Архитектура должна строится на модульном принципе, где каждый компонент отвечает за конкретный домен безопасности.

Предполагаемый состав подсистемы ИБ для MLOps включает шесть специализированных модулей, представленных в таблице 2. Каждый модуль направлен на защиту своего этапа MLOps.

Таблица 2. Концептуальная схема модулей подсистемы ИБ

Название	Функции и задачи	Подходы
Модуль защиты цепочки поставок	Сканирование сторонних моделей и библиотек на уязвимости и вредоносный код	Анализ сторонних библиотек и зависимостей, верификация цифровых подписей артефактов моделей, сканирование моделей на наличие вредоносного кода в сериализованных форматах (например, pickle)
Модуль защиты данных	Проверка целостности обучающих данных, обнаружение признаков отравления, анонимизация персональных данных в соответствии с ФЗ-152	Дифференциальная приватность, хеширование данных, статический анализ распределений, обнаружение признаков отравления и аномалий
Модуль защиты моделей	Защита моделей на этапах обучения и эксплуатации. Предотвращение кражи и инверсии, контроль робастности	Ведение «родословной» модели (Model Lineage), детекция адверсариальных примеров (Adversarial Examples), фильтрация входных данных, защита от атак типа «отказ в обслуживании», анализ объяснимости (XAI)
Модуль хранения и аудита	Хранение доверенных версий моделей, ведение журнала всех изменений	Ролевая модель доступа, журналы событий в неизменяемых хранилищах, управление жизненным циклом модели
Модуль активной защиты	Автоматизированная симуляция атак (промпт-инъекции, состязательные примеры) перед развёртыванием модели, фильтрация трафика к модели в реальном времени, блокировка атак на API	Сценарии автоматизированного тестирования робастности, системы лимитирования запросов (NeMo Guardrails, LLM Guard и т.п.)
Модуль мониторинга и реагирования	Детектирование дрейфа данных и модели, обнаружение аномалий в поведении системы, автоматический откат изменений при обнаружении угроз	Системы мониторинга Prometheus/Grafana, интеграция с SIEM-системами (Security Information and Event Management), автоматическое переобучение при деградации

Центральным элементом подсистемы ИБ является модуль защиты цепочки поставок, который фокусируется на обеспечении безопасности моделей и библиотек, используемых в процессе обучения и развёртывания. Уязвимости сторонних компонентов могут быть использованы для компрометации всей системы, поэтому данный модуль критически важен. Он включает в себя сканирование сторонних моделей и библиотек на наличие уязвимостей и вредоносного кода, анализ SBOM (Software Bill of Materials) для выявления потенциальных уязвимостей и зависимостей, верификацию цифровых подписей артефактов моделей и сканирование моделей на наличие вредоносного кода в сериализованных форматах, таких как pickle [13].

Модуль защиты данных направлен на обеспечение целостности и конфиденциальности обучающих данных. Его основная задача – предотвращение отравления данных, которое может привести к предвзятости или непредсказуемому поведению модели [27]. Для достижения этой цели модуль использует дифференциальную приватность, хеширование данных,

статический анализ распределений и методы обнаружения признаков отравления, включая методы обнаружения аномалий, что позволит выявлять подозрительные данные, способные негативно повлиять на процесс обучения.

Модуль защиты моделей направлен на защиту моделей от атак на этапах обучения и эксплуатации (инференса), предотвращая кражу и инверсию моделей, а также обеспечивая их робастность. Для этого модуль использует ведение «родословной» модели (Model Lineage) для отслеживания всех изменений [3], детекцию адверсариальных примеров [13], фильтрацию входных данных, защиту от атак типа «отказ в обслуживании» и анализ объяснимости (XAI) для понимания принципов принятия решений моделью [28].

