AI Technology Research

IGF Internet Governance Forum RIYADH 2024

Руководство по безопасной разработке и внедрению систем искусственного интеллекта

kaspersky активируй будущее

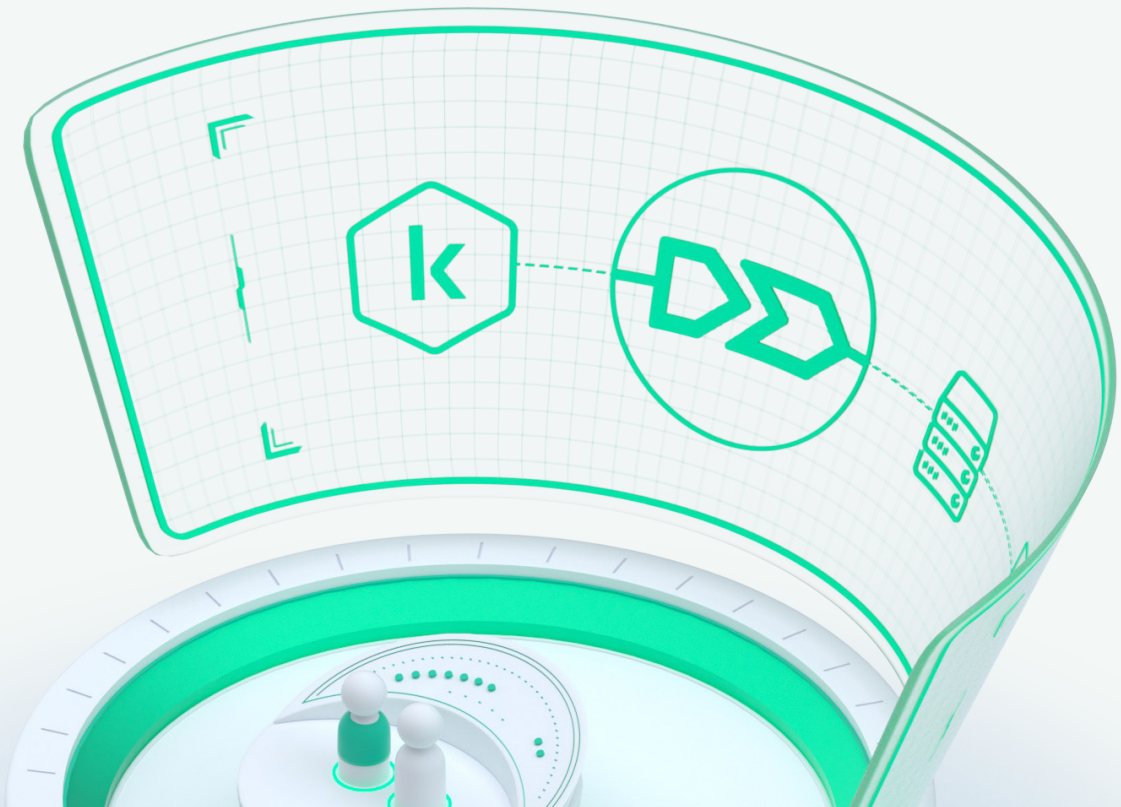
Благодарности

Данный документ был разработан «Лабораторией Касперского» и представлен **на семинаре № 31 «Кибербезопасность в сфере искусственного интеллекта: баланс рисков и инноваций»** на 19-м ежегодном Форуме по управлению интернетом **(Internet Governance Forum, IGF)**, который прошел 15–19 декабря 2024 года.

В работе над документом принимали участие:

- Юрий Шелихов, руководитель отдела кибербезопасности и защиты данных, «Лаборатория Касперского»
- Владислав Тушканов, руководитель группы исследования и разработки, группа исследования технологий машинного обучения, «Лаборатория Касперского»
- Алексей Антонов, руководитель направления исследования данных, группа анализа методов детектирования, «Лаборатория Касперского»
- Эллисон Уайлд (Allison Wylde), член команды по функциональной совместимости Сети по вопросам политики в области искусственного интеллекта Форума ООН по управлению Интернетом
- Д-р Мелодена Стивенс (Melodena Stephens), профессор кафедры управления инновациями и технологиями, Школа государственного управления имени Мохаммеда бин Рашида, ОАЭ
- Серхио Майо Масиас (Sergio Mayo Macías), менеджер инновационных программ, Технологический институт Арагона (ITA), Испания
- Игорь Кумагин, эксперт по кибербезопасности
- Дмитрий Фонарев, старший менеджер по связям с общественностью, «Лаборатория Касперского»

Содержание



Введение

Цель

Новые цифровые технологии создают **новые риски в сфере кибербезопасности, а также новые векторы атак**. Поэтому при внедрении технологий искусственного интеллекта компании должны позаботиться о защите от угроз. Безопасность при разработке систем ИИ стала центральным положением ряда нормативных инициатив, таких как Закон об ИИ Евросоюза или сингапурская Модель управления ИИ для генеративного искусственного интеллекта. Европейский Закон об ИИ устанавливает строгие нормативно-правовые требования, призванные обеспечить прозрачность, безопасность и этичное использование этой технологии. США больше сосредоточены на разработке отраслевых стандартов и поощрении инноваций, чем на строгом регулировании. Китай активно создает стандарты и нормы для развития ИИ-технологий, но при этом ограничивает их использование в некоторых сферах.

