О работе AI Red Team

Д.Е. Намиот, Е.В. Зубарева

***Аннотация* — Быстрое распространение приложений машинного обучения, основанных на больших языковых моделях (ChatGPT и т.п.) привлекло внимание к известной проблеме систем машинного обучения – состязательным атакам. Такие атаки представляют собой специальные модификации данных на разных этапах стандартного конвейера машинного обучения (тренировка, тестирование, использование), которые призваны либо воспрепятствовать работе систем машинного обучения, либо добиться требуемого атакующему специального поведения таких систем. В последнем случае атакующий обычно желает добиться того, чтобы обученная модель специальным (нужным атакующему) способом реагировала на определенным образом подготовленные входные данные. Есть также классы атак на модели машинного обучения, которые специальным образом опрашивают работающие модели с целью получения скрытой информации, использованной при обучении модели. Все перечисленные атаки достаточно просто реализуются и для больших языковых моделей, что открыло глаза бизнес-сообществу на реально существующую проблему – кибербезопасность самих систем машинного обучения (искусственного интеллекта). Ответом стало ускоренное создание подразделений корпоративной кибербезопасности, которые должны тестировать системы искусственного интеллекта – AI Red Team. Принципы построения и работы таких команд и рассматриваются в данной статье.**

***Ключевые слова*—искусственный интеллект, машинное обучение, кибербезопасность.**

I. Введение

Термин Red team (как и Blue team), применительно к кибербезопасности, достаточно давно является уже общеупотребимым и трактуется однозначно. Внутри компании (например, в подразделении кибербезопасности) есть группы сотрудников, которые разделены на две команды: Blue team и Red team. Задача Blue team — защищать ИТ-инфраструктуру компании и предотвращать инциденты информационной безопасности, задача Red team – наоборот, имитировать атаки на инфраструктуру, имитировать действия киберпреступников и находить уязвимости в системе защиты. Регулярные учения такого рода есть неотъемлемая часть работы любой серьезной службы кибербезопасности. Есть даже целый рынок услуг по оценке киберзащищенности, когда сторонние Red team команды тестируют защиту заказчиков[1].

Начиная с лета 2023 года можно наблюдать резкий рост публикаций о создании и работе так называемых AI Red Team. Внезапно оказалось, что такого рода подразделения созданы, успешно работают и даже выходят на первый план в кибербезопасности во многих больших компаниях. Например, Google [1], Microsoft [2], OpenAI [3] и т.д. Это команды Red team, которые атакуют приложения Искусственного интеллекта (AI – Artificial Intelligence).

На практике, Искусственный интеллект сегодня – это машинное обучение. Есть, конечно, так называемый сильный Искусственный интеллект (AGI), но до его практического применения еще относительно далеко, а то, что есть – это также машинное обучение. Например, фреймворк OpenAGI – это большие языковые модели (LLM) и обучение с подкреплением (RL) [4]. Соответственно, с практической точки зрения, AI Red team – это команды, которые атакуют системы машинного обучения.

Да, системы машинного обучения подвержены так называемым состязательным атакам [5]. Такие атаки представляют собой специальные модификации данных на всех этапах стандартного конвейера машинного обучения, которые призваны либо воспрепятствовать работе систем машинного обучения, либо добиться требуемого атакующему специального поведения таких систем (определенной реакции на специально подготовленные данные) [6]. Есть еще формы атак, когда атакующий специальным образом формулирует запросы к моделям (это касается приложений MLaaS - Machine learning as a service), чтобы получить какую-то дополнительную информацию (например, о тренировочных наборах данных и т.п.).

Именно состязательные атаки, связанные с модификацией данных, являются основной проблемой для применения машинного обучения в критических приложениях (авионика, автоматическое вождение и т.д.) [7]. Естественно, что с появлением атак на системы машинного обучения стали разрабатываться и методы защиты. Это касалось как модификации тренировочных данных (атаки отравления [8]), так и модификации входных данных на этапе исполнения (атаки уклонения [9]). Но пока в этом соревновании атак и защит

Статья получена 15 сентября 2023.
Д.Е. Намиот – д.т.н., в.н.с. МГУ имени М.В. Ломоносова (e-mail: dnamiot@gmail.com).
Е.В. Зубарева – к.п.н., с.н.с. МГУ имени М.В. Ломоносова (email: e.zubareva@cs.msu.ru).

[1] Цветовая дифференциация прижилась. Появились также White team, Purple team https://en.wikipedia.org/wiki/Red_team#Terminology.

преимущество на стороне атакующих. Состязательная атака – это модификация данных. Модифицированные данные – это такие же данные, как и исходные. Поэтому нет никакого универсального “антивируса”, который определял бы все возможные атаки. С другой стороны, изменять данные проще, чем изменять код. Это значит, что и состязательных атак будет больше – их проще организовать. Самым большим сдерживающим моментом для атак является, на самом деле, практическая возможность осуществления атаки. Например, для модификации данных на этапе тренировки требуется получить доступ к этому тренировочному набору. Например, если разметку данных отдали на аутсорсинг, или разработчики загрузили уже отравленный набор данных и т.д. В конкретном случае такого вполне может не быть.

По нашему мнению, есть несколько моментов, которые привели к взрывному интересу к направлению AI Red team. Во-первых, это, безусловно, подтверждение того факта, что приложения машинного обучения становятся ядром бизнеса компаний. Реальное подтверждение широкого распространения систем искусственного интеллекта и, соответственно, необходимость перейти к практическим аспектам безопасности систем машинного обучения. Из академической дисциплины это становится абсолютной практикой. Отметим, что машинное обучение широко используется и в кибербезопасности, так что атаки на системы машинного обучения – это, в том числе, и атаки на системы киберзащиты.

Следующий момент, сильно повлиявший на быстрый переход к практике – это большие языковые модели (LLM). Именно ChatGPT и последовавшие за ним аналогичные модели стали распространяться быстрее всех других приложений искусственного интеллекта. А ведь эти системы также подвержены атакам. LLM могут обучаться на специально подготовленных документах (атаки отравления), можно специальным образом готовить запросы (prompt engineering, который будет выступать аналогом атак уклонения). Плюс, такого рода системы могут использоваться для подготовки вредоносного контента [10]. Соответственно, LLM простимулировали интерес к безопасности систем ИИ.