Перед развёртыванием модели в продуктивной среде критически важным этапом является прохождение через модуль активной защиты. Данный модуль обеспечивает автоматизированную защиту от атак в реальном времени, используя многоуровневый подход, включающий симуляцию потенциальных угроз, фильтрацию входящего трафика и блокировку атак на API. Модель помещается в изолированную «песочницу», где подвергается тщательному тестированию, включающему как состязательные примеры, так и сложные попытки инъекций. Для обеспечения надёжности процесса используются сценарии автоматизированного тестирования робастности и системы лимитирования запросов, такие как NeMo Guardrails и LLM Guard и т. п. [29]. Только после подтверждения соответствия заданным метрикам робастности модель получает статус, одобренный для эксплуатации.

Модуль хранения и аудита предназначен для безопасного хранения и отслеживания всех версий моделей и изменений, внесённых в систему, что критически важно для поддержания надёжности и воспроизводимости результатов. В частности, модуль обеспечивает хранение доверенных версий моделей, ведёт подробный журнал всех изменений в неизменяемых хранилищах, гарантируя целостность записей о событиях. Для обеспечения контролируемого доступа к данным и функциям используется ролевая модель доступа. Для обеспечения контроля целостности при передаче моделей из среды разработки в продуктивную среду, модуль использует цифровые подписи для всех артефактов, хранящихся в доверенном реестре.

Модуль мониторинга и реагирования обеспечивает непрерывный мониторинг состояния системы и автоматическое реагирование на обнаруженные угрозы. Он включает в себя детекцию дрейфа данных и модели, обнаружение аномалий в поведении системы и автоматический откат изменений при обнаружении угрозы. Для этого используются инструменты мониторинга, такие как Prometheus/Grafana, интеграция с SIEM-системами (Security Information and Event Management) [3]. Также модуль способен сигнализировать о необходимости автоматического переобучение модели при деградации.

Связь этапов MLOps с модулями проектируемой подсистемы информационной безопасности показана на рис. 2.

Для обеспечения эффективной работы в производственной среде, подсистема ИБ должна быть спроектирована с учётом минимального воздействия на основной код моделей машинного обучения. Обработка запросов не должна приводить к значительной деградации производительности. Также необходимо обеспечить отказоустойчивость, исключающую полную остановку системы при сбое в модуле безопасности, и совместимость с отечественным программным обеспечением и средствами криптографии.

Реализация такой подсистемы позволит обеспечить комплексную защиту конвейеров MLOps и снизить риски, связанные с кибербезопасностью. При этом, необходимо учитывать, что, как отмечается в работах по AI Red Team [30], проверка моделей на тренировочном наборе данных не гарантирует защиту от сдвигов данных при работе на генеральной совокупности, что подчёркивает важность непрерывного мониторинга и адаптации системы безопасности.

