Доверенные платформы искусственного интеллекта

Д.Е. Намиот, Е.А. Ильюшин, О. Г. Пилипенко

***Аннотация*— Разработка и использование систем искусственного интеллекта (машинного обучения) в критических областях (авионика, автономное движение и т.п.) неизбежно ставит вопрос о надежности используемого программного обеспечения. Доверенные вычислительные системы существуют уже достаточно давно. Их смысл состоит в разрешении выполнения только определенных приложений и гарантии от вмешательства в работу таких приложений. Доверие в этом случае состоит в уверенности в том, что назначенные приложения работают так, как они работали при тестировании. Но в случае машинного обучения этого недостаточно. Приложение может работать так, как оно и задумывалось, никаких вмешательств нет, но результатам доверять нельзя просто потому, что изменились данные. В целом, эта проблема является следствием принципиального момента для всех систем машинного обучения – данные на этапе тестирования (эксплуатации) могут отличаться от таковых же данных, на которых система обучалась. Соответственно, нарушение работы системы машинного обучения возможно и без целенаправленных действий, просто потому, что мы столкнулись на этапе эксплуатации с данными, для которых обобщения, достигнутые на этапе обучения, не работают. А есть еще атаки, под которыми понимаются специальные воздействия на элементы конвейера машинного обучения (тренировочные данные, собственно модель, тестовые данные) с целью либо добиться желаемого поведения системы, либо воспрепятствовать ее корректной работе. На сегодняшний день, эта проблема, которая, в общем случае, связана с устойчивостью работы систем машинного обучения, является главным препятствием для использования машинного обучения в критических приложениях.**

***Ключевые слова*— доверенные системы, состязательные атаки, кибербезопасность систем машинного обучения**

I. Введение

В настоящей статье мы хотим остановиться на понятии доверенной платформы искусственного интеллекта (ИИ). Проблема в том, что понятие “доверенный” по отношению к вычислительным системам используется уже достаточно давно и, в принципе, имеет уже устоявшееся определение, понимаемое всеми одинаково. Доверительные вычисления (Trusted Computing) – это, в оригинале, совокупность стандартов, разрабатываемых на их основе технологий, аппаратного и программного обеспечения для создания высокой степени безопасности при работе программ (проведении вычислений). Безопасность гарантируется тем, что система допускает использование только проверенного (сертифицированного) аппаратного и программного обеспечения. Под такой сертификацией (проверкой) понимается некоторая система подтверждающих цифровых сертификатов, которые и определяют, что и какие компоненты могут делать.

Согласно Wikipedia, Trusted Computing - это технология, разработанная и продвигаемая Trusted Computing Group [1]. Этот термин взят из области доверенных систем и имеет особое значение, отличное от области конфиденциальных вычислений [2]. Основная идея доверенных вычислений состоит в том, чтобы дать производителям оборудования контроль над тем, какое программное обеспечение работает и не работает в системе, отказываясь запускать неподписанное программное обеспечение. Благодаря Trusted Computing компьютер будет постоянно вести себя ожидаемым образом, и это поведение будет обеспечиваться компьютерным оборудованием и программным обеспечением. Обеспечение такого поведения достигается за счет загрузки аппаратного обеспечения с уникальным ключом шифрования, который недоступен для остальной части системы и владельца.

Конфиденциальные вычисления (проект Linux Foundation [2]) защищают используемые данные, выполняя вычисления в аппаратной доверенной среде выполнения. Эти безопасные и изолированные среды предотвращают несанкционированный доступ или изменение приложений и данных во время их использования, тем самым повышая гарантии безопасности для организаций, которые управляют конфиденциальными и регулируемыми данными.

Общая идея в том, чтобы гарантировать выполнение только предписанных программ, исключив любые воздействия на процесс вычислений.

При этом в англоязычной литературе разделяют понятия Trusted computing и Trustworthy computing (доверенные и надежные вычисления). NSA (Национальное агентство безопасности США)

Статья получена 30 мая 2022. Исследование выполнено при поддержке Междисциплинарной научно-образовательной школы Московского университета «Мозг, когнитивные системы, искусственный интеллект»

Д.Е. Намиот – МГУ имени М.В. Ломоносова (email: dnamiot@gmail.com)
Е.А. Ильюшин - МГУ имени М.В. Ломоносова (email: john.ilyushin@gmail.com)
О.Г. Пилипенко - МГУ имени М.В. Ломоносова (email: piligol1995@gmail.com)

определяет доверенный компонент (trusted component) как компонент, сбой в котором не влияет на безопасность всей системы, а надежный компонент (trustworthy component) - как компонент, в котором не произойдет сбой. С другой стороны, Министерство обороны США определяет trusted как нечто, чему мы вынуждены доверять за неимением выбора. Таким образом, термин trusted (доверенный) имеет довольно много значений [3].

Эти различия в толковании не меняют базовой идеи – доверия к вычислительному процессу. Как было сказано выше, термин доверительные (доверенные) вычисления (trusted computing) означает, что данному вычислительному процессу может доверять как производитель вычислительного устройства, так и разработчик программного обеспечения. Соответственно, если мы имеем протестированное программное обеспечение, свободное от возможных закладок [4], то мы можем быть уверены, что на доверенной платформе ничто не помешает правильной работе данной системы.

Термин Искусственный интеллект на сегодняшний день является, на практике, синонимом термина машинное обучение. Система машинного обучения есть, конечно, некоторая программа. Но вот с термином доверие или надежность возникают проблемы. Системы машинного обучения в подавляющем большинстве случаев не являются устойчивыми [5]. То есть отсутствие закладок и вмешательства в их работу не позволяет гарантировать результаты работы систем машинного обучения. Доверенная платформа для работы системы машинного обучения не приводит, автоматически, к доверию результатам работы системы машинного обучения. Соответственно, доверенная платформа для искусственного интеллекта есть нечто большее, чем доверенная платформа вычислений.