Несмотря на прогресс в области законотворчества, некоторые важные технические ограничения не позволяют в полной мере применить разрабатываемые стандарты на практике. В данной работе рассматриваются **базовые требования кибербезопасности**, которые должны учитываться при внедрении систем ИИ. Эти требования должны распространяться на **широкий ряд компаний, использующих сторонние ИИ-компоненты** для разработки собственных решений.

Для безопасного внедрения ИИ организациям требуется техническое руководство по разработке и развертыванию соответствующих технологий. Внедрение ИИ без надлежащих инструкций может повлечь за собой существенные риски. Данный документ предлагает рекомендации для разработчиков и администраторов систем ИИ, MLOps и AI DevOps, использующих уже существующие базовые модели для создания ИИ-решений общего назначения, в особенности облачных. Здесь представлены **ключевые аспекты разработки, внедрения систем ИИ, а также управления ими**, включая вопросы проектирования, интеграции и передового опыта в области безопасности, — без углубленного рассмотрения разработки базовых моделей.

Искусственный интеллект (ИИ) стал не только частью повседневной жизни, но и одной из важнейших для мировой экономики технологий. ИИ позволяет автоматизировать рутинные задачи, повысить качество обслуживания клиентов и обеспечить более быстрый и простой доступ к информации для сотрудников организаций.

> 50%

Недавнее исследование «Лаборатории Касперского» показало, что более 50% компаний уже применяют решения на основе ИИ*.

33%

планируют внедрить эту технологию в течение следующих двух лет.

* Статья «Более половины компаний используют ИИ и IoT в своих бизнес-процессах», https://www.kaspersky.com/about/press-releases/more-than-half-of-companies-use-ai-and-iot-in-their-business-processes

С популяризацией ИИ-систем среди организаций развиваются и сопутствующие угрозы. Кибератаки затрагивают все стадии разработки ИИ: от создания наборов данных и алгоритмов до получения результатов моделирования.

Обзор ландшафта угроз для систем ИИ

Согласно исследованию «Лаборатории Касперского»*, ИИ-системы подвержены специфичным угрозам, которые стремительно эволюционируют. Злоумышленники воздействуют на обучающие данные, манипулируют моделями, изменяя их поведение, и нарушают целостность системы. Это подчеркивает острую необходимость обеспечения комплексной безопасности при использовании ИИ.

Проблемы с разработкой моделей

В отличие от традиционного программирования, где код можно протестировать и получить однозначные результаты, модели машинного обучения, особенно модели глубокого обучения с миллионами или миллиардами параметров, сложны для изучения и зачастую функционируют как «черный ящик». Из-за этой сложности невозможно точно предсказывать и интерпретировать поведение моделей. Вследствие этого значительно возрастает риск пропуска ошибок, которые могут иметь серьезные последствия, например нарушить финансовую стабильность банка или даже угрожать жизни пациента.

Еще одной проблемой является то, что ИИ-модели иногда основывают свои решения на **нерелевантных или незначительных свойствах входных данных**, упуская при этом значимые характеристики. Например, модель распознавания изображений может научиться классифицировать животных как гепардов, леопардов или ягуаров, основываясь только на узоре пятен в их окрасе, но не принимая во внимание их анатомию в целом**. По такому принципу одна из моделей распознала диван как леопарда только потому, что обивка имела соответствующий рисунок. Подобные ошибки в классификации могут приводить к неверным результатам в критически важных сферах, таких как здравоохранение, образование, социальное обеспечение, транспорт, государственный сектор и тому подобные.

Несоответствие между данными, использованными для обучения модели, и реальными данными, с которыми ей придется работать, снижает качество результатов. К примеру, если модель обучается на данных с устройств одного вида, а анализировать будет данные с другого, то она может сфокусироваться на особенностях конкретных устройств, а не тех объектов, которые призвана распознавать. Такое несоответствие ведет к ошибочным предположениям и неточной классификации.

С этими проблемами можно столкнуться как при обучении моделей с нуля, так и при дообучении базовых моделей. Хотя наборы данных для дообучения меньше и проще в управлении, они тоже могут создавать ложные корреляции, из-за чего полученная модель не будет соответствовать поставленной цели.

