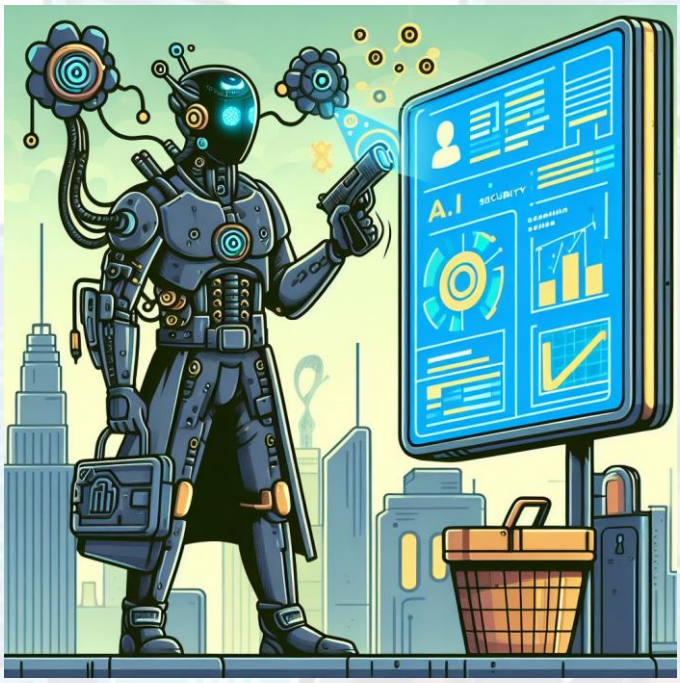

Аннотация – В этом документе представлен анализ DASF, изучается его структура, рекомендации и практические приложения, которые он предлагает организациям, внедряющим решения в области искусственного интеллекта. Этот анализ не только служит качественной экспертизой, но также подчёркивает его важность и практическую пользу для экспертов по безопасности и профессионалов из различных секторов. Внедряя руководящие принципы и средства контроля, рекомендованные DASF, организации могут защитить свои активы искусственного интеллекта от возникающих угроз и уязвимостей.

I. Введение

Databricks AI Security Framework (DASF) представляет собой всеобъемлющее руководство, разработанное для устранения возникающих рисков, связанных с повсеместной интеграцией ИИ по всему миру и направлено на предоставление действенных рекомендаций по защитному контролю для систем ИИ, охватывающих весь жизненный цикл ИИ и облегчающих сотрудничество между бизнесом, ИТ, данными, ИИ и командами безопасности. DASF не ограничивается моделями защиты или конечными точками, но применяет целостный подход к снижению кибер-рисков в системах ИИ.

DASF идентифицирует 55 технических рисков безопасности в 12 основополагающих компонентов общей системы ИИ, ориентированной на данные, включая необработанные данные, подготовку данных, наборы данных, управление каталогом данных, алгоритмы машинного обучения, оценку, модели машинного обучения, управление моделями, обслуживание моделей и вывод, реакцию на вывод, операции машинного обучения (MLOP), а также безопасность данных и платформы искусственного интеллекта. Каждый риск сопоставляется с набором мер по смягчению последствий, которые ранжируются в приоритетном порядке, начиная с обеспечения безопасности периметра и заканчивая безопасностью данных.

Фреймворк анализа данных выделяется как ключевой компонент DASF, предлагающий единую основу для всех данных и управления и включает в себя Mosaic AI, Unity, архитектуру и безопасность платформы Databricks. Mosaic AI охватывает сквозной рабочий процесс ИИ, в то время как Unity Catalog предоставляет унифицированное решение для управления данными и активами ИИ. Архитектура представляет собой гибридный PaaS, не зависящий от данных, а безопасность основана на принципах доверия, технологий и прозрачности.

DASF предназначен для групп безопасности, практиков ML, руководителей и инженерных команд DevSecOps. Это обеспечивает структурированный подход к новым угрозам и способах их устранения, не требуя привлечения глубоких экспертных знаний. DASF также включает подробное руководство по пониманию безопасности и соответствия конкретным системам ML, предлагающее информацию о том, как ML влияет на безопасность системы, применение принципов разработки безопасности к ML и предоставляющее подробное руководство по пониманию безопасности и соответствия конкретным системам ML.

DASF завершается рекомендациями о том, как безопасно управлять моделями искусственного интеллекта и внедрять их в соответствии с основными принципами внедрения машинного обучения: определения бизнес-сценария использования ML, определения модели развёртывания ML, перечисления угроз и рисков для каждого риска и выбора средства контроля для внедрения. В нем также содержится дополнительная информация для расширения знаний в области искусственного интеллекта и фреймворков, рассмотренных в рамках анализа.