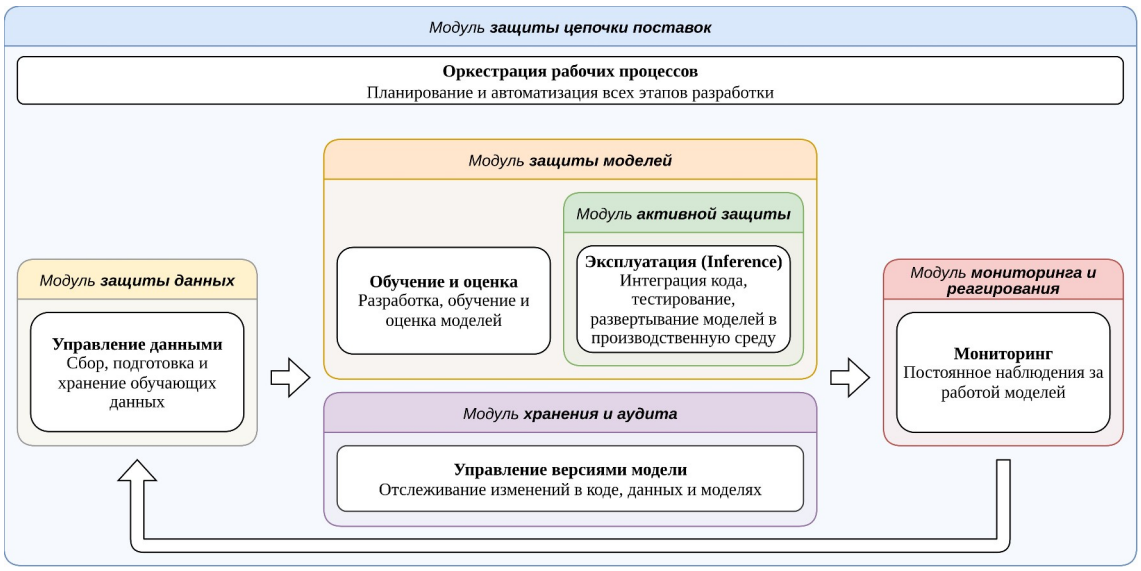

Рис. 2 Связь модулей подсистемы ИБ и основных этапов жизненного цикла MLOps

5.2. Математическая формализация

Внедрение комплексной защиты в MLOps-конвейер требует разработки формализованного подхода к оценке защищенности системы на этапе проектирования. В данном контексте, подсистема информационной безопасности рассматривается не как отдельный компонент, добавляемый к существующей системе, а как неотъемлемая часть архитектуры жизненного цикла модели машинного обучения. Такой подход позволяет интегрировать принципы безопасности на всех этапах разработки и эксплуатации.

Жизненный цикл системы машинного обучения может быть представлен в виде ориентированного графа состояний $G=(S,T)$, где множество вершин $S=\{s_1, s_2, \ldots, s_n\}$ соответствуют ключевым артефактам, таким как обучающие данные, веса нейронной сети и исполняемые контейнеры, а рёбра T – процессам их трансформации. Данный подход позволяет визуализировать и анализировать поток данных и операций, выявляя потенциальные точки уязвимости. С учётом воздействия внешних и внутренних угроз, динамика изменения состояния безопасности на этапе i описывается как (1):

$$S_{i+1} = f_i\left(S_i, \theta_i\right) + \delta_i, \tag{1}$$

где:

S_i – состояние артефакта (набор данных, веса модели, конфигурация);

f_i – целевая функция преобразования данных или модели;

θ_i – параметры доверенной среды исполнения;

δ_i – вектор дестабилизирующего воздействия (атак), способный изменить свойства артефакта (например, внесение уязвимостей в веса модели).

Для количественного обоснования выбора мер защиты может использоваться функция риска, учитывающая технические и регуляторные аспекты (2):

$$R_{total} = \sum \omega_t \cdot \left[Adv\left(M_t\right) + \lambda \cdot NonComp\left(C_t\right) \right], \tag{2}$$

где:

$Adv(Mt)$ (Adversarial Robustness) − техническая составляющая риска, определяемая как математическое ожидание потери точности модели Δ_{Acc} при воздействии состязательных примеров;

$NonComp(C_t)$ (Regulatory Non-Compliance) − функция штрафа за несоответствие требованиям безопасности;

λ − коэффициент регуляторной значимости, отражающий приоритет юридической легитимности системы в контуре КИИ.

Учитывая текущую геополитическую ситуацию, ключевым фактором при проектировании систем ИБ в Российской Федерации является обеспечение импортонезависимости. Для формализации этого требования предлагается использовать индекс технологического суверенитета (ИТС). Индекс ИТС позволяет количественно оценить степень зависимости от зарубежных технологий и служит критерием при выборе архитектурных решений I_{ts} (3):

$$I_{ts} = \frac{\sum_{j=1}^{k} \mu_j \cdot E_j}{\sum_{j=1}^{k} \mu_j}, \quad (3)$$

где:

$E_j \in \{0,1\}$ – индикатор использования доверенного отечественного компонента (например, СУБД из реестра Минцифры или сертифицированной ОС);

μ_j – весовой коэффициент значимости компонента для обеспечения целостности конвейера. Введение данного индекса позволяет ограничить пространство проектных решений только теми вариантами, которые соответствуют требованиям по замещению иностранного ПО.