* Документ «Искусственный интеллект под угрозой» (AI under Attack), https://content.kaspersky-labs.com/se/media/en/business-security/enterprise/machine-learning-cybersecurity-whitepaper.pdf

** Статья «Внезапно появляется диван с леопардовым принтом» (Suddenly, a leopard print sofa appears), https://web.archive.org/web/20200208171948/http://rocknrollnerd.github.io/ml/2015/05/27/leopard-sofa.html

Атаки на модели искусственного интеллекта

Злоумышленники могут атаковать модели искусственного интеллекта различными способами. Ниже представлены примеры того, как атакующие эксплуатируют уязвимости в механизмах проектирования, обучения и взаимодействия систем на базе ИИ.

Способ атаки	Описание
 Отравление данных: компрометация целостности модели	Отравление происходит посредством внедрения вредоносных данных в обучающий набор с целью повлиять на поведение модели. Тщательно подготавливая и добавляя отравленные образцы*, злоумышленники заставляют модель принимать ошибочные решения или неверно классифицировать определенные входные данные. Такой тип атак нарушает целостность модели и снижает ее надежность. Подобные атаки могут происходить в том числе при дообучении моделей.
 Adeversarial-атаки: невидимая манипуляция искусственным интеллектом	В ходе adversarial-атаки во входные данные вносятся небольшие изменения, незаметные для человека**, но заставляющие ИИ-модель давать неверные ответы. Злоумышленники добавляют к входным данным специально созданный шум, из-за чего модель выдает неверные результаты, в то время как для наблюдателя входные данные выглядят неизменными.
 Запоминание данных: риск непреднамеренного воздействия	Современные модели ИИ могут непроизвольно запоминать некоторые элементы из обучающих данных, особенно если те содержат какие-либо уникальные, исключительные образцы. Эти данные могут быть извлечены злоумышленниками, что потенциально ведет к утечке персональных данных или конфиденциальной деловой информации.
 Инъекции промптов: риск для больших языковых моделей	Инъекции промптов — это угроза, характерная для больших языковых моделей (LLM), таких как ChatGPT. Изначально разработчики программируют модели на выполнение задач, давая промпты (то есть подсказки) на естественном языке. Пользователи тоже взаимодействуют с моделью при помощи естественного языка, и у модели нет никакой возможности определить, от кого поступают инструкции. Пользуясь этой уязвимостью, злоумышленник может создать входные данные, которые изменят поведение модели, заставив ее выполнить непредусмотренные действия или раскрыть конфиденциальную информацию. Вредоносные промпты вводятся в модель напрямую или путем встраивания их в данные, предназначенные для обработки, к примеру документы или веб-страницы.

Это только наиболее актуальные тактики. Более полное описание всех возможных атак на системы ИИ выходит за рамки данного документа.

* Анализ атак с отравлением данных, https://bdtechtalks.com/2020/10/07/machine-learning-data-poisoning/
Статья «Атаки на машинное обучение: обзор» (Attacks against machine learning — an overview), https://elie.net/blog/ai/attacks-against-machine-learning-an-overview/

** «Объяснение и использование состязательного промптинга» (Explaining and Harnessing Adversarial Examples), https://arxiv.org/abs/1412.6572 «Состязательные атаки и защита от них в глубоком обучении» (Adversarial Attacks and Defenses in Deep Learning), https://arxiv.org/abs/2201.06192 Статья «Как запутать нейросети антивирусных программ: состязательные атаки и защита от них» (How to Confuse Antimalware Neural Networks: Adversarial Attacks and Protection), https://securelist.com/how-to-confuse-antimalware-neural-networks-adversarial-attacks-and-protection/102949/

Традиционные **уязвимости систем безопасности**

Традиционные недостатки систем безопасности свойственны и для моделей ИИ.

Уязвимости из сторонних ресурсов

Системы искусственного интеллекта часто опираются на модели сторонних разработчиков или на наборы данных, взятые из открытых источников. Подобные ресурсы могут содержать случайные ошибки или бэкдоры, преднамеренно внедренные злоумышленниками. Использование таких скомпрометированных компонентов создает уязвимости в системе ИИ, нарушая ее безопасность.