Ключевые моменты:

- **Совместное использование**: DASF разработан для совместного использования командами обработки данных и искусственного интеллекта вместе с их коллегами по безопасности. Это подчёркивает важность совместной работы этих команд на протяжении всего жизненного цикла искусственного интеллекта для обеспечения безопасности и соответствия требованиям систем искусственного интеллекта.
- **Применимость в разных командах**: Концепции DASF применимы ко всем командам, независимо от того, используют ли они Databricks для создания своих решений в области ИИ. Такая инклюзивность гарантирует, что фреймворк может быть использован широкой аудиторией для повышения безопасности искусственного интеллекта.
- **Руководство по типам моделей ИИ**: предлагается, чтобы организации сначала определили, какие типы моделей искусственного интеллекта создаются или используются. В нем модели в широком смысле подразделяются на прогнозирующие модели ML, современные открытые модели и внешние модели,

обеспечивая основу для понимания конкретных соображений безопасности для каждого типа.

- **Понимание компонентов системы ИИ**: организациям рекомендуется ознакомиться с основополагающими компонентами общей системы искусственного интеллекта, ориентированной на данные, как описано в документе.
- **Идентификация рисков и снижение их последствий**: DASF помогает организациям выявлять соответствующие риски и определять применимые средства контроля на основе всеобъемлющего списка, представленного в документе. Такой структурированный подход помогает расставить приоритеты в мерах безопасности на основе конкретных потребностей организации.
- **Документация и функции в терминологии Databricks**: Цель подхода сделать документ полезным для более широкой аудитории, сохраняя при этом его практичность для пользователей Databricks.

II. Целевая аудитория

- **Команды безопасности**: CISO, руководители служб безопасности, DevSecOps, SRE-инженеры и другие лица, ответственные за безопасность систем. Они могут использовать DASF, чтобы понять, как машинное обучение (ML) повлияет на безопасность системы, и понять основные механизмы ML.
- **Инженеры ML**: инженеры по обработке данных, архитекторы данных, инженеры по обработке данных и специалисты по обработке данных. DASF помогает им понять, как инженерия безопасности и менталитет "защищённости по замыслу" могут быть применены к ML.
- **Governance-сотрудники**: отвечают за обеспечение соответствия данных и методов ИИ в организации соответствующим законам, нормативным актам и политике. DASF предоставляет рекомендации о том, как ML влияет на безопасность системы и соответствие требованиям.
- **Команды инженеров DevSecOps**: отвечают за интеграцию безопасности в процессы разработки и эксплуатации. DASF предлагает этим командам структурированный способ обсуждения новых угроз и мер по их устранению, не требующий обмена глубокими знаниями.

III. Преимущества и Недостатки

DASF предлагает всеобъемлющее и практическое руководство для организаций, стремящихся понять риски безопасности искусственного интеллекта и снизить их. Однако его сложность и ориентированное на блоки данных руководство могут представлять проблемы для некоторых организаций.

A. Преимущества

- **Целостный подход**: DASF применяет целостный подход к безопасности ИИ, устраняя риски на протяжении всего жизненного цикла ИИ и всех компонентов универсальной системы ИИ, ориентированной на данные. Такой комплексный подход помогает организациям более эффективно выявлять риски безопасности и снижать их уровень.
- **Сотрудничество**: Фреймворк предназначен для облегчения взаимодействия между бизнесом, ИТ, данными, искусственным интеллектом и командами безопасности. Это поощряет единый подход к обеспечению безопасности искусственного интеллекта и помогает преодолеть разрыв между различными дисциплинами.
- **Практические рекомендации**: DASF предоставляет практические рекомендации по защитному контролю для каждого выявленного риска, которые могут обновляться по мере появления новых рисков и появления дополнительных средств контроля. Это гарантирует, что организации смогут оставаться в курсе возникающих угроз безопасности искусственного интеллекта.
- **Применимость**: DASF применим к организациям, использующим различные модели ИИ, включая прогнозирующие модели ML, генеративные модели искусственного интеллекта и внешние модели. Такая широкая применимость делает его ценным ресурсом для широкого круга организаций.
- **Интеграция с платформой анализа данных Databricks**: для организаций, использующих платформу анализа данных Databricks, DASF предлагает конкретные рекомендации по использованию средств управления рисками искусственного интеллекта платформы. Это помогает организациям максимально использовать преимущества платформы в области безопасности.

B. Недостатки

- **Сложность**: DASF охватывает широкий спектр рисков для безопасности ИИ и мер по их снижению, которые могут оказаться непосильными для организаций, не знакомых с безопасностью ИИ или имеющих ограниченные ресурсы, а внедрение системы может потребовать значительных затрат времени и усилий.
- **Руководство, ориентированное на Databricks**: хотя DASF предлагает рекомендации для организаций, использующих платформу Databricks Data Intelligence Platform, некоторые рекомендации могут быть менее применимы или неосуществимы для организаций, использующих другие платформы или инструменты искусственного интеллекта.
- **Меняющийся ландшафт**: поскольку ландшафт безопасности ИИ продолжает развиваться,

организациям, возможно, потребуется постоянно обновлять свои средства контроля и практики обеспечения безопасности, чтобы оставаться актуальными.

- **Отсутствие конкретных примеров**: DASF предоставляет высокоуровневый обзор рисков безопасности ИИ и средств контроля за их снижением, но в нем отсутствуют конкретные примеры или тематических исследований, иллюстрирующих, как эти риски и средства контроля применяются в реальных сценариях.