Ну и в качестве последней причины можно назвать явный интерес к регулированию систем ИИ. При этом не только государства, но и руководители крупнейших компаний (OpenAI, Microsoft и Google) публично высказываются за регулирование ИИ и проводят встречи с мировыми лидерами. Генеральный директор OpenAI (ChatGPT) Сэм Альтман отправился в мировое турне, чтобы выразить поддержку новым законам, включая предстоящий Закон Европейского Союза об искусственном интеллекте. Компания OpenAI выделяет гранты разработку сред управления ИИ [11]. Президент Microsoft повторил призывы OpenAI к агентству США по регулированию ИИ. Генеральный директор Google Сундар Пичаи согласился сотрудничать с европейскими законодателями для разработки «пакта об ИИ» — набора добровольных правил, которым разработчики должны следовать до вступления в силу правил ЕС. Несомненно, этот процесс ускорился именно из-за успехов больших языковых моделей.

В данной статье мы хотим описать текущее состояние дел по формированию и работе AI Red Team. Статья является частью серии публикаций, написанных для поддержки магистерской программы факультета ВМК МГУ имени М.В. Ломоносова по кибербезопасности совместно со Сбербанк [43].

Оставшаяся часть статьи структурирована следующим образом. В разделе II рассматриваются задачи AI Red Team. Раздел III посвящен фреймоворку Google SAIF. В разделе IV рассматриваются другие доступные фреймворки и программные инструменты для AI Red Team. И раздел V содержит заключение.

II. Задачи AI Red Team

В статье [12] (взгляд венчурного инвестора) отмечается, что решения на основе машинного обучения отличаются от традиционного программного обеспечения, в первую очередь, недетерминированностью. Соответственно, главными задачами директора по кибербезопасности становятся прозрачность, управление и аудит для таких приложений. Возможные проблемы безопасности предлагается разделить на статические (проблемы архитектуры) и динамические (проблемы использования). Автор рассматривает безопасность систем ИИ по аналогии с безопасностью облачных систем, подчеркивая, что необходимо время, чтобы выработать подходы к безопасности. Безопасность систем ИИ сейчас там, где была безопасность облачных систем в самом начале пути. Возможные проблемы с безопасностью конфигурации или системы исполнения – это и есть направления поиска уязвимостей для AI Red Team. А оценка рисков касательно указанных проблем – это аудит систем Искусственного интеллекта (машинного обучения).

С точки зрения архитектуры (конфигурации) системы предлагается рассматривать следующее:

1. Происхождение данных и политики их использования. Организациям необходимо убедиться, что они (или их поставщики ИИ) соблюдают нормативные требования к политике хранения, хранения и использования данных сообразно имеющимся регуляторным требованиям. На самом базовом уровне им необходимо знать, откуда взялись данные, и являются ли они конфиденциальными или предвзятыми. Кроме того, организации должны знать, представляют ли данные юридический риск из-за авторских прав или лицензионных соглашений с открытым исходным кодом.

Заметим от себя, что любая загрузка сторонних датасетов - это риск получить отравленные данные. Соответственно, нужны процедуры проверки (очистки) загружаемых данных. Ну а для исследователей уязвимостей – это поиск возможности отравления данных.

2. Идентификаторы агентов ИИ. Имеются в виду автономные программные компоненты с системами искусственного интеллекта - по мере того, как они взаимодействуют с информационными системами предприятия и, потенциально, получают разрешения на изменение состояния, они становятся новой областью управления идентификацией. Организации должны знать степень доступа и функций этих программных агентов.

Здесь можно провести параллель с программными роботами (RPA), используемыми для автоматизации [13]. Аналогичные роботы с искусственным интеллектом не будут, очевидно, столь предсказуемыми.

3. Реестры моделей/уязвимости наборов инструментов и риск цепочки поставок: необходимы инструменты для поиска уязвимостей в коде моделей с открытым кодом и в наборах инструментов (например, уязвимость PyTorch [14]). Сторонние зависимости в вышестоящих службах представляют собой еще один фактор риска, Например, ошибка в кеше Redis затронула ChatGPT [15].

Примечание. Загружаемые модели могут содержать вредоносный код, например, в весах [16]. Стандартные фреймворки могут быть компрометированы: так называемая "слепая атака" – модификация функции вычисления потерь затрагивает все модели, запускаемые на такой версии фреймворка [8]. Соответственно, при поиске уязвимостей AI Red Team – это возможные поверхности атаки - рис.1.

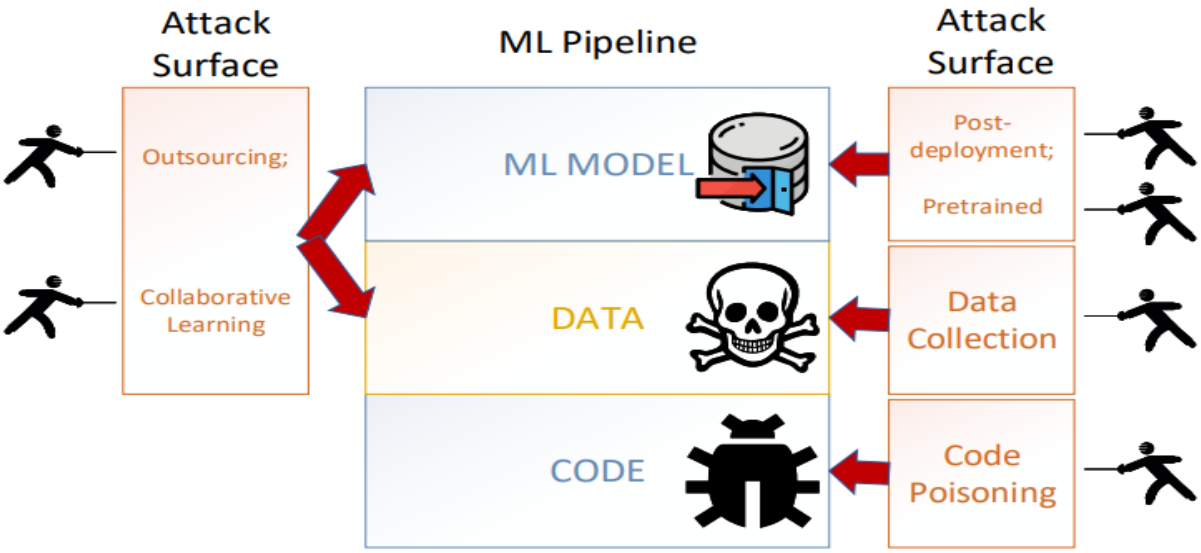

Рис. 1. Поверхности атак отравления [17]