II. О ДОВЕРЕННЫХ ПЛАТФОРМАХ

Просто название Доверенные платформы Искусственного Интеллекта не описывает конечное представление (облик) предлагаемого решения. Такое название допускает множественные трактовки (интерпретации), и необходимо четко специфицировать (описать, представить), что же это из себя представляет.

Сама терминология доверенных вычислений (платформ и т.п.), как было отмечено выше, не является новой. Классически – это гарантированное отсутствие несанкционированных (незапланированных) действий (любой природы) во время вычислений (во время эксплуатации вычислительной системы).

Искусственный интеллект, по крайней мере, на сегодня, является синонимом слов машинное обучение. Машинное обучение, по своей природе, имеет непреодолимые проблемы с данными: принципиально, мы обучаем систему (настраиваем ее поведение) на некотором тренировочном наборе данных, что далее обобщаем на неизвестную нам генеральную совокупность данных. Характеристики генеральной совокупности (реальных данных) могут отличаться от тренировочных. Это и есть основная проблема. Все обобщения, достигнутые на этапе тренировки, валидации и тестирования могут оказаться неверными в силу изменения характеристик данных [6]. А раз это так, то возникает соблазн (естественно, в первую очередь это важно для критических систем) специальным образом изменять (подготавливать) данные на разных этапах конвейера машинного обучения. Такие действия называются атаками на системы машинного обучения. Такие атаки не обнаруживаются стандартными средствами кибербезопасности, что и определило внимание к проблемам кибербезопасности систем искусственного интеллекта [7].

Приведем пару цитат, в подтверждение изложенного выше.

Google (Deepmind) в обзорной публикации своей исследовательской группы Robust and Verified Deep Learning group [8]:

“системы машинного обучения, по умолчанию, не являются надежными. Даже системы, которые превосходят людей в определенной области, могут потерпеть неудачу в решении простых проблем, если будут внесены различия в исходные данные”.

Madry Lab (MIT) on Safe ML [9]:

I. Вы не должны тренироваться на данных, которым не полностью доверяете (из-за возможного отравления данных – изменения данных с целью обмана модели)

II. Вы не должны позволять никому использовать вашу модель (или наблюдать за ее работой), если вы полностью им не доверяете (из-за кражи модели и атак черного ящика). Это можно представить как аналогию декомпилирования или reverse engineering в программных системах – работа (поведение) модели изучается с целью построения состязательного примера.

III. Вы не должны полностью доверять предсказаниям вашей модели (из-за возможных состязательных примеров).

В самом общем случае проблемы с моделями машинного обучения описываются как отсутствие устойчивости. Классически, алгоритм, в котором погрешность, допущенная в начальных данных или допускаемая при вычислениях, с каждым шагом не увеличивается или увеличивается незначительно, называется устойчивым. В противном случае, если погрешность существенно увеличивается от шага к шагу, алгоритм называется неустойчивым. Устойчивость алгоритма – это мера его чувствительности к изменениям в исходных данных. Применительно к машинному обучению, устойчивость – это сохранение показателей, достигнутых при тренировке, тестировании и валидации во время эксплуатации системы (точность не ухудшилась, когда данные изменились по сравнению с тренировочными и т.д.). Естественно, что это важно, в первую очередь, для критических систем. Такие системы запустили в эксплуатацию только исходя из достигнутых во время тренировки показателей, и сохранение таких показателей критично для эксплуатации.

Соответственно, доверенные системы ИИ – это системы, именно результатам работы (а не целостности процесса исполнения) которых мы можем доверять. Или (поскольку различие тренировочных и реальных данных, в общем случае, не устранить) – система, для которой проводились (проводятся во время работы) специальные мероприятия (обработка, тестирование, оценка, мониторинг и т.п.), призванные повысить доверие к ее результатам (результатам ее работы). Пример – отчет MIT для Министерства Обороны США о принципах оценки производительности и устойчивости систем машинного обучения [10]. В нем приводятся именно рекомендации для Министерства обороны по оценке устойчивости моделей машинного обучения:

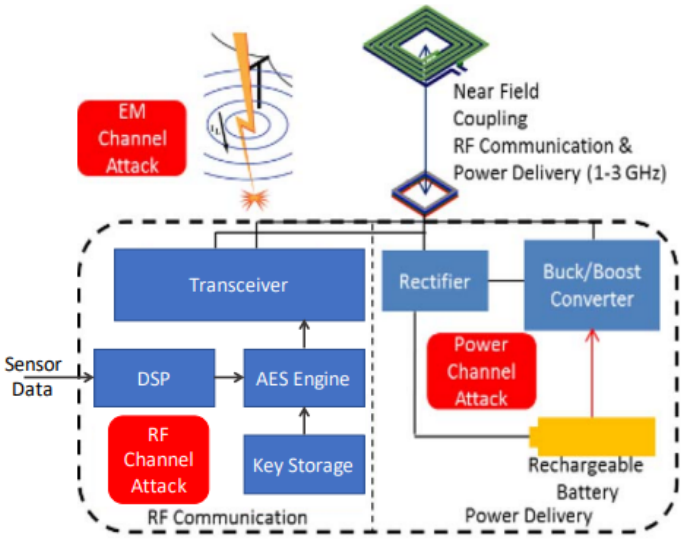
- Создавать тестовые наборы данных с достаточной вариативностью и количеством образцов для эффективного измерения ожидаемой производительности модели на будущих (неизвестных на этапе тренировки) данных после развертывания.
- Поддерживать разделение между данными, используемыми для разработки и оценки (т.е. тестовые данные не используются для разработки модели или обучения ее параметров), чтобы обеспечить честную и беспристрастную оценку возможностей модели.
- Оценить производительность при небольших возмущениях и искажениях входных данных, чтобы оценить чувствительность модели и выявить потенциальные уязвимости.
- Оценить производительность на выборках из распределений данных, которые смещены от предполагаемого распределения, которое использовалось для разработки модели, чтобы оценить, как модель может работать на оперативных данных, которые могут отличаться от данных обучения.