- **Фокус на технических рисках**: DASF в первую очередь фокусируется на технических рисках безопасности и средствах контроля за их снижением. Хотя это важный аспект безопасности искусственного интеллекта, организациям следует также учитывать нетехнические риски, такие как этические, юридические и социальные последствия искусственного интеллекта, которые недостаточно подробно рассматриваются в DASF.

IV. Связь с индустриальными документами

DASF предназначена для дополнения и интеграции с другими практиками безопасности, такими как NIST, HITRUST, ISO / IEC 27001 и 27002, а также критически важными средствами контроля безопасности CIS. DASF применяет целостный подход к снижению рисков безопасности ИИ вместо того, чтобы сосредотачиваться только на безопасности моделей или конечных точек модели. Такой подход соответствует принципам этих фреймворков, которые обеспечивают структурированный процесс выявления, оценки и снижения рисков безопасности.

V. Фреймворк DASF

Фреймворк подразделяет систему ИИ на 12 основных компонентов, каждый из которых связан с конкретными рисками безопасности, выявленными в результате тщательного анализа. Этот анализ включает прогнозирующие модели ML, генеративные базовые модели и внешние модели, основанные на запросах клиентов, оценках безопасности и т.д. Затем идентифицированные риски сопоставляются с соответствующими средствами контроля в рамках фреймворка анализа данных Databricks со ссылками на подробную документацию продукта по каждому риску.

В документе описываются компоненты системы искусственного интеллекта и связанные с ними риски следующим образом:

- **Операции с данными**: этап включает первоначальную обработку необработанных данных, включая приём, преобразование и обеспечение безопасности данных и управления ими. В общей сложности в этой категории идентифицировано 19 конкретных рисков, начиная от недостаточного контроля доступа и заканчивая отсутствием комплексного управления жизненным циклом ML.

- **Операции с моделями**: этап включает в себя создание моделей ML, будь то путём построения прогнозирующих моделей, приобретения моделей на торговых площадках или использования API, таких как OpenAI. Для этого требуется серия экспериментов и механизмы отслеживания для сравнения различных условий и результатов. Выявлено 14 конкретных рисков, включая такие проблемы, как недостаточная воспроизводимость эксперимента и отклонение модели.

- **Развёртывание и обслуживание модели**: основное внимание уделяется безопасному развёртыванию образов моделей, обслуживанию моделей и управлению такими функциями, как автоматическое масштабирование и ограничение скорости. Всего выделено 15 конкретных рисков, включая оперативный ввод и инверсию модели.

- **Операции и платформа**: заключительный этап включает управление уязвимостями платформы, исправление ошибок, изоляцию модели и обеспечение авторизованного доступа к моделям со встроенной в архитектуру защитой. Это также включает в себя операционный инструментарий для CI / CD для поддержания безопасных MLOP в средах разработки, промежуточных и производственных средах. Определены семь конкретных рисков, таких как отсутствие стандартов MLOP и управление уязвимостями.

VI. Необработанные данные

- **Важность необработанных данных**: Необработанные данные являются основой систем ИИ, охватывая корпоративные данные, метаданные и оперативные данные в различных формах, таких как полуструктурированные или неструктурированные данные, пакетные данные или потоковые данные.

- **Безопасность данных**: Защита необработанных данных имеет первостепенное значение для целостности алгоритмов машинного обучения и любых технических деталей развёртывания. Это сопряжено с уникальными проблемами, и весь сбор данных в системе искусственного интеллекта сопряжён как со стандартными проблемами безопасности данных, так и с новыми.

- **Меры по снижению рисков**: описываются конкретные риски, связанные с необработанными данными, и приводятся подробные меры по снижению рисков для каждого из них. Эти средства контроля включают эффективное управление доступом, классификацию данных, обеспечение качества данных, хранение и шифрование, управление версиями данных, происхождение данных, достоверность данных, юридические соображения, обработку устаревших данных и журналов доступа к данным.

- **Управление доступом**: Обеспечение того, чтобы только авторизованные лица или группы могли получить доступ к определённым наборам данных, имеет фундаментальное значение для безопасности данных. Это включает аутентификацию, авторизацию и точно настроенные средства контроля доступа.
- **Классификация данных**: Классификация данных имеет решающее значение для управления, позволяя организациям сортировать и категоризировать данные по степени чувствительности, важности и критичности, что важно для реализации соответствующих мер безопасности и политик управления.
- **Качество данных**: Высокое качество данных имеет решающее значение для принятия надёжных решений на основе данных и является краеугольным камнем управления данными. Организации должны тщательно оценивать ключевые атрибуты данных для обеспечения точности анализа и экономической эффективности.
- **Хранение и шифрование**: Шифрование данных в состоянии покоя и при передаче имеет жизненно важное значение для защиты от несанкционированного доступа и соблюдения отраслевых правил безопасности данных.
- **Управление версиями данных и их происхождение**: Управление версиями данных и отслеживание журналов изменений важны для отката или отслеживания исходных данных в случае повреждения. Data lineage помогает обеспечить соответствие требованиям и готовность к аудиту, обеспечивая чёткое понимание и отслеживаемость данных, используемых для ИИ.
- **Достоверность и юридические аспекты**: Обеспечение достоверности данных и соблюдения юридических требований, таких как GDPR и CCPA, имеет важное значение. Это включает в себя возможность "удалять" определённые данные из систем машинного обучения и переобучать модели, используя чистые наборы данных, подтверждённые владельцами.
- **Устаревшие данные и журналы доступа**: Устранение рисков, связанных с устаревшими данными и отсутствием журналов доступа к данным, важно для поддержания эффективности и безопасности бизнес-процессов. Надлежащие механизмы аудита имеют решающее значение для обеспечения безопасности данных и соблюдения нормативных требований.