4. MLSecOps и тестирование моделей. Оснащение конвейеров разработки ИИ кодом позволит разработчикам как искать уязвимости, так и создавать спецификацию модели для отслеживания и проверки. В этот процесс также можно интегрировать стресс-тесты для оценки устойчивости моделей к различным типам данных и типичным атакам на стадии подготовки к эксплуатации [18, 19] – рис. 2.

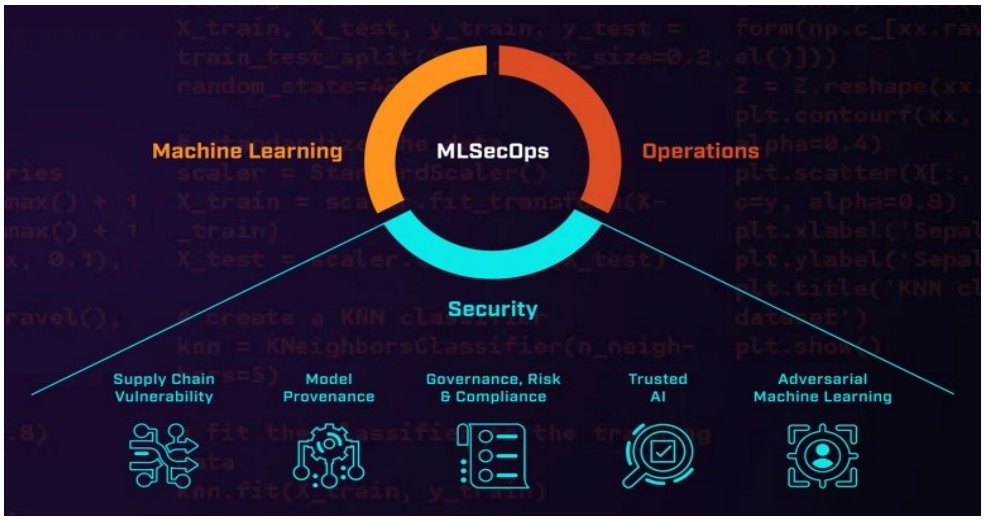

Рис. 2. MLSecOps [19]

Примечание. Здесь автор работы [12], кажется, несколько оптимистичен. Проверять модели на тренировочном наборе данных, конечно, можно. Но, в общем случае, это не спасет, например, от возможных сдвигов данных при работе на генеральной совокупности данных [20]. Такой сдвиг, в частности, может свидетельствовать о полной потере связи между зависимыми и независимыми переменными (сдвиг концепции), а перетренировка, очевидно, не очень подходит для моделей, работающих 24/7 в критических приложениях. Ключевым элементом на рис. 2 является Trusted AI, что соответствует так называемым доверенным платформам машинного обучения [21]. Это тема отдельного рассмотрения.

5. Конфигурация модели. Существует вероятность возникновения рисков в настройках и конфигурациях по мере того, как модели становятся более сложными и продуктивными, что похоже на неправильную настройку облачных сервисов.

Говоря о проблемах периода исполнения (вывода), можно отметить следующее

1. Атаки уклонения. Это касается и больших языковых моделей – там есть prompt injection [22].

2. Отравление данных: необходимо ввести меры контроля против ввода злоумышленниками вредоносных, низкокачественных или неподдерживаемых данных с целью использования модели или снижения ее производительности.
Примечание. Разведочный анализ и очистка данных должны быть частью доверенных платформ машинного обучения [21]. Но проверять (защищать) нужно и скачиваемые предобученные модели [23].

3. Кража модели: необходимы меры безопасности, чтобы не дать злоумышленникам получить веса модели или воспроизвести границу решения или поведение запатентованной модели.
Примечание. Как только модель машинного обучения открывается для внешних запросов, появляется возможность атак, связанных с кражей интеллектуальной собственности или неавторизованного доступа к данным модели [24].

Поскольку, как мы писали выше, развитие темы было стимулировано внедрением LLM, следующие пункты касаются именно больших языковых моделей.

4. Оценка ответа LLM. Отсутствие оценочного контроля неправильных или токсичных ответов представляет собой риск и потенциальный репутационный ущерб для компаний. Личная информация и отказ от авторских прав являются еще одним юридическим риском. Опять же, в основе проблемы оценки ответа лежит переход от детерминированных результатов программного обеспечения к стохастическим.

5. Помимо упомянутых выше prompt injection attack [25], существую еще и так называемые непрямые запросы (Indirect prompt injection attacks). Даже если подсказка сама по себе хорошо построена и не является вредоносной по своей сути, системы с возможностями агента и доступными плагинами представляют дополнительный риск. Например, приглашение может вызывать веб-страницу со скрытыми в ней комментариями, предназначенными для замены исходного приглашения пользователя, или внедрения полезных данных, которые создают скомпрометированное соединение со злонамеренной третьей стороной. Цепочка плагинов также может привести к эскалации разрешений и компрометации [26].

III Google SAIF

Компания Google предложила SAIF (Security AI Framework) [27], который определяет 6 основных задач в области обеспечения безопасности систем AI.