Проектирование подсистемы ИБ предполагает поиск оптимального компромисса между уровнем защищенности, сохранением целевой точности модели машинного обучения и соблюдением требований законодательства (4):

$$\begin{cases} min\, R_{total}(\delta, \lambda) \\ max\, Acc(M) \\ \Delta Lat \le T_{max} \\ I_{ts} \ge I_{min} \end{cases}, \quad (4)$$

где:

$R_{total}(\delta, \lambda)$ – минимизация совокупного риска;

$Acc(M)$ – сохранение целевой точности модели;

$\Delta Lat \le T_{max}$ – ограничение задержки при эксплуатации модели (при инференсе);

$I_{ts} \ge I_{min}$ – соответствие регуляторным нормам.

Предложенный математический аппарат позволяет обосновать выбор архитектурных решений на ранних этапах проектирования, обеспечивая соответствие требованиям безопасности и нормативным актам. Например, если значение индекса ИТС (I_{ts}) не соответствует установленным требованиям, система не может быть допущена к эксплуатации в контуре КИИ независимо от её производительности. Такой подход гарантирует соответствие проектируемой подсистемы ИБ актуальным нормам российского законодательства в области MLSecOps.

Заключение

Разработка и внедрение подсистемы информационной безопасности для систем машинного обучения (MLOps) представляет собой сложную, но крайне важную задачу в современных условиях цифровой трансформации. Анализ рисков кибербезопасности в жизненном цикле MLOps показывает, что системы машинного обучения создают качественно новые вызовы, требующие внедрения специализированных подсистем защиты, выходящих за рамки традиционных средств обеспечения информационной безопасности.

Предложенный состав модулей подсистемы ИБ позволяет реализовать комплексную защиту на всех этапах жизненного цикла модели, соответствуя как техническим требованиям, так и бизнес-задачам по сохранности интеллектуальной собственности. В работе были рассмотрены ключевые принципы построения такой подсистемы. Обоснованность выбора

предлагаемых решений подтверждается изучением международного опыта и анализом существующих решений на рынке MLSecOps.

В заключение следует отметить, что обеспечение безопасности систем машинного обучения является сложной и динамично развивающейся областью. Комплексный подход, включающий применение специализированных методов защиты на всех этапах жизненного цикла модели, а также постоянный мониторинг и адаптацию к новым угрозам, представляется ключевым фактором успешного внедрения и безопасного использования моделей машинного обучения в различных сферах деятельности. Дальнейшие работы предполагаются над технической реализации подсистемы информационной безопасности. Разработка и внедрение предложенной подсистемы информационной безопасности для MLOps внесет существенный вклад в обеспечение надёжности и безопасности ML-систем, позволит снизить риски, связанные с кибербезопасностью, и создать условия для широкого и безопасного внедрения технологий искусственного интеллекта в различных сферах деятельности.

Литература

1. *Kreuzberger D.*, *Kühl N.*, *Hirschl S.* Machine Learning Operations (MLOps): Overview, Definition, and Architecture // IEEE Access. 2023. Vol. 11. P. 31866–31879.
2. *Patel R. et al*. Towards Secure MLOps: Surveying Attacks, Mitigation Strategies, and Research Challenges // arXiv:2506.02032 [cs]. 2026.
3. *Wilson S.* The developer’s playbook for large language model security: building secure AI applications. First edition. Beijing; Boston; Farnham; Sebastopol; Tokyo: O’Reilly, 2024. 200 p.
4. Указ Президента Российской Федерации от 10.10.2019 г. № 490. О развитии искусственного интеллекта в Российской Федерации. (В редакции Указа Президента Российской Федерации от 15.02.2024 № 124).
5. Регуляторные документы РФ по безопасности ИИ – с чем мы вступаем в 2026 год [Электронный ресурс] // Хабр. URL: `https://habr.com/ru/articles/986800/` (дата обращения: 07.02.2026).
6. Федеральная служба по техническому и экспортному контролю. Требования о защите информации, содержащейся в государственных информационных системах, иных информационных системах государственных органов, государственных унитарных предприятий, государственных учреждений: утверждены приказом ФСТЭК России от 11 апреля 2025 г. № 117.
7. Анализ ключевых изменений в требованиях к защите информации согласно Приказу ФСТЭК № 117 [Электронный ресурс]. URL: `https://sec.ussc.ru/fstec_117` (дата обращения: 08.02.2026).
8. Федеральная служба по техническому и экспортному контролю. Методический документ. Методика анализа защищенности информационных систем. Утверждена 25 ноября 2025 г.
9. ГОСТ Р 70462.1-2022/ISO/IEC TR 24029-1-2021. Информационные технологии. Интеллект искусственный. Оценка робастности нейронных сетей.
10. *Намиот Д. Е.* Схемы атак на модели машинного обучения // International Journal of Open Information Technologies. 2023. Т. 11, № 5. С. 68–86.
11. *Lavaur L.*, *Busnel Y.*, *Autrel F.* Investigating the Impact of Label-flipping Attacks against Federated Learning for Collaborative Intrusion Detection // Computers & Security. 2025. Vol. 156. P. 104462.
12. *Hall P.* Machine learning for high-risk applications: approaches to responsible AI. First Edition. Québec: O’Reilly Media, Incorporated, 2023. 466 p.
13. *Sotiropoulos J.* Adversarial AI Attacks, mitigations, and defense strategies: a cybersecurity professional’s guide to AI attacks, threat modeling, and securing AI with MLSecOps. Packt Publishing, 2024. 602 p.