Риски для цепочки поставок при разработке ИИ

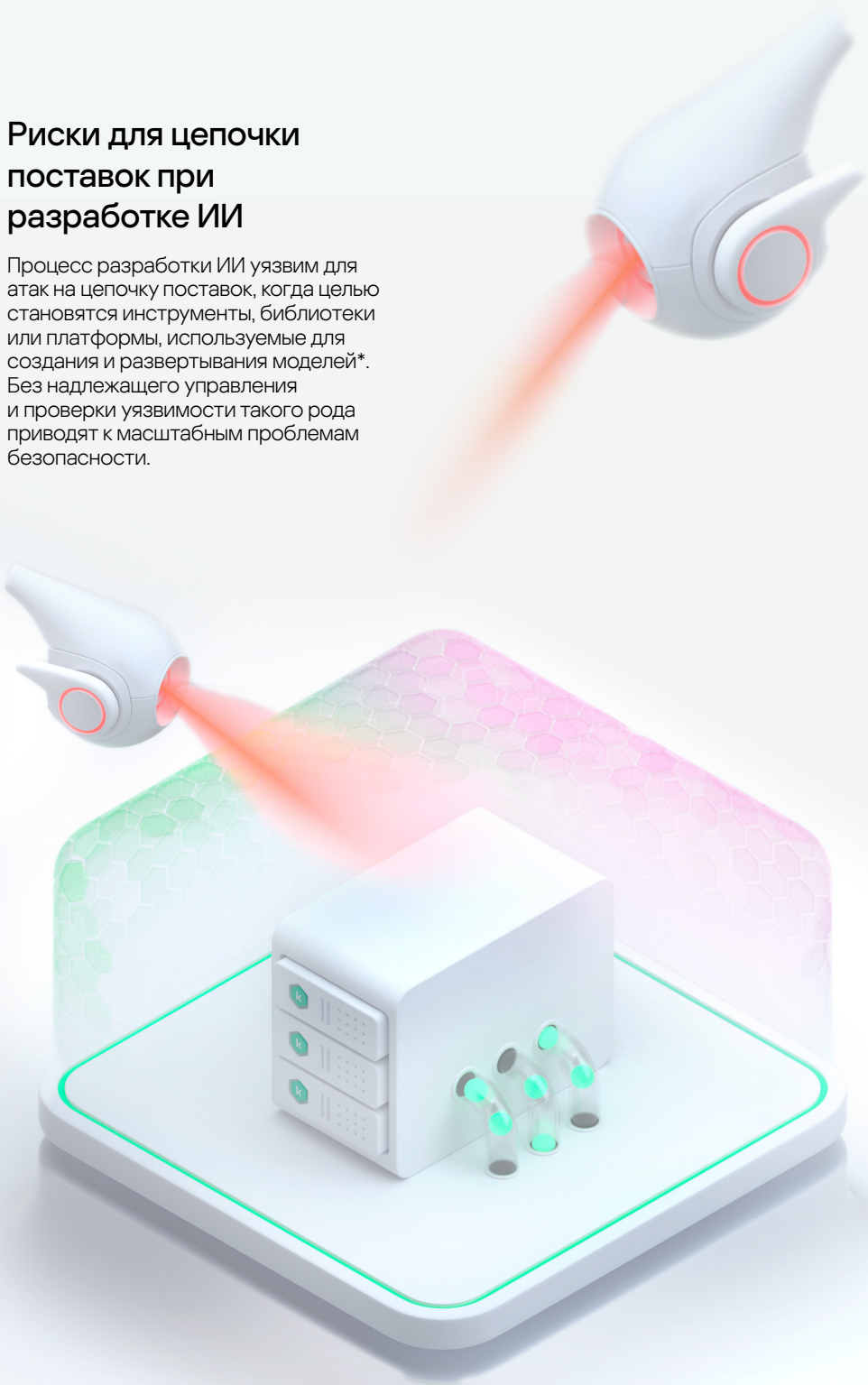
Процесс разработки ИИ уязвим для атак на цепочку поставок, когда целью становятся инструменты, библиотеки или платформы, используемые для создания и развертывания моделей*. Без надлежащего управления и проверки уязвимости такого рода приводят к масштабным проблемам безопасности.

Ошибки в коде, раскрывающие уязвимости ИИ

Ошибки в программном коде интерфейсов доступа к ИИ могут также привести к эксплуатации уязвимостей.

Риски кражи компонентов ИИ

Системы ИИ, а также их ключевые компоненты, такие как модели и наборы данных, без надлежащей защиты становятся уязвимыми к краже.



* Скомпрометирована цепочка зависимостей PyTorch-nightly в период с 25 по 30 декабря 2022 года, https://pytorch.org/blog/compromised-nightly-dependency

Рекомендации

Тренинги и повышение осведомленности о кибербезопасности

Для внедрения новых технологий, таких как ИИ, требуются **поддержка руководства, разработка внутренних политик и системы управления, а также повышение осведомленности сотрудников о сопутствующих рисках и угрозах**. Последнее имеет критическое значение, учитывая стремительное развитие ИИ. Многие доступные разработчикам ИИ-ресурсы не обладают достаточной ИБ-зрелостью или же не в полной мере придерживаются принципов security-by-default (безопасность по умолчанию) и security-by-design (безопасность, заложенная в архитектуру). В результате на разработчиков систем ИИ ложится дополнительная ответственность за устранение потенциальных рисков, о которых им следует сообщить.

Вот что может сделать компания для повышения осведомленности в дополнение к стандартным методам обеспечения безопасности.

1

Руководству организации необходимо повысить свою осведомленность о рисках безопасности, связанных с ИИ, и научиться ими управлять.

2

Политика безопасности компании должна быть обновлена с учетом специфических требований ИИ-сервисов; с обновлениями должны быть ознакомлены все сотрудники и подрядчики.

3

Политика должна описывать риски, связанные с разработкой и использованием ИИ-сервисов, а также местные законодательные ограничения на их применение.

4

Политика должна устанавливать роли и обязанности, связанные с использованием ИИ-сервисов внутри компании.