1. Распространите строгую систему безопасности на компоненты (подсистемы) Искусственного интеллекта
2. Расширьте возможности обнаружения и реагирования, чтобы они охватывали и угрозы системам искусственного интеллекта
3. Автоматизируйте защиту, чтобы идти в ногу с существующими и появляющимися угрозами.
4. Гармонизируйте элементы управления на уровне платформы, чтобы обеспечить единообразную безопасность во всей организации.
5. Адаптируйте элементы управления для корректировки мер по снижению рисков и создания более быстрых циклов обратной связи для развертывания ИИ.
6. Оцените риски системы ИИ в окружающих бизнес-процессах

Соображения по реализации SAIF изложены в работе [28]. Если следовать комментариям из работы [29], то можно, например, отметить следующее.

Для пунктов 1 и 2 хорошим источником является атлас атак MITRE [30] – это как раз ответ на вопрос что защищать в ИИ. Другой источник знаний – это рассматриваемый ниже GAIA top 10.

Для пункта 3 в качестве источника информации можно указать свежий документ от того же MITRE - ATLAS mitigations [31]. Базой автоматизации является мониторинг работы конвейера машинного обучения для отслеживания угроз и аномалий.

Риски, которые упоминаются в пункте 5, проистекают из модели угроз для конвейера машинного обучения. Их снижение – это рассмотрение следующих вопросов:

- Безопасность данных. Данные, используемые для обучения и развертывания моделей, должны быть надлежащим образом защищены.
- Безопасность модели. Модели ИИ должны быть надлежащим образом защищены с помощью таких мер, как проверка входных данных, очистка выходных данных и мониторинг работы моделей.
- Безопасность окружения. Среда, в которой развертываются модели ИИ, должна быть должным образом защищена с использованием таких мер, как безопасность и проверка программного обеспечения, сегментация сети и контроль доступа.

Есть вариант от Google оценки угроз системам ИИ: GAIA (Good AI Assessment - свой вариант OWASP top 10) [29]. Это список наиболее распространенных атак и слабых мест ИИ, которые злоумышленники могут использовать против концептуального конвейера ИИ (базовый конвейер: тренировка – тестирование – эксплуатация) или имеющейся генеративной модели. В содержании этих пунктов также очевидно влияние LLM.

G01 – Оперативная инъекция
Здесь злоумышленник попытается ввести неверные данные или информацию в командную строку, чтобы заставить вашу модель делать то, что вы не хотите, например, попытаться получить доступ к базовой операционной системе или заставить ее выдавать смущающие результаты, которые могут поделиться в социальных сетях.

G02 – Раскрытие конфиденциальных данных
Здесь злоумышленник может получить доступ к конфиденциальным данным из-за недостаточной обработки обучающих данных или из-за того, что злоумышленник получает доступ к базовому технологическому стеку.

G03 — Нарушение целостности данных
Здесь злоумышленник может внедрить состязательные данные в модель или базу данных внедрений после того, как злоумышленник получит доступ к базовому технологическому стеку.

G04 — Плохой контроль доступа
В этом случае базовый технологический стек имеет недостаточный контроль доступа, и злоумышленник может загрузить модель. Другой вариант - API-интерфейсы не были разработаны с учетом контроля доступа.

G05 – Недостаточная фильтрация подсказок и галлюцинаций (*последний термин для LLM описывает ситуацию, когда ответ сильно выходит за рамки вопроса, предвзят или просто неверен*).
Здесь фильтры подсказок не были должным образом проверены или не сопоставлены со случаями злоупотреблений, или галлюцинации общих данных не были должным образом проверены или не сопоставлены со случаями злоупотреблений.

G06 – Чрезмерный доступ для агента
Это когда публичный программный агент имеет доступ к частным/ограниченным внутренним API, или публичный агент имеет доступ к частным/ограниченным моделям, или агент имеет доступ к финансовым системам.

G07 – Атаки на цепочку поставок
Подобно стекам технологий разработки программного обеспечения, стеки технологий искусственного интеллекта полагаются на множество сторонних библиотек (особенно библиотек Python). Если вы используете библиотеки с открытым исходным кодом, эти библиотеки могут быть скомпрометированы злонамеренными третьими лицами. Кроме того, сторонние репозитории моделей ИИ могли быть

скомпрометированы. Стоит отметить, что сама модель, если она построена с использованием Python, может иметь конфигурацию по умолчанию, состоящую из смеси кода и данных, и потенциально может запускать код злоумышленника при установке (*это то, что называется отравление моделей*).

G08 – Атаки типа «отказ в обслуживании»
В этом случае регулирование или ограничение скорости отсутствует, или балансировка нагрузки недостаточна.

G09 – Недостаточное журналирование
Как и в случае со стандартными стеками технологий, существуют различные точки, в которых полезные данные журналов могут быть собраны и отправлены в централизованную систему SIEM, которая может помочь защитникам идентифицировать продолжающуюся атаку. Ведение журнала часто является второстепенным вопросом для конвейеров ИИ.

G10 – Небезопасное публичное развертывание
Примерами случаев небезопасного публичного развертывания могут быть модели, развернутые непосредственно на незащищенном сервере вывода или доступные для прямой загрузки. Кроме того, API вывода или веб-служба уязвимы, не исправлены и не обновлены, а также существуют учетные записи с чрезмерными разрешениями.

На рис. 3 показаны части конвейера ИИ, которые могут быть уязвимы для топ-10 GAIA.

Организациям следует учитывать свой технологический стек вместе с Топ-10 GAIA, когда задается вопрос «как нам это обеспечить» Учитывая, что конвейер ИИ построен на известных технологиях, и, следовательно, любые принимаемые меры по смягчению последствий атак представляют собой, в основном, модификации существующих средства контроля безопасности.