14. *Намиот Д. Е.*, *Ильюшин Е. А.*, *Чижов И. В.* Атаки на системы машинного обучения – общие проблемы и методы // International Journal of Open Information Technologies. 2022. Т. 10, № 3. С. 17–22.
15. *Kovacs E.* «EchoLeak» AI Attack Enabled Theft of Sensitive Data via Microsoft 365 Copilot [Электронный ресурс]. URL: https://www.securityweek.com/echoleak-ai-attack-enabled-theft-of-sensitive-data-via-microsoft-365-copilot/ (дата обращения: 08.02.2026).
16. *Gao Y. et al*. Backdoor Attacks and Countermeasures on Deep Learning: A Comprehensive Review // arXiv:2007.10760 [cs]. 2020.
17. *Sutton O. J. et al*. Staining and locking computer vision models without retraining // arXiv:2507.22000 [cs]. 2025.
18. MLOps Market Outlook from 2026 to 2033: Trends by Application, by Region, and 12.8% CAGR Forecast [Электронный ресурс]. URL: https://www.linkedin.com/pulse/mlops-market-outlook-from-2026-2033-trends-application-region-dswuf (дата обращения: 02.03.2026).
19. Giskard – LLM Agent Testing & Evaluation Platform [Электронный ресурс]. URL: https://docs.giskard.ai/index.html (дата обращения: 08.02.2026).
20. Center S. I. S. ModelScan – Protection Against Model Serialization Attacks [Электронный ресурс]. URL: https://isc.sans.edu/diary/31692 (дата обращения: 08.02.2026).
21. LLM Monitoring для GenAI-приложений | AI Security Lab ИТМО [Электронный ресурс]. URL: https://hivetrace.ru/ (дата обращения: 08.02.2026).
22. LLAMATOR [Электронный ресурс]. URL: https://llamator-core.github.io/llamator/ (дата обращения: 08.02.2026).
23. Гарда DLP – система предотвращения утечек информации | Продажа и внедрение от Российского разработчика «Гарда» [Электронный ресурс]. URL: https://garda.ai/products/data-protection/dlp (дата обращения: 08.02.2026).
24. PT Dephaze – автопентест, который покажет, что сможет сделать хакер в инфраструктуре [Электронный ресурс]. URL: https://ptsecurity.com/products/dephaze/ (дата обращения: 08.02.2026).
25. *Kim J. et al.* Security for the scientific data services framework // 2015 IEEE International Conference on Big Data (Big Data). 2015. P. 1871–1875.
26. *Spadari V. et al.* An MLOps Framework for Explainable Network Intrusion Detection with MLflow // 2024 IEEE Symposium on Computers and Communications (ISCC). Paris, France: IEEE, 2024. P. 1–6.
27. *Tete S. B.* Threat Modelling and Risk Analysis for Large Language Model (LLM)-Powered Applications // arXiv:2406.11007 [cs]. 2024.
28. *Hassija V. et al.* Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence // Cognitive Computation. 2024. Vol. 16, no. 1. P. 45–74.
29. Awesome MLSecOps [Электронный ресурс]. URL: https://github.com/RiccardoBiosas/awesome-MLSecOps (дата обращения: 04.02.2026).
30. *Намиот Д. Е.*, *Зубарева Е. В.* О работе AI Red Team // International Journal of Open Information Technologies. 2023. Т. 11, № 10. С. 130–139.