VII. Подготовка данных

- **Определение и важность**: Подготовка данных определяется как процесс преобразования необработанных входных данных в формат, который могут интерпретировать алгоритмы машинного обучения. Этот этап имеет решающее значение, поскольку он напрямую влияет на безопасность и объяснимость системы ML.
- **Риски безопасности и меры по их устранению**: В разделе описываются различные риски безопасности, связанные с подготовкой данных, и приводятся подробные меры по их устранению для каждого. Эти риски включают целостность предварительной обработки, манипулирование функциями, критерии исходных данных и состязательные разделы.
- **Целостность предварительной обработки**: Обеспечение целостности предварительной обработки включает числовые преобразования, агрегирование данных, кодирование текста или изображений и создание новых функций. Меры по смягчению последствий включают настройку единого входа (SSO) с помощью поставщика идентификационных данных (IdP) и многофакторной аутентификации (MFA), ограничение доступа с использованием списков доступа IP и реализацию частных ссылок для ограничения источника входящих запросов.
- **Манипулирование объектами**: Этот риск связан с возможностью того, что злоумышленники могут манипулировать тем, как данные аннотируются к объектам, что может поставить под угрозу целостность и точность модели. Элементы управления включают защиту функций модели для предотвращения несанкционированных обновлений и использование ориентированных на данные MLOP и LLMOP для продвижения моделей в виде кода.
- **Критерии необработанных данных**: Понимание критериев отбора необработанных данных важно для предотвращения внесения злоумышленниками вредоносного ввода, который ставит под угрозу целостность системы. Элементы управления включают использование списков контроля доступа и ориентированного на данные MLOP для модульного тестирования и интеграции.
- **Состязательные разделы**: это связано с риском того, что злоумышленники повлияют на разделение наборов данных, используемых при обучении и оценке, потенциально косвенно контролируя систему ML. Смягчение последствий включает отслеживание и воспроизведение обучающих данных, используемых для обучения модели ML, и идентификацию моделей ML и запусков, полученных на основе определённого набора данных.
- **Комплексные стратегии смягчения последствий**: В разделе подчёркивается важность комплексного подхода к обеспечению безопасности процесса подготовки данных, включая использование строгих мер безопасности для защиты от манипуляций, которые могут подорвать целостность и надёжность систем ML

VIII. Наборы данных

- **Значимость наборов данных**: Наборы данных имеют решающее значение для обучения, валидации и тестирования моделей машинного обучения. Ими необходимо тщательно управлять, чтобы обеспечить целостность и эффективность систем искусственного интеллекта.
- **Риски безопасности**: В разделе описываются различные риски безопасности, связанные с наборами данных, включая отравление данных, неэффективное хранение и шифрование, а также переворачивание этикеток. Эти риски могут поставить под угрозу надёжность и производительность моделей машинного обучения.
- **Отравление данными**: Этот риск связан с тем, что злоумышленники манипулируют обучающими данными, чтобы повлиять на выходные данные модели на этапе вывода. Стратегии смягчения последствий включают надёжный контроль доступа, проверки качества данных и мониторинг цепочки данных для предотвращения несанкционированных манипуляций с данными.
- **Неэффективное хранение и шифрование**: Надлежащее хранение и шифрование данных имеют решающее значение для защиты наборов данных от несанкционированного доступа и утечек. Фреймворк рекомендует шифрование данных в состоянии покоя и при передаче, а также строгий контроль доступа.
- **Переключение меток**: Этот специфический тип отравления данными включает изменение меток в обучающих данных, что может ввести модель в заблуждение во время обучения и снизить её производительность. Для снижения этого риска рекомендуется использовать шифрование и безопасный доступ к наборам данных.
- **Меры по смягчению последствий**: для каждого выявленного риска DASF предоставляет подробные меры по смягчению последствий. Эти средства контроля включают использование единого входа (SSO) с поставщиками идентификационных данных (IdP), многофакторной аутентификации (MFA), списков доступа по IP, частных ссылок и шифрования данных для повышения безопасности наборов данных.
- **Комплексное управление рисками**: В этом разделе подчёркивается важность комплексного подхода к управлению безопасностью набора данных, начиная с первоначального сбора данных и заканчивая внедрением моделей машинного обучения. Это включает регулярные аудиты, обновления протоколов безопасности и постоянный мониторинг целостности данных.