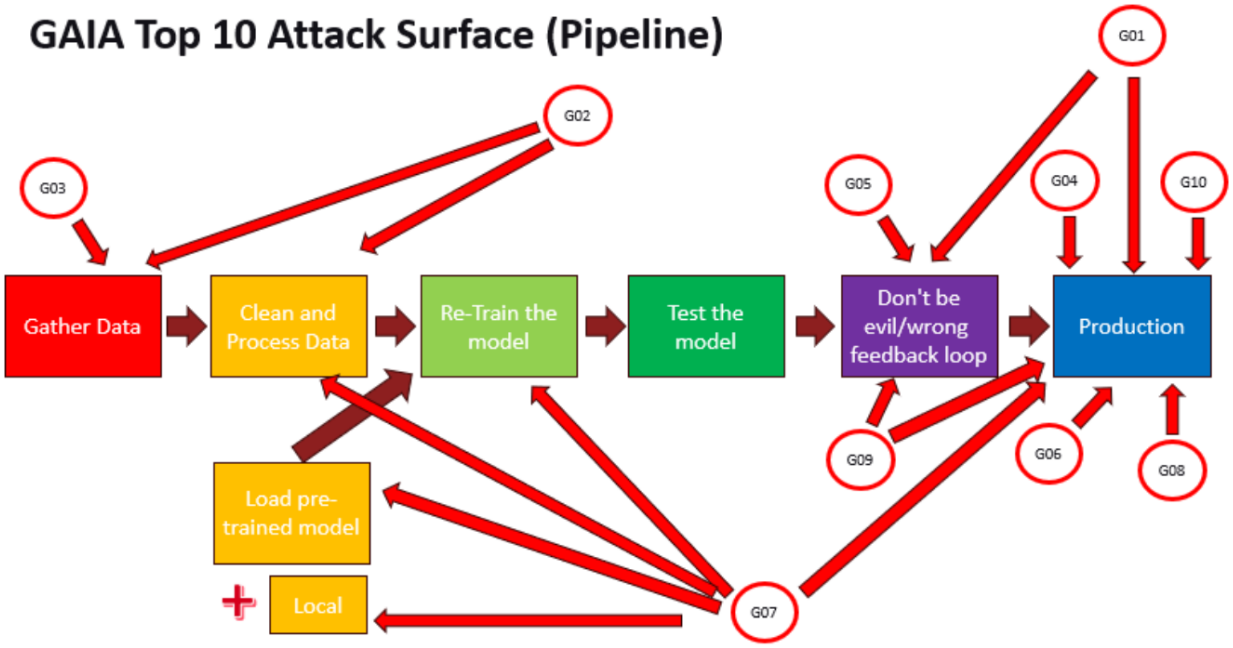

Рис. 3. GAIA top 10 атаки [29]

IV Другие фреймворки и инструменты AI Red Team

Из других фреймворков для AI Red Team можно отметить инструменты Microsoft для LLM [32] (часть сервиса Azure OpenAI). OpenAI имеет группу около 50 человек для GPT Red Team [33].

Microsoft (напомним, что именно эта компания стоит за созданием Атласа MITRE по состязательным атакам), в свой статье о AI Red Team [34] приводит ссылки на ряд полезных инструментов.

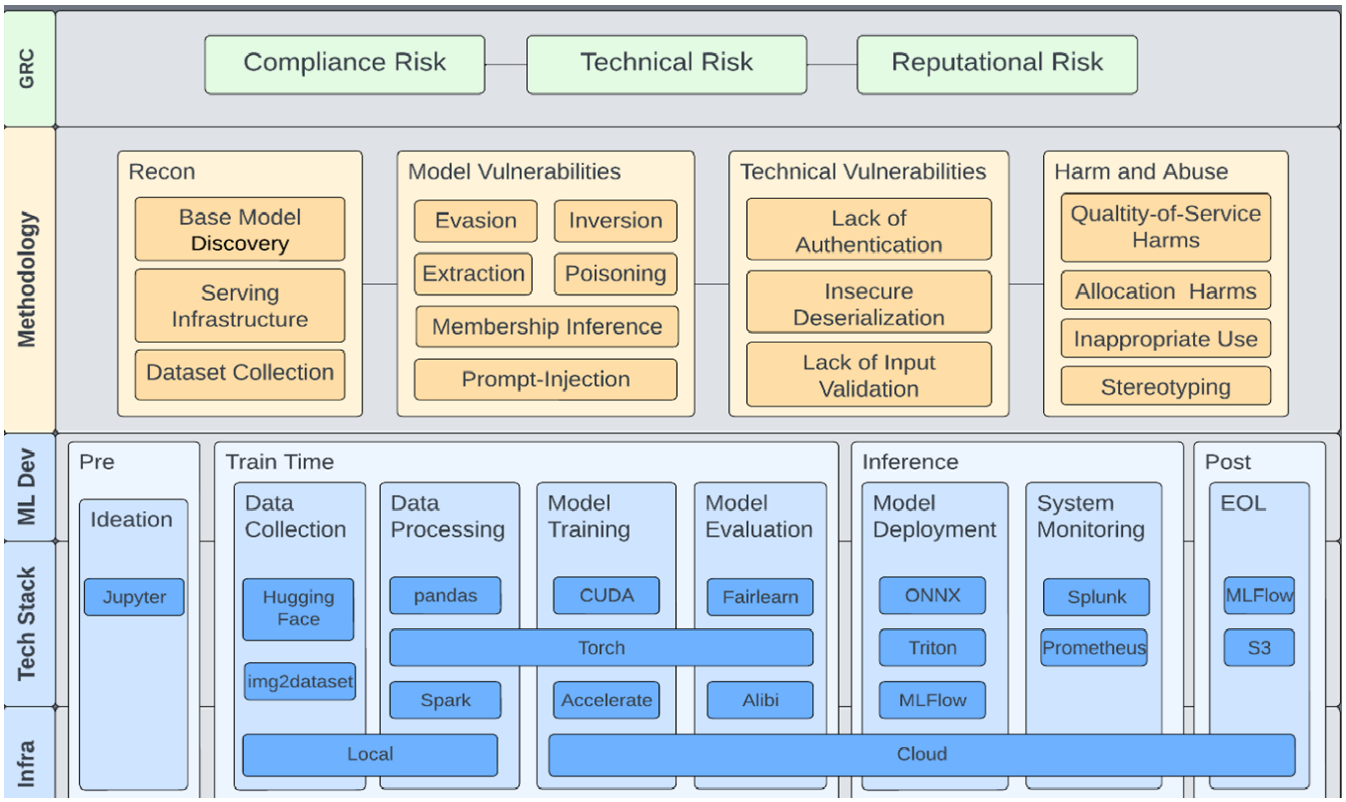

Рис. 4. Что оценивает AI Red Team [40].

