ФИЗИКА ЭЛЕМЕНТАРНЫХ ЧАСТИЦ И АТОМНОГО ЯДРА
2024. Т. 55, вып. 3. С. 530–536

О СЕРТИФИКАЦИИ СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

*Д. Е. Намиот *, Е. А. Ильюшин*

Московский государственный университет им. М. В. Ломоносова, Москва

Системы машинного обучения в настоящее время являются основными примерами использования искусственного интеллекта в самых разнообразных областях. С практической точки зрения можно сказать, что машинное обучение является синонимом понятия «искусственный интеллект». Распространение технологий машинного обучения приводит к необходимости их применения в так называемых критических областях: авионике, атомной энергетике, автоматическом вождении и т. д. Традиционное программное обеспечение, например, в авионике проходит специальные процедуры сертификации, которые не могут быть прямо перенесены на модели машинного обучения. Рассматриваются подходы к сертификации моделей машинного обучения.

Machine learning systems are today the main examples of the use of Artificial Intelligence in a wide variety of areas. From a practical point of view, we can say that machine learning is synonymous with the concept of Artificial Intelligence. The spread of machine learning technologies leads to the need for their application in the so-called critical areas: avionics, nuclear energy, automatic driving, etc. Traditional software, for example, in avionics, undergoes special certification procedures that cannot be directly transferred to machine learning models. The paper discusses approaches to certification of machine learning models.

PACS: 07.05.Tp

ВВЕДЕНИЕ

Модели машинного обучения зависят от данных, на которых они обучались. Изменение данных на этапе обучения ведет, например, к изменению параметров модели. Изменение входных данных (по отношению к данным, на которых модель обучалась) ведет к изменению результатов работы. Такие изменения могут быть весьма существенными и качественного характера (например, изменение классификации объектов и т. п.) или могут привести к снижению точности работы системы. Соответственно, исходя из этого и возникают так называемые состязательные атаки на модели машинного обучения — сознательные модификации данных на разных этапах конвейера, которые призваны либо помешать работе системы машинного обучения, либо, наоборот, добиться желаемого для атакующего результата работы.

* E-mail: dnamiot@gmail.com

Google (Deepmind) в обзорной публикации собственной исследовательской группы Robust and Verified Deep Learning group отмечает, что «системы машинного обучения по умолчанию не являются надежными. Даже системы, которые превосходят людей в определенной области, могут претерпеть неудачу в решении простых проблем, если будут внесены различия в исходные данные» [1].

Презентация Madry-lab (MIT) представила три заповеди Secure/Safe ML [2]:

1. Вы не должны тренироваться на данных, которым не полностью доверяете (из-за возможного «отравления» данных — изменения данных с целью обмана модели).

2. Вы никому не должны позволять использовать вашу модель (или наблюдать за ее работой), если полностью 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.

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

Особую значимость такие состязательные примеры имеют, естественно, для критических приложений (в области авионики, автоматического вождения, ядерной энергетики и т. п.). Последствия ошибок здесь всегда серьезные, и для подобного рода систем могут найтись заинтересованные в подобных атаках лица.

При этом важно отметить, что модели машинного обучения могут не показывать на реальных данных той производительности, которая была на этапе тренировки безо всякого специального воздействия. Это связано с так называемой проблемой сдвига данных и вопросом робастности моделей, который, на самом деле, является даже более общим, чем состязательные атаки. Итак, в целом можно сказать, что машинное обучение (в настоящей форме) дает возможность получать результаты, но имеет проблемы с их гарантированием. А для критических систем нужны именно гарантии. В данной работе приводится обзор возможных источников для гарантий результатов работы моделей машинного обучения.

ЗАКОНОДАТЕЛЬСТВО, АУДИТ И СЕРТИФИКАЦИЯ

Обеспечение гарантий результатов работы естественным образом входит в различные регулирующие акты для систем искусственного интеллекта (ИИ, ML). Такого рода акты начали активно готовиться на уровне как государств (США, Китай), межгосударственных структур (закон Европейского союза об искусственном интеллекте), групп государств (G7 — Хиросимский процесс как межправительственная целевая группа

для исследования рисков генеративного ИИ), так и частных компаний (OpenAI, Microsoft и Google).

Законы должны опираться на соответствующие процедуры анализа. Например, для Европейского центра алгоритмической прозрачности (ECAT) определены три основные задачи:

— расследование — это оценка функционирования алгоритмов «черного ящика»;

— исследование — это анализ возможностей алгоритмов рекомендаций для распространения незаконного контента, нарушения прав человека, нанесения ущерба демократии или вреда здоровью пользователей и т. п.;

— создание центра обмена информацией и передовым опытом между исследователями в академических кругах и промышленностью.