IX. Управление каталогом данных

- **Комплексный подход к управлению**: Каталог данных и управление включают управление информационными активами организации на протяжении всего их жизненного цикла, что включает принципы, практики и инструменты эффективного управления.
- **Централизованный контроль доступа**: Управление данными и активами искусственного интеллекта обеспечивает централизованный контроль доступа, аудит, происхождение, данные и возможности обнаружения моделей, что ограничивает риск дублирования данных или моделей, ненадлежащего использования секретных данных для обучения, потери происхождения и кражи моделей.
- **Конфиденциальность и безопасность данных**: при работе с наборами данных, которые могут содержать конфиденциальную информацию, крайне важно обеспечить надлежащую защиту личной информации (PII) и других конфиденциальных данных для предотвращения взломов и утечек. Это особенно важно в секторах с жесткими нормативными требованиями.
- **Контрольные журналы и прозрачность**: Надлежащее управление каталогом данных позволяет проводить контрольные журналы и отслеживать происхождение и преобразования данных, используемых для обучения моделей искусственного интеллекта. Такая прозрачность способствует укреплению доверия и подотчётности, снижает риск предвзятости и улучшает результаты искусственного интеллекта.
- **Соответствие нормативным требованиям**: Обеспечение надлежащей защиты конфиденциальной информации в наборах данных имеет важное значение для соблюдения таких нормативных актов, как GDPR и CCPA. Это включает в себя возможность демонстрировать безопасность данных и вести журналы аудита.
- **Панель мониторинга совместной работы**: для проектов компьютерного зрения с участием нескольких заинтересованных сторон наличие простого в использовании инструмента маркировки с панелью мониторинга совместной работы важно для того, чтобы все были в курсе событий в режиме реального времени и не допускали искажения миссии.
- **Автоматизированные конвейеры данных**: для проектов с большими объемами данных автоматизация конвейеров данных путём подключения наборов данных и моделей с помощью API может упростить процесс и ускорить обучение моделей ML.
- **Рабочие процессы контроля качества**: важно иметь настраиваемые и управляемые рабочие процессы контроля качества для проверки меток и

аннотаций, уменьшения ошибок и предвзятости, а также исправления ошибок в наборах данных. Автоматизированные инструменты аннотирования могут помочь в этом процессе

X. Алгоритмы машинного обучения

- **Техническое ядро систем ML**: Алгоритмы машинного обучения описываются как техническое ядро любой системы ML, имеющее решающее значение для функциональности и безопасности системы.
- **Меньший риск для безопасности**: отмечается, что атаки на алгоритмы машинного обучения обычно представляют значительно меньший риск для безопасности по сравнению с данными, используемыми для обучения, тестирования и последующей эксплуатации.
- **Автономные и онлайновые системы**: В этом разделе проводится различие между автономными и онлайновыми алгоритмами машинного обучения. Автономные системы обучаются на фиксированном наборе данных и затем используются для прогнозирования, в то время как онлайн-системы постоянно обучаются и адаптируются посредством итеративного обучения с новыми данными.
- **Преимущества автономных систем в плане безопасности**: считается, что автономные системы обладают определёнными преимуществами в плане безопасности из-за их фиксированного, статичного характера, который уменьшает поверхность атаки и со временем сводит к минимуму подверженность уязвимостям, связанным с данными.
- **Уязвимости онлайн-систем**: Онлайн-системы постоянно подвергаются воздействию новых данных, что повышает их восприимчивость к атакам отравления, состязательному вводу данных и манипулированию процессами обучения.
- **Тщательный выбор алгоритмов**: подчёркивается важность тщательного рассмотрения выбора между автономными и онлайн-алгоритмами обучения на основе конкретных требований безопасности и операционной среды системы ML

XI. Оценка

- **Критическая роль оценки**: Оценка необходима для оценки эффективности систем машинного обучения в достижении их предполагаемых функциональных возможностей. Это включает в себя использование выделенных наборов данных для систематического анализа производительности обученной модели с учётом её конкретной задачи.
- **Отравление оценочных данных**: существует риск атак на данные, когда данные подделываются перед их использованием для машинного обучения, что значительно усложняет обучение и оценку моделей ML. Эти атаки могут повредить или изменить данные таким образом, что исказит процесс обучения, что приведёт к созданию ненадёжных моделей.
- **Неполнота данных для оценки**: Наборы данных для оценки также могут быть слишком маленькими или слишком похожими на данные для обучения, чтобы быть полезными. Некачественные оценочные данные могут привести к предвзятости, галлюцинациям и токсическому эффекту. Трудно эффективно оценивать большие языковые модели (LLM), поскольку эти модели редко имеют маркировку объективной истинности.
- **Меры по смягчению последствий**:
  - Внедрение единого входа (SSO) с поставщиком идентификационных данных (IdP) и многофакторной аутентификации (MFA) для ограничения доступа к данным и платформе искусственного интеллекта.
  - Использование списков IP-доступа для ограничения IP-адресов, которые могут проходить аутентификацию в Databricks.
  - Шифрование данных в состоянии покоя и при передаче.
  - Отслеживать изменения в данных и системе искусственного интеллекта с помощью единого окна и принимать меры при возникновении изменений.
- **Важность надёжной оценки**: Эффективная оценка имеет решающее значение для обеспечения надёжности и точности моделей машинного обучения. Это помогает выявить расхождения или аномалии в процессе принятия решений в модели и даёт представление о производительности модели.