Например, шкала ошибок жизненного цикла разработки систем ИИ [35], подход к оценке рисков ИИ [36], материалы по моделированию угроз для систем ИИ [37], портал, посвященный ответственному ИИ [38], таксономия по умышленным и непреднамеренным сбоям систем ИИ [39].

NVIDIA [40] достаточно подробно расписала задачи и работу своей AI Red Team. На рисунке 4 представлен фреймворк, который используется для оценки в Red Team.

На верхнем уровне основные задачи Red Team описываются англоязычной аббревиатурой GRC (governance, risk, and compliance - управление, риски и соответствие требованиям). GRC является высшим уровнем усилий по обеспечению информационной безопасности, обеспечивая перечисление, передачу и реализацию требований бизнес-безопасности. К рискам высокого уровня относятся:

- Технический риск: системы или процессы ML подвергаются риску в результате технической уязвимости или недостатка.
- Репутационный риск: производительность или поведение модели плохо отражаются на организации. В этой новой парадигме это может включать в себя выпуск модели, которая будет иметь широкое социальное воздействие.
- Риск несоблюдения требований: система ML не соответствует требованиям, что приводит к штрафам или снижению конкурентоспособности.

Эти категории риска высокого уровня присутствуют во всех информационных системах, включая системы ML.

V Заключение

Тема AI Red Team, стремительно возникшая летом 2023 года, несомненно, получит свое развитие. Бурный рост этой темы является свидетельством того, что направление безопасности систем Искусственного Интеллекта (машинного обучения) перешло в практическую область. Это свидетельство того, что системы машинного обучения массово внедряются в бизнес-процессы компаний и, соответственно, становятся интересными объектами для атак.

Большую роль в столь стремительном развитии этого направления сыграло успешное распространение ChatGPT и других языковых моделей. Это реально самый демократичный и понятный на сегодня практический инструмент искусственного интеллекта, который оказался открытым для атак.

Как следующий этап развития этого направления следует, очевидно, ожидать появления (развития) средств автоматизации. Состязательные атаки, в принципе, строятся с помощью систем машинного обучения, так что автоматизация здесь выглядит вполне естественно. Как существующие инструменты автоматизации работы AI Red team можно указать, например, Microsoft Counterfit [41] и GARD project [42].

Благодарности

Авторы благодарны сотрудникам кафедры Информационной безопасности факультета ВМК МГУ имени М.В. Ломоносова за ценные обсуждения.

Исследование выполнено при поддержке Междисциплинарной научно-образовательной школы Московского университета «Мозг, когнитивные системы, искусственный интеллект».

Библиография

[1] Google's AI Red Team: the ethical hackers making AI safer https://blog.google/technology/safety-security/googles-ai-red-team-the-ethical-hackers-making-ai-safer/ Retrieved: 07.09.2023.
[2] Microsoft AI Red Team building future of safer AI https://www.microsoft.com/en-us/security/blog/2023/08/07/microsoft-ai-red-team-building-future-of-safer-ai/ Retrieved: 07.09.2023.
[3] OpenAI's red team: the experts hired to 'break' ChatGPT https://archive.is/xu0wS#selection-1437.0-1437.55 Retrieved: 07.09.2023.
[4] Ge, Yingqiang, et al. "Openagi: When llm meets domain experts." arXiv preprint arXiv:2304.04370 (2023).
[5] Ilyushin, Eugene, Dmitry Namiot, and Ivan Chizhov. "Attacks on machine learning systems-common problems and methods." International Journal of Open Information Technologies 10.3 (2022): 17-22. (in Russian)
[6] Namiot, Dmitry. "Schemes of attacks on machine learning models." International Journal of Open Information Technologies 11.5 (2023): 68-86. (in Russian)
[7] Namiot, Dmitry, and Eugene Ilyushin. "On the robustness and security of Artificial Intelligence systems." International Journal of Open Information Technologies 10.9 (2022): 126-134. (in Russian)
[8] Namiot, Dmitry. "Introduction to Data Poison Attacks on Machine Learning Models." International Journal of Open Information Technologies 11.3 (2023): 58-68. (in Russian)
[9] Kostyumov, Vasily. "A survey and systematization of evasion attacks in computer vision." International Journal of Open Information Technologies 10.10 (2022): 11-20.(in Russian)
[10] Mozes, Maximilian, et al. "Use of LLMs for Illicit Purposes: Threats, Prevention Measures, and Vulnerabilities." arXiv preprint arXiv:2308.12833 (2023).