5

Необходимо разработать или приобрести корпоративный курс по безопасному использованию ИИ-сервисов и сделать его прохождение обязательным для всех сотрудников, в том числе новых. Программа курса должна затрагивать политику организации, существующие угрозы, примеры инцидентов, меры по защите от угроз, соответствующие законы и другие вопросы. При необходимости следует включить практические задания на основе сценариев. По окончании курса сотрудники должны проходить тестирование. Курс должен регулярно обновляться.

6

Существующие курсы по информационной безопасности должны быть дополнены новыми техниками злоумышленников, использующих ИИ-сервисы для атак на компании. К таким методам относятся генерация текста, клонирование голоса, фальсификация фотографий и видео.

7

Необходимо отслеживать законодательство в области безопасного использования ИИ и своевременно обновлять внутренние политики и программы тренингов.

Моделирование угроз помогает выявлять, понимать и снижать потенциальные риски безопасности на ранних этапах разработки системы ИИ.

Моделирование угроз и оценка рисков

Искусственный интеллект — это новая и активно развивающаяся технология, и риски в этой области тоже постоянно эволюционируют. Оценка рисков позволяет быть на шаг впереди опасностей, а с помощью моделирования угроз разработчики с небольшим опытом в создании систем ИИ могут лучше подготовиться к возможным трудностям. Моделирование также помогает выявлять и устранять уязвимые места в ИИ-системе до того, как их обнаружат злоумышленники.

Чтобы обеспечить эффективный процесс моделирования угроз, организации необходимо придерживаться **следующих рекомендаций**:

Выберите методологию оценки рисков (например, STRIDE, DREAD, LINDDUN, PASTA, TRIKE), а затем разработайте процедуры оценки риска и моделирования угроз для ИИ-сервисов. Методология оценки рисков должна включать схему определения уровней рисков, стратегии управления рисками в организации (включая допустимые пороги рисков), процедуры мониторинга и назначение ответственных для контроля за процессом.

При управлении рисками рекомендуется классифицировать угрозы по следующим категориям:

- угрозы, возникающие в результате неиспользования ИИ-сервисов;
- угрозы, возникающие в результате несоблюдения требований;
- угрозы, возникающие в результате неправильного использования ИИ-сервисов пользователями;
- угрозы для ИИ-моделей и обучающих наборов данных;
- угрозы для услуг, создаваемые ИИ-моделями;
- угрозы для данных;
- угрозы экологического, социального и управленческого характера (ESG).

Выполняйте оценку рисков для всех существующих и внедряемых ИИ-сервисов.

Включайте в оценку не только выявленные угрозы и риски, но и потенциальные источники опасности и злоумышленников. Для идентификации известных угроз и рисков следует использовать такие справочные материалы, как NIST-AI-600-1, MITRE ATLAS, «OWASP: 10 главных угроз для LLM-приложений», DASF и CSA.

Передавайте информацию об обнаруженных рисках руководству компании.

Защита инфраструктуры (облако)

Чаще всего ИИ-сервисы предоставляются в виде облачных служб и нередко требуют специализированной инфраструктуры, например серверов, оснащенных графическими процессорами, программируемыми логическими матрицами (FPGA), специализированными интегральными схемами (ASIC) или тензорными процессорами (TPU). Учитывая чувствительность ИИ-систем, им необходима защита в соответствии с наиболее продвинутыми стандартами в области кибербезопасности, например NIST Cybersecurity Framework. Как правило, ИИ-сервисы используют открытое или бесплатное программное обеспечение, такое как TensorFlow, PyTorch или Keras, а также такие библиотеки, как Pandas, NumPy и SciPy. Чтобы обеспечить безопасность такой среды, следует соблюдать **ряд требований**:

1

Идентифицируйте все информационные активы. Ведите учет наборов данных для обучения и тестирования моделей, данных для дообучения, баз данных, карточек данных, моделей и рисков, входящих и исходящих данных сервиса, весов и гиперпараметров модели, а также данных журналов LLM-систем.

2

Контролируйте доступ на всех уровнях, включая сеть, операционные системы, базы данных, программное обеспечение, данные и модели. Установите двухфакторную аутентификацию (2FA) для административного доступа.

3

Журналируйте все события и обеспечьте защиту данных журналов. Отслеживайте инциденты и возможные нарушения безопасности.

4

Установите защиту от вредоносного ПО и других типов атак. Регулярно устанавливайте исправления безопасности на компонентах инфраструктуры.

5

Сегментируйте сеть, чтобы защитить уязвимые области. Используйте шифрование данных — как передаваемых, так и хранимых.

6

Обеспечьте целостность критически важных данных, проверяйте подлинность используемых библиотек и моделей.

7

Обеспечьте резервирование для серверов и каналов коммуникации. Регулярно создавайте резервные копии данных и следите за способностью их корректного восстановления.

8

Храните ключи в Key Vault.

9

Применяйте принципы минимальных привилегий и «нулевого доверия» во всей инфраструктуре.