XII. Модели машинного обучения

- **Безопасность модели**: Модели машинного обучения являются ядром систем ИИ, и их безопасность имеет решающее значение для обеспечения целостности и надёжности системы. В этом разделе обсуждаются различные риски, связанные с моделями машинного обучения, и предлагаются меры по снижению каждого риска.
- **Бэкдорное машинное обучение / Троянская модель**: риск связан с внедрением злоумышленником бэкдора в модель во время обучения, который может быть использован позже для манипулирования поведением модели. Меры по смягчению последствий включают мониторинг производительности модели, использование надёжных обучающих данных и внедрение средств контроля доступа.
- **Утечка ресурсов модели**: риск связан с несанкционированным раскрытием ресурсов модели, таких как архитектура модели, веса и обучающие данные. Меры по смягчению

последствий включают шифрование, контроль доступа и мониторинг на предмет несанкционированного доступа.

- **Уязвимости цепочки поставок ML**: риск возникает из-за уязвимостей в цепочке поставок ML, таких как библиотеки сторонних производителей и зависимости. Меры по смягчению последствий включают регулярные оценки уязвимостей, использование надёжных источников и внедрение безопасных методов разработки.
- **Атака под контролем исходного кода**: риск связан с получением несанкционированного доступа к хранилищу исходного кода и модификацией кода для внедрения уязвимостей или бэкдоров. Меры по смягчению последствий включают контроль доступа, проверку кода и мониторинг на предмет несанкционированного доступа.
- **Указание принадлежности к модели**: риск связан с несанкционированным использованием модели без надлежащего указания принадлежности к её первоначальным создателям. Меры по смягчению последствий включают использование цифровых водяных знаков, ведение надлежащей документации и обеспечение соблюдения лицензионных соглашений.
- **Кража модели**: риск связан с кражей злоумышленником модели путём реверс-инжиниринга её поведения или прямого доступа к её коду. Меры по смягчению последствий включают шифрование, контроль доступа и мониторинг на предмет несанкционированного доступа.
- **Жизненный цикл модели без HITL**: риск возникает из-за отсутствия участия человека в цикле (HITL) в жизненном цикле модели, что может привести к предвзятым или неверным прогнозам. Меры по смягчению последствий включают регулярную валидацию модели, анализ со стороны персонала и непрерывный мониторинг.
- **Инверсия модели**: риск связан с тем, что злоумышленник получает конфиденциальную информацию об обучающих данных путём анализа поведения модели. Меры по смягчению последствий включают использование дифференцированной конфиденциальности, контроль доступа и мониторинг на предмет несанкционированного доступа.

XIII. Управление моделями

- **Управление моделями**: Управление моделями — это процесс организации, отслеживания и поддержки моделей машинного обучения на протяжении всего их жизненного цикла, от разработки до развёртывания и вывода из эксплуатации.
- **Риски безопасности**: риски безопасности связаны с управлением моделями, включая атрибуцию модели, кражу модели, жизненный цикл модели без участия человека в цикле (HITL) и инверсию модели.
- **Указание принадлежности к модели**: риск связан с несанкционированным использованием модели без надлежащего указания принадлежности к её первоначальным создателям. Меры по смягчению последствий включают использование цифровых водяных знаков, ведение надлежащей документации и обеспечение соблюдения лицензионных соглашений.
- **Кража модели**: риск связан с кражей злоумышленником модели путём реверс-инжиниринга её поведения или прямого доступа к её коду. Меры по смягчению последствий включают шифрование, контроль доступа и мониторинг на предмет несанкционированного доступа.
- **Жизненный цикл модели без HITL**: риск возникает из-за отсутствия участия человека в цикле (HITL) в жизненном цикле модели, что может привести к предвзятым или неверным прогнозам. Меры по смягчению последствий включают регулярную валидацию модели, анализ со стороны персонала и непрерывный мониторинг.
- **Инверсия модели**: риск связан с тем, что злоумышленник получает конфиденциальную информацию об обучающих данных путём анализа поведения модели. Меры по смягчению последствий включают использование дифференцированной конфиденциальности, контроль доступа и мониторинг на предмет несанкционированного доступа.

XIV. Запросы на обслуживание моделей

- **Обслуживание модели**: Обслуживание модели – это процесс развёртывания обученной модели машинного обучения в производственной среде для генерации прогнозов на основе новых данных.
- **Запросы на вывод**: Запросы на вывод — это входные данные, отправляемые в развёрнутую модель для генерации прогнозов.
- **Риски безопасности**: различные риски безопасности, связанные с обслуживанием модели и запросами вывода, включая оперативный ввод, инверсию модели, прерывание модели, циклический ввод, определение принадлежности к обучающим данным, обнаружение онтологии модели ML, отказ в обслуживании (DoS), галлюцинации LLM, контроль входных ресурсов и случайное попадание неавторизованных данных в модели.
- **Оперативное внедрение**: Этот риск связан с тем, что злоумышленник вводит вредоносный ввод в модель для манипулирования её поведением или извлечения конфиденциальной информации.