Нагибин Дмитрий Викторович

аспирант кафедры автоматики и телемеханики, Пермский национальный исследовательский политехнический университет (ПНИПУ, 614990, Пермь, Комсомольский проспект, д. 29), e-mail: dvnagibin@pstu.ru.

Кокоулин Андрей Николаевич

к.т.н., доцент кафедры автоматики и телемеханики, Пермский национальный исследовательский политехнический университет (ПНИПУ, 614990, Пермь, Комсомольский проспект, д. 29); ФБУН «Федеральный научный центр медико-профилактических технологий управления рисками здоровью населения» (614045, Пермь, ул. Монастырская, 82), e-mail: a.n.kokoulin@at.pstu.ru.

Южаков Александр Анатольевич

д.т.н., профессор, заведующий кафедрой автоматики и телемеханики, Пермский национальный исследовательский политехнический университет (ПНИПУ, 614990, Пермь, Комсомольский проспект, д. 29), e-mail: uz@at.pstu.ac.ru.

Авторы прочитали и одобрили окончательный вариант рукописи.

Авторы заявляют об отсутствии конфликта интересов.

Вклад соавторов: *Каждый автор внес равную долю участия как во все этапы проводимого теоретического исследования, так и при написании разделов данной статьи.*

Key principles for building an MLOps information security subsystem

Dmitry V. Nagibin[1], Andrey N. Kokoulin[1,2], Alexander A. Yuzhakov[1]

[1] Perm National Research Polytechnic University (PNRPU)
[2] Federal Scientific Center for Medical and Preventive Health Risk Management Technologies

Abstract: The article examines the key principles for building an information security subsystem for MLOps (Machine Learning Operations), taking into account the increasing relevance and vulnerability of machine learning systems in modern conditions. The work considers the main regulatory requirements for information security of systems using machine learning, including an analysis of the requirements of the Russian Federal Service for Technical and Export Control (FSTEC). It argues for the need to adapt modern MLSecOps solutions to the conditions of Russian legislation. The analysis covers the key cybersecurity risks characteristic of each stage of the MLOps lifecycle, as well as existing methods and approaches for mitigating them. The work presents a mathematical formalization for assessing the security of the system, taking into account technical, regulatory, and import substitution requirements. The conceptual structure of the information security subsystem is presented. The results of the research are aimed at creating an effective and reliable protection system for MLOps pipelines and reducing the risks associated with the use of machine learning technologies.

Keywords: MLOps, MLSecOps, information security, machine learning security, cybersecurity risks in MLOps, security subsystem in MLOps, russian regulatory documents on AI security.

For citation: Nagibin D. V., Kokoulin A. N., Yuzhakov A. A. Klyuchevye printsipy postroeniya podsistemy informatsionnoi bezopasnosti MLOps [Key principles for building an MLOps information security subsystem]. *Vestnik SibGUTI*, 2026, vol. 20, no. 2, pp. 77-92. https://doi.org/10.55648/1998-6920-2026-20-2-77-92.

The article was submitted: 18.03.2026;
accepted for publication 22.04.2026.

References

1. Kreuzberger D., Kühl N., Hirschl S. Machine Learning Operations (MLOps): Overview, Definition, and Architecture. *IEEE Access*, 2023, vol. 11, pp. 31866–31879.