В зависимости от используемой инфраструктуры могут потребоваться дополнительные меры:

Используйте шлюз API для управления доступом к моделям и выполнения аутентификации.

Применяйте специальные меры безопасности для сред Kubernetes.

Следуйте передовым практикам обеспечения безопасности облачных сервисов.

Обеспечьте целостность исходного кода, обучающих данных, моделей и сценариев автоматизации.

Изолируйте обучающие данные, модели и среды обучения, чтобы предотвратить утечку или заражение.

Цепочки поставок и безопасность данных

Атаки на цепочки поставок представляют угрозу для инфраструктуры организаций, и системы искусственного интеллекта — не исключение. Известны случаи, когда целью подобных атак становились специализированные библиотеки для обучения нейросетей. Однако в случае с ИИ возникает особая проблема, связанная одновременно с безопасностью поставщика услуг и с защитой моделей машинного обучения.

Доступ к продвинутым моделям, большим языковым моделям в частности, во многом зависим от решений на базе облака. Тем не менее недоступность некоторых моделей в определенных регионах, а также другие ограничения могут заставить сотрудников компаний и разработчиков, использующих ИИ для повседневных задач, обращаться к сервисам-посредникам (прокси), которые перепродают доступ к моделям ИИ через API. Такая практика несет в себе значительные риски, от создания дополнительного источника утечки данных в случае инцидента безопасности в прокси-сервисе до неэтичного использования полученных данных для перепродажи или обучения собственных версий LLM. **Важно понимать эти риски** и тщательно изучить политику конфиденциальности как основного поставщика, так и прокси-сервиса, а также информировать всех сотрудников об опасностях использования сторонних сервисов для выполнения рабочих задач.

В целях обеспечения безопасности организация может развернуть языковую модель в локальной инфраструктуре. Так обрабатываемая моделью конфиденциальная информация при соблюдении определенных условий (сетевая изоляция, отключение телеметрии и т. д.) не выйдет за пределы компании. Однако помимо рисков, связанных с уязвимостями платформ машинного обучения, этот метод несет в себе еще и угрозу наличия бэкдоров в моделях. Дело в том, что форматы данных, используемые для распространения моделей машинного обучения, имеют разные уровни безопасности. Некоторые форматы позволяют встраивать произвольный код, который в дальнейшем выполняется при запуске моделей. Исследование показало, что некоторые (впрочем, немногочисленные) модели из общедоступных репозиториев могут выполнять произвольный код при запуске. При этом отсутствие доступа к оригинальным моделям из-за региональных или лицензионных ограничений вынуждает обращаться к сторонним поставщикам. Использование защищенных форматов, таких как safetensors, способно решить эту проблему. Однако необходимо повышать осведомленность разработчиков и аналитиков данных о важности выбора надежных источников для получения моделей и использования защищенных форматов для загрузки весов.

Убедитесь, что модели искусственного интеллекта получены из надежных, легитимных источников. Избегайте использования сторонних репозиториев.

Загружайте веса моделей в безопасных форматах, таких как safetensors, чтобы избежать риска выполнения произвольного кода.

Внедрите меры обнаружения и реагирования на атаки на цепочки поставок, связанные с компонентами ИИ.

Оценивайте и пересматривайте политики конфиденциальности сторонних сервисов, используемых для доступа к моделям ИИ, чтобы убедиться в их соответствии стандартам безопасности.

При локальном внедрении моделей ИИ создайте условия, обеспечивающие конфиденциальность данных: изолируйте сеть и отключите функции телеметрии.

Внедрите протоколы безопасного развертывания локальных моделей, чтобы минимизировать риски, связанные с возможным наличием бэкдоров в моделях машинного обучения.

Регулярно устанавливайте на платформы машинного обучения обновления и патчи, чтобы закрывать известные уязвимости.

Внедрите меры контроля за конфиденциальными данными, обрабатываемыми моделями ИИ, чтобы те не покидали инфраструктуру организации.

При использовании сторонних API проводите аудиты безопасности в соответствии с ведущими международными стандартами, такими как «OWASP: 10 главных угроз для API».

Тестирование и валидация

После проведения оценки и определения рисков необходимо понять, как избежать случайных или преднамеренных ошибок при обучении и применении модели. Для этого организации следует прибегнуть к следующим мерам:

1

Оцените потенциальный ущерб, который может быть нанесен случайными или преднамеренными ошибками в системе. Определите ценность данных, используемых для обучения модели, и данных, которые она обрабатывает.

2

Определите, используются ли для построения системы платформы, модели или наборы данных из открытых источников.

3

Определите потенциальных пользователей инструмента: сотрудники компании, клиенты, любые интернет-пользователи.

4

Убедитесь, что при построении модели применялись передовые практики в области машинного обучения. Проверьте, что наборы данных правильно разделены на обучающие, тестовые и валидационные в соответствии с принципами работы модели.