Именно доверие и является основной проблемой внедрения (использования) систем ИИ в критических приложениях (автономное вождение, авионика и т.п.). Отсутствие вмешательства в работу вычислительной системы здесь является, конечно, необходимым требованием, но недостаточным. Большая часть проблем в плане доверия происходит из-за самого машинного обучения. Более того, для этих проблем на сегодняшний день и вовсе нет полного решения.

Поэтому, говоря о доверенном (надежном) ИИ можно отметить следующее. Необходимо разделять доверенные платформы (trusted platforms) и доверенный Искусственный интеллект (Trusted AI, Trustworthy AI)

Доверенная платформа, в первую очередь, это термин, описывающий вычислительную систему, устойчивую к атакам по сторонним каналам (электромагнитное воздействие и т.п.)

Вот иллюстративный пример из работы [11]



Security vulnerabilities of IoT nodes include:

1. **Power side channel attack during AES encryption**
2. **EM side channel attack via near and far fields**
3. **RF channel attack**

Рис.1 Доверенная платформа [11]

Можно сказать, что в этом плане ИИ ничего не добавляет к имеющимся (использующимся) доверенным платформам. В вычислительной технике Trusted Platform Module (TPM) — название спецификации, описывающей криптопроцессор, в котором хранятся криптографические ключи для защиты информации, а также обобщённое наименование реализаций указанной спецификации, например, в виде «чипа TPM» или «устройства безопасности TPM» (Dell). Спецификация TPM разработана некоммерческой организацией Trusted Computing Group. Модуль TPM может использоваться, чтобы подтвердить подлинность аппаратных средств. Так как каждый чип TPM уникален для специфического устройства, это делает возможным однозначное установление подлинности платформы. Например, чтобы проверить, что система, к которой осуществляется доступ — ожидаемая система.

Аппаратные реализации существуют, естественно, для исполнения программ. Критически важным для

безопасности приложениям требуется безопасная, надежная платформа для выполнения, охватывающая программное, микропрограммное и аппаратное обеспечение. Самый нижний уровень, с которым приложение взаимодействует напрямую, это доверенная операционная система (ОС). Доверие к ОС зависит от двух факторов: ее надежности с точки зрения безопасности и уверенности в том, что ОС была загружена и настроена правильно и никогда не подделывалась. Доверие к ОС также частично зависит от доверенных функций до ОС, таких как микропрограмма безопасной загрузки, которая выполняется до ОС (рис.2).

В этом плане (trusted platforms) искусственный интеллект не добавляет ничего нового. Собранное приложение, использующее машинное обучение, является компьютерной программой, которая должна выполняться для критических приложений на доверенной связке OS + вычислитель.

Соответственно, как и для других программ в критической области, для приложений ИИ остается задача использования доверенных компиляторов, общих библиотек и средств разработки (сопровождения).

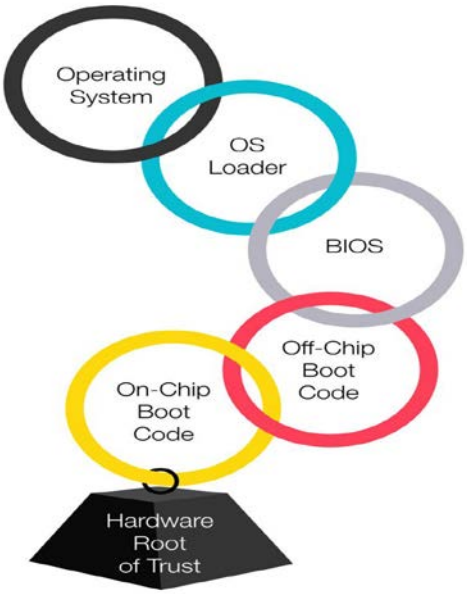


Рис. 2. Доверенные ОС и аппаратура [12]

III О ДОВЕРЕННОМ ИИ

Что касается именно доверенного ИИ, то это не есть некоторая единая реализация какой-либо программы. Это подход (методология) по оценке рисков, измерению характеристик, принципов реализации и т.п., поддерживаемый на различных этапах разными программными средствами (компонентами, фреймворками и т.д.) призванный снизить (или, по крайней мере, оценить) риски использования систем ИИ. В первую очередь, это используется для критических применений, хотя и не ограничивается ими. Например, генерация фальшивых отзывов в системе электронной коммерции есть, фактически, атака на систему машинного обучения в рекомендательной системе, которая снижает уровень доверия к рекомендациям.