- **Инверсия модели**: Этот риск связан с попыткой злоумышленника восстановить исходные обучающие данные или конфиденциальные функции путём наблюдения за выходными данными модели.
- **Нарушение модели**: Этот риск связан с использованием злоумышленником уязвимостей в среде, обслуживающей модель, для получения несанкционированного доступа к базовой системе или данным.
- **Циклический ввод**: Этот риск связан с тем, что злоумышленник вводит повторяющийся или закольцованный ввод в модель, что приводит к исчерпанию ресурсов или снижению производительности системы.
- **Определение принадлежности к обучающим данным**: Этот риск связан с попыткой злоумышленника определить, использовалась ли конкретная точка данных в обучающих данных модели.
- **Обнаружение онтологии модели ML**: Этот риск связан с попыткой злоумышленника извлечь информацию о внутренней структуре или функциональности модели.
- **Отказ в обслуживании (DoS)**: Этот риск связан с тем, что злоумышленник отправляет большой объем запросов на вывод, чтобы перегрузить инфраструктуру, обслуживающую модель, и вызвать перебои в обслуживании.
- **Галлюцинации LLM**: Этот риск связан с тем, что модель генерирует неверные или вводящие в заблуждение выходные данные из-за присущей ей неопределённости или ограничений базовых алгоритмов.
- **Контроль входных ресурсов**: Этот риск связан с тем, что злоумышленник манипулирует входными данными для использования чрезмерных ресурсов в процессе вывода.
- **Случайное предоставление неавторизованных данных моделям**: Этот риск связан с непреднамеренным предоставлением конфиденциальных или неавторизованных данных модели в процессе вывода.
- **Меры по смягчению последствий**: для каждого выявленного риска DASF предоставляет подробные меры по смягчению последствий. Эти средства контроля включают использование единого входа (SSO) с поставщиками идентификационных данных (IdP), многофакторной аутентификации (MFA), списков доступа по IP, частных ссылок и шифрования данных для повышения безопасности обслуживания модели и вывода запросов

XV. Обслуживание модели

- **Обслуживание модели**: Обслуживание модели — это процесс развёртывания обученной модели машинного обучения в производственной среде для генерации прогнозов на основе новых данных.
- **«Ответ на логический вывод»**: относится к выходным данным, генерируемым развёрнутой моделью в ответ на входные данные, отправленные для прогнозирования.
- **Риски безопасности**: различные риски безопасности, связанные с обслуживанием модели, включая отсутствие аудита и мониторинга качества вывода, манипулирование выводом, обнаружение онтологии модели ML, обнаружение семейства моделей ML и атаки с использованием черного ящика.
- **Недостаточное качество аудита и мониторинга выводов**: риск связан с отсутствием надлежащих механизмов мониторинга и аудита для обеспечения качества и точности прогнозов модели.
- **Манипулирование выходными данными**: риск связан с тем, что злоумышленник манипулирует выходными данными модели для получения неверных или вводящих в заблуждение прогнозов.
- **Обнаружение онтологии модели ML**: риск связан с попыткой злоумышленника извлечь информацию о внутренней структуре или функциональности модели путём анализа выходных данных.
- **Обнаружение семейства моделей ML**: риск связан с попыткой злоумышленника идентифицировать конкретный тип или семейство моделей, используемых в системе, путём анализа выходных данных.
- **Атаки с использованием черного ящика**: риск связан с тем, что злоумышленник использует уязвимости модели, рассматривая её как черный ящик и манипулируя входными данными для получения желаемых результатов.
- **Меры по смягчению последствий**: для каждого выявленного риска DASF предоставляет подробные меры по смягчению последствий. Эти средства контроля включают использование единого входа (SSO) с поставщиками идентификационных данных (IdP), многофакторной аутентификации (MFA), списков доступа по IP, частных ссылок и шифрования данных для повышения безопасности обслуживания модели и вывода ответа

XVI. Машинное обучение (MLOps)

- **MLOps Определение**: MLOps – это практика объединения машинного обучения (ML), DevOps и разработки данных для автоматизации и стандартизации процесса развёртывания, обслуживания и обновления моделей ML в производственных средах.

- **Риски безопасности**: различные риски безопасности, связанные с MLOP, включая отсутствие MLOP, повторяющиеся принудительные стандарты и несоответствие требованиям.
- **Отсутствие MLOP**: риск связан с отсутствием стандартизированного и автоматизированного процесса развёртывания, обслуживания и обновления моделей ML, что может привести к несоответствиям, ошибкам и уязвимостям в системе безопасности.
- **Стандарты**: Соблюдение стандартов имеет решающее значение для обеспечения безопасности и надёжности моделей ML в производственных средах. Это включает внедрение системы контроля версий, автоматизированного тестирования и конвейеров непрерывной интеграции и развёртывания (CI / CD).
- **Несоблюдение требований**: риск связан с несоблюдением соответствующих нормативных актов и отраслевых стандартов, что может привести к юридическим и финансовым последствиям для организации.
- **Меры по смягчению последствий**: для каждого выявленного риска DASF предоставляет подробные меры по смягчению последствий. Эти средства контроля включают использование единого входа (SSO) с поставщиками идентификационных данных (IdP), многофакторной аутентификации (MFA), списков доступа по IP, частных ссылок и шифрования данных для повышения безопасности MLOP