[11] Democratic inputs to AI https://openai.com/blog/democratic-inputs-to-ai Retrieved: 08.09.2023
[12] Securing AI The Next Platform Opportunity in Cybersecurity https://greylock.com/greymatter/securing-ai/ Retrieved: 08.09.2023
[13] Namiot, Dmitry, et al. "Information robots in enterprise management systems." International Journal of Open Information Technologies 5.4 (2017): 12-21. (in Russian)
[14] Compromised PyTorch-nightly dependency chain between December 25th and December 30th, 2022 https://pytorch.org/blog/compromised-nightly-dependency/ Retrieved: 08.09.2023
[15] OpenAI Reveals Redis Bug Behind ChatGPT User Data Exposure https://thehackernews.com/2023/03/openai-reveals-redis-bug-behind-chatgpt.html Incident Retrieved: 08.09.2023
[16] Bidzhiev, Temirlan, and Dmitry Namiot. "Research of existing approaches to embedding malicious software in artificial neural networks." International Journal of Open Information Technologies 10.9 (2022): 21-31. (in Russian)
[17] Gao, Yansong, et al. "Backdoor attacks and countermeasures on deep learning: A comprehensive review." arXiv preprint arXiv:2007.10760 (2020).
[18] Kalin, Josh, David Noever, and Matthew Ciolino. "Color Teams for Machine Learning Development." arXiv preprint arXiv:2110.10601 (2021).
[19] MLSecOps https://mlsecops.com/ Retrieved: 08.09.2023
[20] Namiot, Dmitry, and Eugene Ilyushin. "Data shift monitoring in machine learning models." International Journal of Open Information Technologies 10.12 (2022): 84-93. (in Russian)
[21] Namiot, Dmitry, Eugene Ilyushin, and Oleg Pilipenko. "On Trusted AI Platforms." International Journal of Open Information Technologies 10.7 (2022): 119-127. (in Russian)
[22] Liu, Yi, et al. "Prompt Injection attack against LLM-integrated Applications." arXiv preprint arXiv:2306.05499 (2023).
[23] Wong, Sheng, et al. "MLGuard: Defend Your Machine Learning Model!." arXiv preprint arXiv:2309.01379 (2023).
[24] Song, Junzhe, and Dmitry Namiot. "A Survey of the Implementations of Model Inversion Attacks." International Conference on Distributed Computer and Communication Networks. Cham: Springer Nature Switzerland, 2022.
[25] Zou, Andy, et al. "Universal and transferable adversarial attacks on aligned language models." arXiv preprint arXiv:2307.15043 (2023).
[26] Compromising LLMs using Indirect Prompt Injection https://github.com/greshake/llm-security Retrieved: 08.09.2023
[27] Introducing Google's Secure AI Framework https://blog.google/technology/safety-security/introducing-googles-secure-ai-framework/ Retrieved: 08.09.2023
[28] Secure AI Framework Approach. A quick guide to implementing the Secure AI Framework (SAIF) https://services.google.com/fh/files/blogs/google_secure_ai_framework_approach.pdf Retrieved: 11.09.2023
[29] Securing AI Pipeline https://www.mandiant.com/resources/blog/securing-ai-pipeline Retrieved: 11.09.2023
[30] Atlas MITRE https://atlas.mitre.org/ Retrieved: 11.09.2023
[31] ATLAS mitigations https://atlas.mitre.org/mitigations/ Retrieved: 11.09.2023
[32] Introduction to red teaming large language models (LLMs) https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/red-teaming Retrieved: 11.09.2023
[33] OpenAI's red team: the experts hired to 'break' ChatGPT https://www.ft.com/content/0876687a-f8b7-4b39-b513-5fee942831e8 Retrieved: 11.09.2023
[34] Microsoft AI Red Team building future of safer AI https://www.microsoft.com/en-us/security/blog/2023/08/07/microsoft-ai-red-team-building-future-of-safer-ai/ Retrieved: 13.09.2023
[35] Bug-bar https://learn.microsoft.com/ru-ru/security/engineering/bug-bar-aiml Retrieved: 13.09.2023
[36] AI-Security-Risk-Assessment https://github.com/Azure/AI-Security-Risk-Assessment/blob/main/AI_Risk_Assessment_v4.1.4.pdf Retrieved: 13.09.2023
[37] AI threat modeling https://learn.microsoft.com/ru-ru/security/engineering/threat-modeling-aiml Retrieved: 13.09.2023
[38] Responsible https://www.microsoft.com/en-us/ai/responsible-ai AI Retrieved: 13.09.2023
[39] Failure modes taxonomy https://learn.microsoft.com/ru-ru/security/engineering/failure-modes-in-machine-learning Retrieved: 13.09.2023
[40] NVIDIA AI Red Team: An Introduction https://developer.nvidia.com/blog/nvidia-ai-red-team-an-introduction/ Retrieved: 13.09.2023
[41] Microsoft Counterfit https://github.com/Azure/counterfit/ Retrieved: 13.09.2023
[42] GARD project https://www.gardproject.org/ Retrieved: 13.09.2023
[43] Магистерская программа Кибербезопасность https://cyber.cs.msu.ru/ Retrieved: 13.09.2023