Как отмечено в [13], несмотря на растущее согласие в отношении того, что ИИ должен быть этичным и заслуживающим доверия, разработка функциональности ИИ опережает возможности разработчиков по обеспечению ее такими характеристиками, как прозрачность, безопасность, проверяемость. Иными словами, мы получаем результаты без возможности их обосновать и гарантировать воспроизводимость. Это требует от каждой организации, занимающейся созданием систем ИИ разработать модель управления ИИ. В эту модель должны входить политика и стандарты проектирования, инвентаризация всех алгоритмов, включая ключевые детали ИИ, которые создаются с помощью программного обеспечения. Каждый алгоритм в перечне должен подвергаться оценке воздействия для оценки рисков. Должны использоваться инструменты и методы проверки, гарантирующие, что алгоритмы работают должным образом и обеспечивают получение точных, справедливых и непредвзятых результатов. Эти инструменты также должны использоваться для отслеживания изменений в структуре принятия решений алгоритма, равно как и для аудита таких алгоритмов. Модель управления должна использоваться на протяжении всего жизненного цикла ИИ, от выявления проблем до обучения и эксплуатации моделей. Это и есть (так появляется), собственно говоря, надежный (доверенный) ИИ. И реализация такого рода систем – это всегда некоторая последовательность шагов. Например, в [14] – это MLTRL (Machine Learning Technology Readiness Levels).

Доверенный ИИ поддерживается (разрабатывается) как на уровне отдельных компаний, так и на государственном (межгосударственном) уровне. Ниже приведены примеры таких реализаций.

Европейский проект ALTAI - The Assessment List on Trustworthy Artificial Intelligence [15]. Он основан на принятом Евросоюзом общем подходе к построению систем ИИ [16] и представляет собой развитый опросный лист, процесс заполнения и оценки которого призван привлечь внимание разработчиков к отдельным аспектам (архитектуре, характеристикам) своих систем. Базовые темы опроса (они соответствуют принципам, выработанным в [16]) включают в себя, в частности:

- Взаимодействие с персоналом (операторами и т.п.) и контроль системы. Системы искусственного интеллекта должны расширять возможности людей, позволяя им принимать обоснованные решения. В то же время необходимо обеспечить надлежащие механизмы надзора (human-in-the-loop).
- Техническая надежность и безопасность. Один из

основных моментов для критических применений

- Конфиденциальность и управление данными. Защита данных относится к этому пункту
- Прозрачность бизнес-модели данных, системы. Системы ИИ и их решения должны быть объяснены таким образом, чтобы они были адаптированы для заинтересованных сторон. Отметим, что без такой прозрачности (объяснимости), вообще говоря, нельзя обосновать устойчивость систем.
- Отсутствие дискриминации (предвзятости) и справедливость. В технических системах – отсутствие смещений и равномерное представление данных
- Подотчетность. Необходимость механизмов для обеспечения ответственности и подотчетности систем ИИ и их результатов. Возможность аудита, который позволяет оценивать алгоритмы, данные и процессы проектирования, играет здесь ключевую роль, особенно в критически важных приложениях. Без мониторинга эксплуатация систем невозможна.

Европейский проект STAR [17], финансируемый в рамках программы EC Horizon 2020, представляет собой совместный проект 15 европейских партнеров, направленный на разработку новых технологий, позволяющих развертывать основанные на стандартах безопасные, безопасные, надежные и надежные системы искусственного интеллекта, ориентированные на человека, в производственной среде. Одно из основных направлений – сохранность (неизменность) наборов данных, используемых для обучения в производственных системах. Довольно подробное описание как архитектуры, так методов предотвращения состязательных атак (защиты системы) есть в книге [18].

IV О ПЛАТФОРМАХ ИИ

Примером компании, разрабатывающей платформы доверенного ИИ, является Datarobot [19]. Дословно: “В DataRobot мы предоставляем опыт и инструменты для тестирования ваших систем по нескольким параметрам доверия, чтобы разработать искусственный интеллект, который будет работать исключительно эффективно, поддерживать операционное превосходство и отражать ваши ценности. Это надежный ИИ.” То есть доверенный ИИ, как и было указано выше – это методология, поддержанная программными инструментами.

В связи с этим продуктом можно отметить следующее. Такого рода системы (автоматизирующие решение задач на основе машинного обучения) также называют платформами машинного обучения. Таких решений на сегодняшний день – десятки, если не сотни. Sagemaker, Automation AI, Roboflow, Supervisely и т.д. Kaggle, например, в этом смысле, также автоматизирует построение и исполнение моделей. Интерфейсы с пользователями и разработчиками у таких систем могут быть самыми разными (low code, визуальное программирование и т.п.). DataRobot выделен только потому, что там прямо говорится о тестировании построенных моделей на предмет доверия.

В альманахе ИИ, издаваемом в МФТИ [22] есть ссылка на описание российской платформы от компании ГосНИИАС [23]. Насколько это работающая система по описанию сказать нельзя, но в рекламном проспекте говорится именно о “типовых решениях на основе нейронных сетей, которые реализуют готовые технологии алгоритмов обучения”, то есть именно об автоматизации построения систем машинного обучения.

В принципе, все, относящиеся к категории AutoML [20] можно назвать платформой искусственного интеллекта. Например, Google Cloud Video Intelligence, Azure AutoML – это все можно назвать (и называется) платформой.

AutoML здесь выступает некоторой конечной (недостижимой в обозримом будущем) целью. Полная автоматизация построения моделей невозможна, но какие-то элементы могут быть, конечно, автоматизированы. И этот процесс автоматизации, в целом, идет по нарастающей.