XVII. Безопасность данных и платформы ИИ

- **Неотъемлемые риски и выгоды**: Выбор платформы, используемой для создания и развёртывания моделей ИИ, может иметь неотъемлемые риски и выгоды. Реальные данные свидетельствуют о том, что злоумышленники часто используют простые тактики для компрометации систем, основанных на ML.
- **Отсутствие реагирования на инциденты**: Приложения AI / ML критически важны для бизнеса, и поставщики платформ должны быстро и эффективно решать проблемы безопасности. Рекомендуется сочетание автоматического мониторинга и ручного анализа для устранения общих угроз и угроз, специфичных для ML (DASF 39 Platform security — Incident Response Team).
- **Несанкционированный привилегированный доступ**: Внутренние злоумышленники, такие как сотрудники или подрядчики, могут представлять серьёзную угрозу безопасности. Они могут получить несанкционированный доступ к частным учебным данным или ML-моделям, что приведёт к утечке конфиденциальной информации, злоупотреблениям бизнес-процессами и потенциальному саботажу ML-систем. Внедрение строгих мер внутренней безопасности и протоколов мониторинга имеет решающее значение для снижения инсайдерских рисков (Безопасность платформы DASF 40 — внутренний доступ).
- **Низкий уровень безопасности в жизненном цикле разработки ПО (SDLC)**: Безопасность программной платформы является важной частью любой прогрессивной программы обеспечения безопасности. Хакеры часто используют ошибки в платформе, на которой построен ИИ. Безопасность ИИ зависит от безопасности платформы (DASF 41 Platform security — secure SDLC).
- **Несоблюдение требований**: По мере того, как приложения с ИИ становятся все более распространёнными, они становятся все более объектом пристального внимания и нормативных актов, таких как GDPR и CCPA. Использование сертифицированной платформы может стать значительным преимуществом для организаций, поскольку эти платформы специально разработаны для соответствия нормативным стандартам и предоставляют необходимые инструменты и ресурсы, помогающие организациям создавать и развёртывать приложения ИИ, соответствующие этим требованиям

XVIII. Фреймворк сбора данных Databricks

Databricks Data Intelligence Platform — это комплексное решение для ИИ и управления данными.

- **Mosaic AI:** компонент платформы охватывает комплексный рабочий процесс искусственного интеллекта, от подготовки данных до развёртывания модели и мониторинга.
- **Каталог Unity**: унифицированное решение для управления данными и активами искусственного интеллекта. Он обеспечивает обнаружение данных, их привязку и детальный контроль доступа.
- **Архитектура**: представляет собой гибридный PaaS, не зависящий от данных и поддерживающий широкий спектр типов данных и источников.
- **Безопасность**: Безопасность платформы основана на принципах доверия, технологии и прозрачности. Он включает в себя такие функции, как шифрование, контроль доступа и мониторинг.
- **Средства контроля за снижением рисков ИИ**: Databricks выявила 55 технических рисков безопасности в 12 основополагающих компонентах общей системы ИИ, ориентированной на данные. Для каждого риска фреймворк предоставляет руководство по контролю за смягчением последствий с помощью ИИ и ML, общую ответственность Databricks и организации, а также соответствующую техническую документацию Databricks.

XIX. Блоки данных для управления рисками ИИ

- **Средства управления рисками ИИ Databricks**: Databricks выявила 55 технических рисков безопасности в 12 основополагающих компонентах общей системы ИИ, ориентированной на данные. Для каждого риска DASF предоставляет руководство по контролю за смягчением последствий ИИ и ML, разделению ответственности Databricks и организации, а также соответствующую техническую документацию Databricks.
- **Общая ответственность**: Ответственность за внедрение мер по смягчению последствий разделяется между Databricks и организацией, использующей платформу. Databricks предоставляет инструменты и ресурсы, необходимые для внедрения элементов управления, в то время как организация несёт ответственность за их настройку и управление в соответствии со своими конкретными потребностями.
- **Комплексный подход**: Средства управления рисками ИИ охватывают широкий спектр рисков безопасности, от защиты данных и контроля доступа до развёртывания моделей и мониторинга. Такой комплексный подход помогает организациям снизить общий риск в процессах разработки и внедрения систем искусственного интеллекта.
- **Применимость**: Средства управления рисками ИИ применимы ко всем типам моделей ИИ, включая прогнозирующие модели ML, генеративные модели ИИ и внешние модели. Это гарантирует, что организации смогут внедрить соответствующие средства контроля на основе конкретных моделей ИИ, которые они используют.
- **Оценка усилий**: Каждый элемент управления помечен как "Готовый", "Конфигурация" или "Реализация", что помогает командам оценить усилия, затраченные на внедрение элемента управления на платформе Databricks Data Intelligence Platform. Это позволяет организациям расставлять приоритеты в своих усилиях по обеспечению безопасности и эффективно распределять ресурсы