5

При валидации моделей ИИ и их метрик (числа ложноположительных и ложноотрицательных срабатываний) убедитесь, что разделение набора данных было выполнено правильным образом, например, для временных данных использовалось хронологическое разделение, а во время обучения данные были защищены от утечек.

6

Определите, какие функции модель использует для принятия решений и насколько они согласуются с мнением экспертов в области. Используйте методы интерпретации, такие как SHAP-векторы, чтобы понять процесс принятия решений моделью.

7

Оцените работу модели в реальных условиях, чтобы убедиться в ее способности выдавать ожидаемые результаты. Постоянно контролируйте модель, поскольку распределение входных данных может меняться с течением времени, потенциально ухудшая качество работы модели.

8

Адаптируйте план тестирования так, чтобы проверить, подвержена ли модель уязвимостям, характерным для моделей машинного обучения (например, adversarial-атакам и отравлению данных).

Отчетность об уязвимостях

Искусственный интеллект — сравнительно новая, стремительно развивающаяся технология. Она дает множество преимуществ, но при этом многие подобные системы подвержены специфическим уязвимостям, характерным именно для ИИ.

Одна из главных проблем заключается в потенциальной способности злоумышленников получить несанкционированный доступ к данным за счет эксплуатации уязвимостей систем. Еще один пример несовершенства ИИ — предвзятость. Она возникает в том случае, если модели обучались на нерепрезентативных, субъективных данных. Если в набор данных для алгоритма попали стереотипы, ошибочные социальные суждения либо он был неполным, система может подвергнуться влиянию предубеждений. В результате она будет принимать несправедливые, дискриминирующие решения, тем самым создавая негативный опыт для пользователей и подрывая доверие к технологии ИИ в целом.

Решением этой проблемы может быть внедрение механизма, позволяющего пользователям **сообщать об обнаруженных уязвимостях** и предубеждениях в системе. Механизм отчетности даст возможность быстро получать обратную связь и принимать необходимые меры.

1

Установите общедоступную политику выявления уязвимостей, описывающую механизм отчетности.

2

Предоставьте пользователям надежные способы отправки отчетов, например зашифрованные веб-формы или специальные адреса электронной почты.

3

Определите процедуры оперативной оценки, приоритизации и устранения обнаруженных уязвимостей.

4

Информируйте сообщившего пользователя о статусе проблемы и ее решении.

5

Информируйте пользователей об известных уязвимостях и мерах по их устранению для поддержания доверия и демонстрации ответственности.

6

Предлагайте исследователям в области безопасности вознаграждение за найденные ошибки (программа Bug Bounty). Это поможет вам оставаться в курсе новых угроз и передовых методов обеспечения безопасности ИИ.

Защита от атак, нацеленных на системы машинного обучения

Учитывая быстрый прогресс технологии ИИ, неудивительно, что уже существуют атаки, специфичные для систем машинного обучения. Эксплуатируя уязвимости, атакующие могут намеренно вводить в модель искаженные данные или скрытые команды. Организациям, использующим бесплатные платформы для разработки собственных систем, следует знать о существовании подобных рисков. Чтобы защититься от атак на системы машинного обучения, необходимо **придерживаться следующих рекомендаций по безопасности**:

Включение в обучающий набор данных примеров состязательных промптов*, чтобы научить модель эффективнее обрабатывать подобные входные данные.

Применение техник дистилляции для повышения устойчивости модели к вредоносным входным данным за счет упрощения процесса принятия решений.

Возможное использование монотонных моделей** для повышения стабильности и снижения восприимчивости к вредоносным входным данным.

Внедрение систем обнаружения вредоносных или аномальных входных данных в запросах пользователей: это позволит модели выявлять и отклонять подобные запросы до обработки данных.

Для защиты от отравления данных разработчики систем ИИ должны **проверять обучающие образцы** на наличие аномальных объектов и сравнивать производительность новых моделей с производительностью предыдущих версий, чтобы выявить резкие изменения в характеристиках.

Для защиты от инъекций промптов в LLM разработчики могут внедрить систему для анализа входящих запросов пользователей или других входных данных третьих сторон. Еще один подход — анализ откликов на подобные запросы с дальнейшей оценкой их соответствия текущей задаче системы.

* Статья «Как запутать нейросети антивирусных программ: adversarial-атаки и защита от них» (How to Confuse Antimalware Neural Networks: Adversarial Attacks and Protection), https://securelist.com/how-to-confuse-antimalware-neural-networks-adversarial-attacks-and-protection/102949/

** «Монотонные модели для обнаружения вредоносных программ в реальном времени» (Monotonic models for real-time dynamic malware detection), https://arxiv.org/pdf/1804.03643

Регулярные обновления безопасности и обслуживание

Сфера ИИ, особенно в части применения больших языковых моделей, считается достаточно молодой, и качество кода систем не всегда соответствует высоким стандартам. Поэтому многие платформы и инструменты, использующие машинное обучение, могут содержать **значительное количество уязвимостей**. К счастью, в настоящее время большинство популярных платформ активно дорабатываются до производственного качества: регулярно выпускаются новые версии и обновления безопасности. Более того, распространение программ Bug Bounty, направленных на поиск уязвимых мест в инфраструктуре машинного обучения, способствует ускоренному решению проблем.