Перечисленные системы автоматизируют построение и тренировку моделей. После чего, эта модель может запускаться в облаке для обработки представленных данных, либо – можно выгрузить сериализованную (тем или иным способом, например, с помощью ONNX [21]) модель и запускать ее (только исполнение – inference) уже на какой-то другой вычислительной платформе. Именно так это будет работать для встраиваемых систем. Естественно, на каждой такой вычислительной платформе нужно будет один раз реализовать интерпретатор для сохраненной модели или воспользоваться каким-то открытым решением, например, упомянутым ранее ONNX. Это не автоматизирует на 100% решение для конечной платформы в плане использования системы машинного обучения, но выполнит большую его часть. Останется реализовать ввод данных и передачу их интерпретатору модели. А для обеспечения надежности нужно будет организовать мониторинг работы системы, чтобы отслеживать возможный сдвиг данных.

В такой конфигурации очевидно место доверенной вычислительной платформы. И тренировать сети на платформе ИИ необходимо так, чтобы в построенных моделях не оказалось закладок, и исполнение созданной модели должно быть защищено от внешних воздействий. Такая защита не устранит собственные проблемы систем машинного обучения (устойчивость), то есть не будет достаточной, но будет, очевидно, необходимой.

Соответственно, на сегодняшний день общепринятая трактовка понятия платформа ИИ – это инструментарий по автоматизации построения решений (практически – моделей машинного обучения). А доверенной такая платформа станет тогда, когда будет включать элементы

повышения доверия к создаваемым моделям (анализировать исходные данные, проводить состязательное тестирование, строить объяснения, предлагать решения для проблемы сдвига данных/концепции и т.д.).

V О ДОВЕРЕННОМ ИИ

В отчете Комиссии США по вопросам национальной безопасности и искусственного интеллекта (NSCAI) о конкурентоспособности страны в сфере ИИ [24] в разделе, описывающем экосистему ИИ для DoD (Министерства обороны) дословно сказано следующее:

“These are platform environments with ready-made workflows that can be tailored and launched depending on user type (e.g., researcher, industry partner, operator) and use case (e.g., development, TEVV [test, evaluation, validation, and verification], fielding)”

Готовые рабочие процессы (Ready-made workflows) для машинного обучения – это как раз и есть системы автоматизация конвейера ML, то есть некоторый AutoML, как об этом говорилось выше.

Одно из наиболее четких представлений о том, какого рода инструменты входят в надежный (доверенный) ИИ дает проект IBM Trusted AI [25]. Следующие разделы включены в этот проект (каждый из них включает примеры инструментов и исследовательские работы):

Тестирование ИИ: разработка инструментов, которые помогут убедиться, что системы ИИ заслуживают доверия, надежны и могут оптимизировать бизнес-процессы. Создание тестов для имитации реальных сценариев и локализации сбоев в системах ИИ. Автоматизация тестирования, отладки и исправления моделей ИИ в самых разных сценариях.

Состязательная устойчивость и сохранение конфиденциальности: даже передовые системы искусственного интеллекта могут быть уязвимы для атак со стороны противника. Создание инструментов для защиты ИИ и сертификации его надежности, включая количественную оценку уязвимости нейронных сетей и разработку новых атак для повышения эффективности защиты. Помощь системам ИИ в соблюдении требований конфиденциальности.

Объяснимый ИИ – объяснения имеют большое значения для доверия системам ИИ. Создание инструментов для отладки ИИ, когда системы могут объяснить, что они делают и на основе чего они это делают. Объяснимый ИИ можно разделить на три части: объяснимость данных – какие данные (или признаки) наиболее релевантны для решения задачи; объяснение предсказаний ИИ – на основе каких признаков экземпляра данных было принято решение; объяснение работы модели – благодаря чему и как ИИ обработал входные данные. Этот этап включает в себя обучение высокооптимизированных, непосредственно интерпретируемых моделей, а также объяснение моделей черного ящика и визуализацию информационных потоков нейронной сети.

Чувствительность ИИ к изменениям: мера влияния изменений по определенным, выученным системой ИИ, признакам. Отдельный этап объяснимости систем с ИИ, суть которого сводится в определении наиболее отклоняемых от референсных значений малыми возмущениями признаков.

Справедливость, подотчетность, прозрачность: предвзятость (выражаемая как смещения данных, например) может возникнуть на любом этапе жизненного цикла разработки ИИ. Чтобы повысить подотчетность систем искусственного интеллекта с высоким уровнем риска, нужны технологии, повышающие их сквозную прозрачность и справедливость.

Надежная и доверенная генерация: данные являются ключом к технологическим инновациям, поэтому очень важны теоретические и алгоритмические основы для генеративного ИИ, а также для синтеза реалистичных, разнообразных и целевых данных. Эти подходы призваны упростить дополнение данных для надежного машинного обучения и ускорить создание новых разработок.

Количественная оценка неопределенности: когда ИИ может объяснить, что он не уверен, то он добавляет важный уровень прозрачности для его безопасного развертывания и использования. Разработка способов поощрения и оптимизации общепринятых практик количественной оценки, улучшения информирования о неопределенности в жизненном цикле разработки приложений ИИ.

К сожалению, часто соблюдение свойств из перечисленных выше влекут снижение таких важных характеристик как:

Переносимость: способность системы ИИ работать на высоком уровне качества на различных доверенных устройствах.

Обобщаемость: способность системы ИИ работать на высоком уровне качества с различными данными.

Известна, например, связь устойчивости модели и ее точности. Повышение точности модели ведет к ухудшению реакции на возможные изменения данных. На рисунке 3 как раз показано такое соответствие точности и устойчивости.