2. Patel R., et al. Towards Secure MLOps: Surveying Attacks, Mitigation Strategies, and Research Challenges. *arXiv*:2506.02032 [cs], 2026.
3. Wilson S. *The Developer's Playbook for Large Language Model Security: Building Secure AI Applications*. First edition. Beijing, Boston, Farnham, Sebastopol, Tokyo, O'Reilly, 2024. 200 p.
4. Ukaz Prezidenta Rossiiskoi Federatsii ot 10.10.2019 g. no. 490. O razvitii iskusstvennogo intellekta v Rossiiskoi Federatsii (V redaktsii Ukaza Prezidenta Rossiiskoi Federatsii ot 15.02.2024 no. 124) [Decree of the President of the Russian Federation No. 490 of October 10, 2019. On the Development of Artificial Intelligence in the Russian Federation (As Amended by Decree of the President of the Russian Federation No. 124 of February 15, 2024)].
5. Regulyatornye dokumenty RF po bezopasnosti II – s chem my vstupaem v 2026 god [Regulatory Documents of the Russian Federation on AI Security – What We Are Entering in 2026]. *Khabr*. Available at: https://habr.com/ru/articles/986800/ (accessed: 07.02.2026).
6. Federal'naya sluzhba po tekhnicheskomu i eksportnomu kontrolyu. Trebovaniya o zashchite informatsii, soderzhashcheisya v gosudarstvennykh informatsionnykh sistemakh, inykh informatsionnykh sistemakh gosudarstvennykh organov, gosudarstvennykh unitarnykh predpriyatii, gosudarstvennykh uchrezhdenii: utverzhdeny prikazom FSTEK Rossii ot 11 aprelya 2025 g. no. 117 [Federal Service for Technical and Export Control. Requirements for the Protection of Information Contained in State Information Systems, Other Information Systems of State Bodies, State Unitary Enterprises, and State Institutions: Approved by Order of FSTEC of Russia No. 117 of April 11, 2025].
7. Analiz klyuchevykh izmenenii v trebovaniyakh k zashchite informatsii soglasno Prikazu FSTEK no. 117 [Analysis of Key Changes in Information Protection Requirements According to FSTEC Order No. 117]. Available at: https://sec.ussc.ru/fstec_117 (accessed: 08.02.2026).
8. Federal'naya sluzhba po tekhnicheskomu i eksportnomu kontrolyu. Metodicheskii dokument. Metodika analiza zashchishchennosti informatsionnykh system. Utverzhdena 25 noyabrya 2025 g. [Federal Service for Technical and Export Control. Methodological Document. Methodology for Analyzing the Security of Information Systems. Approved November 25, 2025].
9. GOST R 70462.1-2022/ISO/IEC TR 24029-1-2021. Informatsionnye tekhnologii. Intellekt iskusstvennyi. Otsenka robastnosti neironnykh setei [GOST R 70462.1-2022/ISO/IEC TR 24029-1-2021. Information Technology. Artificial Intelligence. Assessment of Neural Network Robustness].
10. Namiot D. E. Skhemy atak na modeli mashinnogo obucheniya [Attack Schemes on Machine Learning Models]. *International Journal of Open Information Technologies*, 2023, vol. 11, no. 5, pp. 68–86.
11. Lavaur L., Busnel Y., Autrel F. Investigating the Impact of Label-flipping Attacks against Federated Learning for Collaborative Intrusion Detection. *Computers & Security*, 2025, vol. 156, pp. 104462.
12. Hall P. *Machine Learning for High-Risk Applications: Approaches to Responsible AI*. First Edition. Québec, O'Reilly Media, Incorporated, 2023. 466 p.
13. Sotiropoulos J. *Adversarial AI Attacks, Mitigations, and Defense Strategies: A Cybersecurity Professional's Guide to AI Attacks, Threat Modeling, and Securing AI with MLSecOps*. Packt Publishing, 2024. 602 p.
14. Namiot D. E., Il'yushin E. A., Chizhov I. V. Ataki na sistemy mashinnogo obucheniya – obshchie problemy i metody [Attacks on Machine Learning Systems – General Problems and Methods]. *International Journal of Open Information Technologies*, 2022, vol. 10, no. 3, pp. 17–22.
15. Kovacs E. "EchoLeak" AI Attack Enabled Theft of Sensitive Data via Microsoft 365 Copilot. Available at: https://www.securityweek.com/echoleak-ai-attack-enabled-theft-of-sensitive-data-via-microsoft-365-copilot/ (accessed: 08.02.2026).
16. Gao Y., et al. Backdoor Attacks and Countermeasures on Deep Learning: A Comprehensive Review. *arXiv*:2007.10760 [cs], 2020.
17. Sutton O. J., et al. Staining and locking computer vision models without retraining. *arXiv*:2507.22000 [cs], 2025.
18. MLOps Market Outlook from 2026 to 2033: Trends by Application, by Region, and 12.8% CAGR Forecast. Available at: https://www.linkedin.com/pulse/mlops-market-outlook-from-2026-2033-trends-application-region-dswuf (accessed: 02.03.2026).
19. Giskard – LLM Agent Testing & Evaluation Platform. Available at: https://docs.giskard.ai/index.html (accessed: 08.02.2026).
20. Center S. I. S. ModelScan – Protection Against Model Serialization Attacks. Available at: https://isc.sans.edu/diary/31692 (accessed: 08.02.2026).