Все это еще раз подчеркивает важность постоянного мониторинга состояния инфраструктуры, начиная с платформ для отслеживания экспериментов и заканчивая библиотеками, предназначенными для взаимодействия с облачными сервисами. Однако поддержание актуальности инфраструктуры в этой сфере может занимать больше времени, чем в случае проектов в медленно развивающихся областях. Кроме того, использование последних версий библиотек может приводить к проблемам совместимости, что потребует еще больших вложений в разработку и поддержание функциональности кода, который опирается на эти библиотеки. Эти затраты необходимо учитывать при планировании инициатив в области ИИ.

Отдельный риск, связанный с использованием моделей на базе облачных сервисов (в частности, больших языковых моделей), — это **короткий жизненный цикл версий**. Поставщик платформы может довольно быстро обновить версию модели, выбранной для определенного проекта. Предполагается, что качество модели в общем должно улучшиться, однако ее поведение на конкретных задачах и возможность противостоять атакам могут измениться. Например, новая версия может быть подвержена инъекциям промптов или попыткам несанкционированного получения выходных данных через взлом. Поэтому для обеспечения плавного перехода между моделями без ухудшения эффективности и безопасности требуется тщательное планирование.

1. Убедитесь, что инфраструктура получила последние исправления безопасности и обновления платформы.
2. Организуйте программы Bug Bounty, используйте инструменты для обнаружения уязвимых мест в платформах машинного обучения и инфраструктуре ИИ.
3. Регулярно проверяйте и применяйте исправления безопасности для инструментов и библиотек машинного обучения, чтобы снизить риски, связанные с известными уязвимостями.
4. Помните о возможных проблемах совместимости при использовании последних версий библиотек и платформ — выделяйте ресурсы на разработку и тестирование.
5. Внедрите стратегию управления жизненным циклом облачных моделей ИИ, включая план по переходу на новые версии по мере их выпуска поставщиком.

Соответствие международным стандартам

Правовое регулирование в области искусственного интеллекта стремительно развивается, поэтому соблюдение соответствующих законов и следование передовым практикам становятся все важнее. При этом данные для обучения ИИ могут браться из разных источников, находящихся в различных юрисдикциях, что с правовой точки зрения усложняет обработку и использование таких данных. Кроме того, сами модели часто находятся в **открытых репозиториях**, что вносит значительную степень неопределенности в отношении их соответствия нормативным требованиям конкретной юрисдикции.

Таким образом, разработчики ИИ сталкиваются с непростой задачей: обеспечить соблюдение всех юридических требований в странах, где система будет использоваться. Наиболее надежная стратегия в данной ситуации — придерживаться **стандартов лидеров в этой области**, в частности Китая, Европейского союза или США. Многие страны уже делятся своими подходами и внедряют подобные требования, что дает разработчикам возможность заранее подготовиться к внедрению систем ИИ по всему миру.

1

Разработайте руководства по этичным использованию и разработке ИИ, чтобы обеспечить прозрачность и подотчетность в смежных процессах.

2

Убедитесь, что все собираемые данные соответствуют законам о защите данных в каждой юрисдикции, в частности — Общему регламенту по защите данных (GDPR) в Европе или Закону Калифорнии о защите персональных данных потребителей (CCPA) в США.

3

При использовании моделей ИИ из открытых репозиториев проверяйте их соответствие законам об интеллектуальной собственности.

4

Следуйте ведущим нормативным стандартам, таким как Закон об ИИ Европейского союза (AI Act) или Билль о правах в области ИИ США (AI Bill of Rights), — другие страны часто используют их в качестве ориентиров.

5

Следите за новыми и обновляющимися нормативными актами в области регуляции ИИ в разных странах.

6

Регулярно проверяйте модели и системы ИИ на соответствие международным стандартам, чтобы выявлять и снижать потенциальные юридические и этические риски.

Заключение

Как и большинство технологий, искусственный интеллект представляет не только большие возможности, но и значительные угрозы. Риски кибербезопасности, связанные с ИИ и его влиянием на общество, зависят от действий и намерений разработчика. Для безопасного внедрения ИИ организациям требуется техническое руководство по разработке и развертыванию соответствующих технологий в инфраструктуре. Осуществление этого процесса без надлежащих инструкций может представлять значительные риски. При этом критически важно придерживаться культуры безопасности и ответственности на всех этапах жизненного цикла ИИ и использовать инструменты контроля безопасности — от оценки рисков и тестирования систем до защиты цепочек поставок и регулярного обслуживания. Реализация описанных требований поможет **снизить риски**, связанные с внедрением систем ИИ в рабочие процессы компании.

Подробнее

www.kaspersky.ru

#kaspersky
#активируйбудущее