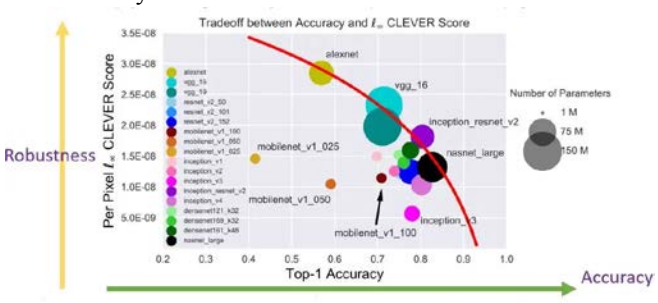


Рис. 3. Точность против устойчивости [32]

То, что называют перетренировкой модели, в большинстве случаев есть именно отсутствие устойчивости. Это приводит к необходимости

выработки каких-то компромиссов при создании практических моделей (что, кстати, является одной из причин невозможности полной автоматизации).

Примером сборки различных инструментов для поддержки надежного ИИ является базовая работа OECD (Организация экономического сотрудничества и развития) [26]. В этом отчете представлена структура (фреймворк) для сравнения инструментов и методов внедрения надежных систем искусственного интеллекта (рис. 4).

Доверенная платформа здесь – это цели (первый столбец), их описания (третий столбец) и примеры инструментария. Цель - помочь собирать, структурировать и обмениваться информацией, знаниями и извлеченными на сегодняшний день уроками в отношении инструментов, методов и подходов к внедрению надежного ИИ. Этот фреймворк послужит основой для разработки интерактивной общедоступной базы данных OECD.AI Policy Observatory.

Обзор большого количества программ, включая разработку инструментальных средств, в области устойчивого машинного обучения приводится в работе [27].

Objective	Tool	Description
Fairness	AT&T software System to Integrate Fairness Transparently (SIFT)	Software system to integrate mechanised and human-in-the-loop components in bias detection, mitigation, and documentation of projects at various stages of the machine learning lifecycle.
	Microsoft Fairlearn	Open-source toolkit to assess and improve the fairness of machine learning models. Contains an interactive visualisation dashboard and bias mitigation algorithms to help navigate trade-offs between fairness and model performance.
	LinkedIn Fairness Toolkit (LiFT)	Open-source toolkit to enable measurement of fairness according to a multitude of fairness definitions in large-scale machine learning workflows.
	Google What-If Tool	Open-source software tool to visually inspect and explore machine learning model performance and data across multiple hypothetical situations, with minimal coding required.
	IBM AI Fairness 360	Open-source toolkit to help detect and mitigate unwanted bias in machine learning models and datasets. Provides approximately 70 metrics to test for biases, and 10 algorithms to mitigate bias in datasets and models.
Transparency	IEEE Standard for Transparency of Autonomous Systems	Technical standard to describe measurable and testable levels of transparency, so that autonomous systems can be assessed and levels of compliance determined.
	Google Model Card Toolkit	Documentation framework for sharing the essential facts of a machine learning model in a structured, accessible way, providing an overview of what the model is intended to do, how it was architected, trained, and its limitations.
Explainability	Google Cloud Explainable AI service	Software to help developers get explanations on the outcomes of their models. Can be applied to the AI models trained on tabular, image, and text data. Not open source.
	IBM AI Explainability 360 Toolkit	Open-source toolkit of algorithms, code, guides, tutorials, and demos to support the interpretability and explainability of machine learning models.
	Microsoft InterpretML	Open-source toolkit containing machine learning interpretability algorithms to help understand model predictions.
Robustness	IBM Adversarial Robustness 360 Toolkit	Open-source toolkit for machine learning security. It provides tools to evaluate, defend, certify and verify machine learning models and applications against the adversarial threats of evasion, poisoning, extraction, and inference.

Рис. 4. Пример набора инструментов.

Отдельно необходимо отметить политику открытости в отношении инструментов для доверенного ИИ. Например, проект GARD (Guaranteeing AI Robustness to Deception - Гарантия устойчивости ИИ к обману) – развивается под эгидой оборонного агентства США DARPA. Программа DARPA GARD направлена на создание теоретических основ системы машинного обучения для выявления уязвимостей системы, описания свойств, которые повысят надежность системы, и поощрения создания эффективных средств защиты. DARPA отмечает, что в настоящее время средства защиты систем машинного обучения, как правило, очень специфичны и эффективны только против конкретных атак. GARD стремится разработать средства защиты, способные защитить от широкого круга атак. Кроме того, современные парадигмы оценки надежности ИИ часто фокусируются на упрощенных мерах, которые могут не иметь отношения к безопасности. Чтобы проверить актуальность для безопасности и широкую применимость, средства защиты, созданные в рамках GARD, будут измеряться на новом испытательном стенде с использованием оценок на основе сценариев [28]. Все инструменты в рамках этого проекта – открыты.

Другое отдельное направление представляет собой сертификация систем ИИ. Критические системы давно сертифицируются по специальным программам, из

которых DO-178 является наиболее известным. Если система Искусственного интеллекта есть, в конечном счете, программа, то она, по смыслу, должна сертифицироваться, как, например, и все остальные программы в авионике [29]. Что, очевидно, представляет собой проблему в силу недетерминированной природы результатов. Это отдельная тема.

Благодарности

Мы благодарны сотрудникам кафедры Информационной безопасности факультета Вычислительной математики и кибернетики МГУ имени М.В. Ломоносова за ценные обсуждения данной работы.

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

Статья является продолжением серии публикаций, посвященных устойчивым моделям машинного обучения [5, 27, 30]. Она подготовлена в рамках проекта кафедры Информационной безопасности факультета ВМК МГУ имени М.В. Ломоносова по созданию и развитию магистерской программы "Искусственный интеллект в кибербезопасности" [31].