About AI Red Team

Dmitry Namiot, Elena Zubareva

***Abstract* — The proliferation of machine learning applications based on large language models (ChatGPT, etc.) has brought attention to a well-known problem in machine learning systems: adversarial attacks. Such attacks are special modifications of data at different stages of the standard machine learning pipeline (training, testing, use), which are designed to either prevent the operation of machine learning systems or achieve the special behavior of such systems required by the attacker. In the latter case, the attacker usually wants to ensure that the trained model reacts in a special way (needed by the attacker) to input data prepared in a certain way. There are also classes of attacks on machine learning models that specifically interrogate running models in order to obtain hidden information used in training the model. All of the above attacks can be implemented quite simply for large language models, which opened the eyes of the business community to a real problem - the cybersecurity of machine learning (artificial intelligence) systems themselves. The answer was the accelerated creation of corporate cybersecurity units that should test artificial intelligence systems - AI Red Teams. The principles of the construction and operation of such teams are discussed in this article.**

***Keywords*— artificial intelligence, machine learning, cybersecurity.**

REFERENCES

[1] Google's AI Red Team: the ethical hackers making AI safer https://blog.google/technology/safety-security/googles-ai-red-team-the-ethical-hackers-making-ai-safer/ Retrieved: 07.09.2023.
[2] Microsoft AI Red Team building future of safer AI https://www.microsoft.com/en-us/security/blog/2023/08/07/microsoft-ai-red-team-building-future-of-safer-ai/ Retrieved: 07.09.2023.
[3] OpenAI's red team: the experts hired to 'break' ChatGPT https://archive.is/xu0wS#selection-1437.0-1437.55 Retrieved: 07.09.2023.
[4] Ge, Yingqiang, et al. "Openagi: When llm meets domain experts." arXiv preprint arXiv:2304.04370 (2023).
[5] Ilyushin, Eugene, Dmitry Namiot, and Ivan Chizhov. "Attacks on machine learning systems-common problems and methods." International Journal of Open Information Technologies 10.3 (2022): 17-22. (in Russian)
[6] Namiot, Dmitry. "Schemes of attacks on machine learning models." International Journal of Open Information Technologies 11.5 (2023): 68-86. (in Russian)
[7] Namiot, Dmitry, and Eugene Ilyushin. "On the robustness and security of Artificial Intelligence systems." International Journal of Open Information Technologies 10.9 (2022): 126-134. (in Russian)
[8] Namiot, Dmitry. "Introduction to Data Poison Attacks on Machine Learning Models." International Journal of Open Information Technologies 11.3 (2023): 58-68. (in Russian)
[9] Kostyumov, Vasily. "A survey and systematization of evasion attacks in computer vision." International Journal of Open Information Technologies 10.10 (2022): 11-20.(in Russian)
[10] Mozes, Maximilian, et al. "Use of LLMs for Illicit Purposes: Threats, Prevention Measures, and Vulnerabilities." arXiv preprint arXiv:2308.12833 (2023).
[11] Democratic inputs to AI https://openai.com/blog/democratic-inputs-to-ai Retrieved: 08.09.2023
[12] Securing AI The Next Platform Opportunity in Cybersecurity https://greylock.com/greymatter/securing-ai/ Retrieved: 08.09.2023
[13] Namiot, Dmitry, et al. "Information robots in enterprise management systems." International Journal of Open Information Technologies 5.4 (2017): 12-21. (in Russian)
[14] Compromised PyTorch-nightly dependency chain between December 25th and December 30th, 2022 https://pytorch.org/blog/compromised-nightly-dependency/ Retrieved: 08.09.2023
[15] OpenAI Reveals Redis Bug Behind ChatGPT User Data Exposure https://thehackernews.com/2023/03/openai-reveals-redis-bug-behind-chatgpt.html Incident Retrieved: 08.09.2023
[16] Bidzhiev, Temirlan, and Dmitry Namiot. "Research of existing approaches to embedding malicious software in artificial neural networks." International Journal of Open Information Technologies 10.9 (2022): 21-31. (in Russian)
[17] Gao, Yansong, et al. "Backdoor attacks and countermeasures on deep learning: A comprehensive review." arXiv preprint arXiv:2007.10760 (2020).
[18] Kalin, Josh, David Noever, and Matthew Ciolino. "Color Teams for Machine Learning Development." arXiv preprint arXiv:2110.10601 (2021).
[19] MLSecOps https://mlsecops.com/ Retrieved: 08.09.2023
[20] Namiot, Dmitry, and Eugene Ilyushin. "Data shift monitoring in machine learning models." International Journal of Open Information Technologies 10.12 (2022): 84-93. (in Russian)
[21] Namiot, Dmitry, Eugene Ilyushin, and Oleg Pilipenko. "On Trusted AI Platforms." International Journal of Open Information Technologies 10.7 (2022): 119-127. (in Russian)
[22] Liu, Yi, et al. "Prompt Injection attack against LLM-integrated Applications." arXiv preprint arXiv:2306.05499 (2023).
[23] Wong, Sheng, et al. "MLGuard: Defend Your Machine Learning Model!." arXiv preprint arXiv:2309.01379 (2023).
[24] Song, Junzhe, and Dmitry Namiot. "A Survey of the Implementations of Model Inversion Attacks." International Conference on Distributed Computer and Communication Networks. Cham: Springer Nature Switzerland, 2022.
[25] Zou, Andy, et al. "Universal and transferable adversarial attacks on aligned language models." arXiv preprint arXiv:2307.15043 (2023).
[26] Compromising LLMs using Indirect Prompt Injection https://github.com/greshake/llm-security Retrieved: 08.09.2023
[27] Introducing Google's Secure AI Framework https://blog.google/technology/safety-security/introducing-googles-secure-ai-framework/ Retrieved: 08.09.2023
[28] Secure AI Framework Approach. A quick guide to implementing the Secure AI Framework (SAIF) https://services.google.com/fh/files/blogs/google_secure_ai_framework_approach.pdf Retrieved: 11.09.2023
[29] Securing AI Pipeline https://www.mandiant.com/resources/blog/securing-ai-pipeline Retrieved: 11.09.2023
[30] Atlas MITRE https://atlas.mitre.org/ Retrieved: 11.09.2023
[31] ATLAS mitigations https://atlas.mitre.org/mitigations/ Retrieved: 11.09.2023
[32] Introduction to red teaming large language models (LLMs) https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/red-teaming Retrieved: 11.09.2023
[33] OpenAI's red team: the experts hired to 'break' ChatGPT https://www.ft.com/content/0876687a-f8b7-4b39-b513-5fee942831e8 Retrieved: 11.09.2023
[34] Microsoft AI Red Team building future of safer AI https://www.microsoft.com/en-us/security/blog/2023/08/07/microsoft-ai-red-team-building-future-of-safer-ai/ Retrieved: 13.09.2023
[35] Bug-bar https://learn.microsoft.com/ru-ru/security/engineering/bug-bar-aiml Retrieved: 13.09.2023
[36] AI-Security-Risk-Assessment https://github.com/Azure/AI-Security-Risk-Assessment/blob/main/AI_Risk_Assessment_v4.1.4.pdf Retrieved: 13.09.2023
[37] AI threat modeling https://learn.microsoft.com/ru-ru/security/engineering/threat-modeling-aiml Retrieved: 13.09.2023
[38] Responsible https://www.microsoft.com/en-us/ai/responsible-ai AI Retrieved: 13.09.2023
[39] Failure modes taxonomy https://learn.microsoft.com/ru-ru/security/engineering/failure-modes-in-machine-learning Retrieved: 13.09.2023

[40] NVIDIA AI Red Team: An Introduction https://developer.nvidia.com/blog/nvidia-ai-red-team-an-introduction/ Retrieved: 13.09.2023
[41] Microsoft Counterfit https://github.com/Azure/counterfit/ Retrieved: 13.09.2023
[42] GARD project https://www.gardproject.org/ Retrieved: 13.09.2023
[43] Master program Cybersecurity https://cyber.cs.msu.ru/ Retrieved: 13.09.2023

Authors:
Dmitry Namiot, Leading Researcher of the Open Information Technologies Lab, Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University (1, Leninskie gory, Moscow 119991, Russian Federation), Dr.Sci. (Eng.), ORCID: http://orcid.org/0000-0002-4463-1678, (e-mail: dnamiot@gmail.com).

Elena Zubareva, Senior Researcher of the Open Information Technologies Lab, Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University (1, Leninskie gory, Moscow 119991, Russian Federation), Cand.Sci. (Ped.), Associate Professor, ORCID: http://orcid.org/0000-0002-9997-4715, (email: e.zubareva@cs.msu.ru).