21. LLM Monitoring dlya GenAI-prilozhenii | AI Security Lab ITMO [LLM Monitoring for GenAI Applications | AI Security Lab ITMO]. Available at: https://hivetrace.ru/ (accessed: 08.02.2026).
22. LLAMATOR. Available at: https://llamator-core.github.io/llamator/ (accessed: 08.02.2026).
23. Garda DLP – sistema predotvrashcheniya utechek informatsii | Prodazha i vnedrenie ot Rossiiskogo razrabotchika «Garda» [Garda DLP – Information Leak Prevention System | Sales and Implementation by the Russian Developer "Garda"]. Available at: https://garda.ai/products/data-protection/dlp (accessed: 08.02.2026).
24. PT Dephaze – avtopentest, kotoryi pokazhet, chto smozhet sdelat' khaker v infrastrukture [PT Dephaze – An Auto-Penetration Test That Will Show What a Hacker Can Do in the Infrastructure]. Available at: https://ptsecurity.com/products/dephaze/ (accessed: 08.02.2026).
25. Kim J., et al. Security for the scientific data services framework. *2015 IEEE International Conference on Big Data (Big Data)*, 2015, pp. 1871–1875.
26. Spadari V., et al. An MLOps Framework for Explainable Network Intrusion Detection with MLflow. *2024 IEEE Symposium on Computers and Communications (ISCC)*, Paris, France, IEEE, 2024, pp. 1–6.
27. Tete S. B. Threat Modelling and Risk Analysis for Large Language Model (LLM)-Powered Applications. *arXiv*:2406.11007 [cs], 2024.
28. Hassija V., et al. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. *Cognitive Computation*, 2024, vol. 16, no. 1, pp. 45–74.
29. Awesome MLSecOps. Available at: https://github.com/RiccardoBiosas/awesome-MLSecOps (accessed: 04.02.2026).
30. Namiot D. E., Zubareva E. V. O rabote AI Red Team [On the Work of AI Red Team]. *International Journal of Open Information Technologies*, 2023, vol. 11, no. 10, pp. 130–139.

Dmitry V. Nagibin

Post-graduate student of the Department of «Automation and Telemechanics», Perm National Research Polytechnic University (PNRPU, Russia, 614990, Perm, Komsomolsky prospekt, 29), e-mail: dvnagibin@pstu.ru

Andrey N. Kokoulin

Candidate of Technical Sciences, Associate Professor of the Department of «Automation and Telemechanics», Perm National Research Polytechnic University (PNRPU, Russia, 614990, Perm, Komsomolsky prospekt, 29); Federal Scientific Center for Medical and Preventive Health Risk Management Technologies (Russia, 614045, Perm, Monastyrskaya St., 82), e-mail: a.n.kokoulin@at.pstu.ru

Alexander A. Yuzhakov

Dr. of Sci. (Engineering), Professor, Head of the Department of «Automation and Telemechanics», Perm National Research Polytechnic University (PNRPU, Russia, 614990, Perm, Komsomolsky prospekt, 29). e-mail: uz@at.pstu.ac.ru.