Библиография

[1] Trusted Computing https://en.wikipedia.org/wiki/Trusted_Computing
[2] Confidential computing https://confidentialcomputing.io/
[3] Доверительные вычисления https://intuit.ru/studies/courses/955/285/lecture/7166
[4] Марков, А. С., and А. А. Фадин. "Систематика уязвимостей и дефектов безопасности программных ресурсов." Защита информации. Инсайд 3 (2013): 56-61.
[5] Намиот, Д. Е., Е. А. Ильюшин, and И. В. Чижов. "Основания для работ по устойчивому машинному обучению." International Journal of Open Information Technologies 9.11 (2021): 68-74.
[6] Gama, Joao, et al. "Learning with drift detection." Brazilian symposium on artificial intelligence. Springer, Berlin, Heidelberg, 2004.
[7] Kurakin, Alexey, Ian Goodfellow, and Samy Bengio. "Adversarial machine learning at scale." arXiv preprint arXiv:1611.01236 (2016).
[8] Robust and Verified Deep Learning group https://deepmindsafetyresearch.medium.com/towards-robust-and-verified-ai-specification-testing-robust-training-and-formal-verification-69bd1bc48bda
[9] Madry Lab https://people.csail.mit.edu/madry/6.S979/files/lecture_4.pdf Retrieved: May, 2022
[10] Principles for evaluation of AI/ML model performance and robustness file:///C:/temp/principles-evaluation-aiml-model-performance-brown-md-62.pdf Retrieved: May, 2022
[11] Machine Learning for Trusted Platform Design A. Raychowdhury & M. Swaminathan, GaTechhttp://publish.illinois.edu/caeml-industry/files/2018/07/2A2-Machine-Learning-for-Trusted-Platform-Design-1.pdf Retrieved: May, 2022
[12] Creating a trusted platform for embedded security-critical applications https://militaryembedded.com/avionics/software/creating-a-trusted-platform-for-embedded-security-critical-applications Retrieved: May, 2022
[13] How do you teach AI the value of trust? https://assets.ey.com/content/dam/ey-sites/ey-com/en_gl/topics/digital/ey-how-do-you-teach-ai-the-value-of-trust.pdf Retrieved: May, 2022
[14] Lavin, Alexander, et al. "Technology readiness levels for machine learning systems." arXiv preprint arXiv:2101.03989 (2021).
[15] ALTAI https://futurium.ec.europa.eu/en/european-ai-alliance/pages/altai-assessment-list-trustworthy-artificial-intelligence Retrieved: May, 2022
[16] Ethics guidelines for trustworthy AI https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai Retrieved: May, 2022
[17] STAR https://star-ai.eu/ Retrieved: May, 2022
[18] Trusted Artificial Intelligence in Manufacturing; Trusted Artificial Intelligence in Manufacturing https://library.oapen.org/handle/20.500.12657/52612 Retrieved: May, 2022
[19] Datarobot https://www.datarobot.com/platform/trusted-ai/ Retrieved: May, 2022
[20] He, Xin, Kaiyong Zhao, and Xiaowen Chu. "AutoML: A survey of the state-of-the-art." Knowledge-Based Systems 212 (2021): 10662
[21] ONNX https://onnx.ai/ Retrieved: May, 2022
[22] AI Report https://aireport.ru/ Retrieved: May, 2022
[23] Унифицированная программная платформа для разработки конечно ориентированных программных комплексов автоматического распознавания объектов на основе нейросетевых подходов https://www.gosniias.ru/pages/d/platforma-fh.pdf Retrieved: May, 2022
[24] NSCAI report https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf Retrieved: May, 2022
[25] Trusted AI https://research.ibm.com/teams/trusted-ai Retrieved: May, 2022
[26] Tools for trustworthy AI https://www.oecd-ilibrary.org/content/paper/008232ec-en Retrieved: May, 2022
[27] Намиот, Д. Е., Е. А. Ильюшин, И. В. Чижов. "Текущие академические и индустриальные проекты, посвященные устойчивому машинному обучению." International Journal of Open Information Technologies 9.10 (2021): 35-46.
[28] Holistic Evaluation of Adversarial Defenses | GARD Project https://www.gardproject.org/ Retrieved: May, 2022
[29] Artificial Intelligence and avionics software per DO-178C https://afuzion.com/artificial-intelligence-and-avionics-software-per-do-178c/ Retrieved: May, 2022
[30] ЕА Ильюшин, ДЕ Намиот, ИВ Чижов. "Атаки на системы машинного обучения - общие проблемы и методы." International Journal Of Open Information Technologies 10.3 (2022): 17-22.
[31] Artificial Intelligence in Cybersecurity. http://master.cmc.msu.ru/?q=ru/node/3496 (in Russian) Retrieved: May, 2022.
[32] AI Tradeoff: Accuracy or Robustness? https://www.eetimes.com/ai-tradeoff-accuracy-or-robustness/ Retrieved: May, 2022.

On Trusted AI Platforms

Dmitry Namiot, Eugene Ilyushin, Oleg Pilipenko

***Abstract*— The development and use of artificial intelligence systems (machine learning) in critical areas (avionics, autonomous movement, etc.) inevitably raise the question of the reliability of the software used. Trusted computing systems have been around for a long time. Their meaning is to allow the execution of only certain applications and guarantee against interference with the work of such applications. Trust in this case is the confidence that the assigned applications work as they did when tested. But in the case of machine learning, this is not enough. The application can work as intended, there is no intervention, but the results cannot be trusted simply because the data has changed. In general, this problem is a consequence of a fundamental point for all machine learning systems - the data at the testing (operation) stage may differ from the same data on which the system was trained. Accordingly, a violation of the machine learning system is possible without targeted actions, simply because we encountered data at the operational stage for which the generalization achieved at the training stage does not work. And there are also attacks, which are understood as special actions on the elements of the machine learning pipeline (training data, the model itself, test data) in order to either achieve the desired behavior of the system or prevent it from working correctly. Today, this problem, which is generally associated with the stability of machine learning systems, is the main obstacle to the use of machine learning in critical applications.**

***Keywords*— trusted systems, adversarial attacks, machine learning cybersecurity**

REFERENCES

[1] Trusted Computing https://en.wikipedia.org/wiki/Trusted_Computing
[2] Confidential computing https://confidentialcomputing.io/
[3] Doveritel'nye vychislenija https://intuit.ru/studies/courses/955/285/lecture/7166
[4] Markov, A. S., and A. A. Fadin. "Sistematika ujazvimostej i defektov bezopasnosti programmnyh resursov." Zashhita informacii. Insajd 3 (2013): 56-61.
[5] Namiot, D. E., E. A. Il'jushin, and I. V. Chizhov. "Osnovanija dlja rabot po ustojchivomu mashinnomu obucheniju." International Journal of Open Information Technologies 9.11 (2021): 68-74.
[6] Gama, Joao, et al. "Learning with drift detection." Brazilian symposium on artificial intelligence. Springer, Berlin, Heidelberg, 2004.
[7] Kurakin, Alexey, Ian Goodfellow, and Samy Bengio. "Adversarial machine learning at scale." arXiv preprint arXiv:1611.01236 (2016).
[8] Robust and Verified Deep Learning group https://deepmindsafetyresearch.medium.com/towards-robust-and-verified-ai-specification-testing-robust-training-and-formal-verification-69bd1bc48bda
[9] Madry Lab https://people.csail.mit.edu/madry/6.S979/files/lecture_4.pdf Retrieved: May, 2022
[10] Principles for evaluation of AI/ML model performance and robustness file:///C:/temp/principles-evaluation-aiml-model-performance-brown-md-62.pdf Retrieved: May, 2022
[11] Machine Learning for Trusted Platform Design A. Raychowdhury & M. Swaminathan, GaTechhttp://publish.illinois.edu/caeml-industry/files/2018/07/2A2-Machine-Learning-for-Trusted-Platform-Design-1.pdf Retrieved: May, 2022
[12] Creating a trusted platform for embedded security-critical applications https://militaryembedded.com/avionics/software/creating-a-trusted-platform-for-embedded-security-critical-applications Retrieved: May, 2022
[13] How do you teach AI the value of trust? https://assets.ey.com/content/dam/ey-sites/ey-com/en_gl/topics/digital/ey-how-do-you-teach-ai-the-value-of-trust.pdf Retrieved: May, 2022
[14] Lavin, Alexander, et al. "Technology readiness levels for machine learning systems." arXiv preprint arXiv:2101.03989 (2021).
[15] ALTAI https://futurium.ec.europa.eu/en/european-ai-alliance/pages/altai-assessment-list-trustworthy-artificial-intelligence Retrieved: May, 2022
[16] Ethics guidelines for trustworthy AI https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai Retrieved: May, 2022
[17] STAR https://star-ai.eu/ Retrieved: May, 2022
[18] Trusted Artificial Intelligence in Manufacturing; Trusted Artificial Intelligence in Manufacturing https://library.oapen.org/handle/20.500.12657/52612 Retrieved: May, 2022
[19] Datarobot https://www.datarobot.com/platform/trusted-ai/ Retrieved: May, 2022
[20] He, Xin, Kaiyong Zhao, and Xiaowen Chu. "AutoML: A survey of the state-of-the-art." Knowledge-Based Systems 212 (2021): 10662
[21] ONNX https://onnx.ai/ Retrieved: May, 2022
[22] AI Report https://aireport.ru/ Retrieved: May, 2022
[23] Unificirovannaja programmnaja platforma dlja razrabotki konechno orientirovannyh programmnyh kompleksov avtomaticheskogo raspoznavanija ob"ektov na osnove nejrosetevyh podhodov https://www.gosniias.ru/pages/d/platforma-fh.pdf Retrieved: May, 2022
[24] NSCAI report https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf Retrieved: May, 2022
[25] Trusted AI https://research.ibm.com/teams/trusted-ai Retrieved: May, 2022
[26] Tools for trustworthy AI https://www.oecd-ilibrary.org/content/paper/008232ec-en Retrieved: May, 2022
[27] Namiot, D. E., E. A. Il'jushin, I. V. Chizhov. "Tekushhie akademicheskie i industrial'nye proekty, posvjashhennye ustojchivomu mashinnomu obucheniju." International Journal of Open Information Technologies 9.10 (2021): 35-46.
[28] Holistic Evaluation of Adversarial Defenses | GARD Project https://www.gardproject.org/ Retrieved: May, 2022
[29] Artificial Intelligence and avionics software per DO-178C https://afuzion.com/artificial-intelligence-and-avionics-software-per-do-178c/ Retrieved: May, 2022
[30] EA Il'jushin, DE Namiot, IV Chizhov. "Ataki na sistemy mashinnogo Obuchenija - obshhie problemy i metody." International Journal Of Open Information Technologies 10.3 (2022): 17-22.
[31] Artificial Intelligence in Cybersecurity. http://master.cmc.msu.ru/?q=ru/node/3496 (in Russian) Retrieved: May, 2022.
[32] AI Tradeoff: Accuracy or Robustness? https://www.eetimes.com/ai-tradeoff-accuracy-or-robustness/ Retrieved: May, 2022.

.