По европейскому законодательству системы ИИ с неприемлемым уровнем риска для безопасности людей будут запрещены.

Но все перечисленные акты определяют итоговые требования к законченному продукту. Они не определяют практических шагов по достижении этих требуемых характеристик. Также законодательные акты не определяют и метрики, которые должны использоваться при оценке этих требуемых характеристик.

Понятия «аудит» и «сертификация» уже непосредственно относятся к практической области. Классическая интерпретация: аудит представляет собой процесс инспекции (проверки), а сертификация — это уже подтверждение (гарантия) данных (результатов работы).

Аудит систем машинного обучения — новая и достаточно быстро развивающаяся область. Причины — указанные выше проблемы с гарантированием результатов работы. Отчет Game Changers среди девяти технологий, которые изменят каждую индустрию, на первом месте называет именно AI-аудит [3].

По факту аудит для систем машинного обучения — это набор лучших практик по тому, что и как проверять для готовых систем. Аудит — это некоторый чек-лист, что должно быть сделано для проверки модели. Именно действия, которые должны быть выполнены, но не гарантия того, что будет достигнут заданный результат.

Аудит для системы машинного обучения (ИИ) — это оценка алгоритмов, моделей, данных и процессов проектирования. Такая оценка приложений ИИ внутренними и внешними аудиторами помогает обосновать надежность системы ИИ, продемонстрировать ответственность проектировщиков и повысить обоснованность прогнозов, сделанных моделями. Аудит ИИ охватывает [4]:

— оценку моделей, алгоритмов и потоков данных;

— анализ операций, результатов и обнаруженных аномалий;

— технические аспекты систем ИИ для оценки точности результатов;

— этические аспекты систем ИИ для справедливости, законности и конфиденциальности.

Все это соответствует общепринятому определению того, что аудит — это инструмент для опроса сложных процессов, для определения того, соответствуют ли они политике компании, отраслевым стандартам или правилам. Стандарт IEEE для разработки программного обеспечения определяет аудит как «независимую оценку соответствия программных продуктов и процессов применяемым нормам, стандартам, руководствам, планам, спецификациям и процедурам» [5].

Есть готовые фреймворки, которые помогают ориентироваться в этих задачах. Под словом «фреймворк» в данном случае понимается некоторая структурированная спецификация действий (шагов), которые необходимо выполнить. В качестве примеров можно назвать IIA Artificial Intelligence Auditing Framework [6], Deloitte's Trustworthy AI Framework [7], Gartner AI TRiSM (Artificial Intelligence (AI) Trust, Risk, and Security Management — управление доверием, рисками и безопасностью искусственного интеллекта) [8].

К теме аудита технически нужно отнести и так называемые доверенные платформы для разработки ИИ-приложений [9]. Идея доверенных платформ в компьютерных науках не нова. Основной смысл доверенных вычислений состоит в том, чтобы дать производителям оборудования контроль над тем, какое программное обеспечение работает (не работает) в системе, отказываясь запускать неподписанное программное обеспечение. Благодаря доверенным вычислениям компьютер будет постоянно вести себя ожидаемым образом, и это поведение будет обеспечиваться компьютерным оборудованием и программным обеспечением. Поддержка такого поведения достигается за счет загрузки аппаратного обеспечения с уникальным ключом шифрования, который недоступен для остальной части системы и ее владельца. Эта концепция необходима и для систем машинного обучения в критических применениях, поскольку есть, например, атаки, которые ориентированы на фреймворки машинного обучения. Изменение, например, функции вычисления потерь в конкретном фреймворке будет затрагивать все модели машинного обучения на такой платформе [10]. Но для систем машинного обучения это лишь самая малая из проблем, поскольку основная проблема происходит именно из-за отсутствия доверия к обработке данных. И доверенные платформы — это платформы, инструменты которых позволяют повысить доверие к моделям машинного обучения, платформы, которые позволяют анализировать тренировочные данные, противостоять состязательным атакам, определять сдвиги данных при работе системы и т. д. Примерами таких платформ являются Datarobot или IBM Trustworthy.

Сертификация, как отмечено выше, — это уже гарантия результатов работы системы. При этом возникает коллизия с сертификацией ML-систем (моделей). Для ML-систем (моделей) сертификация — это получение оценок выбранных метрик (в том числе и вероятностных оценок). Для программ — это именно гарантия работоспособности. Дословно можно сказать так: «Системы авионики должны безопасно

выполнять свои функции по назначению во всех прогнозируемых условиях эксплуатации и окружающей среды» [11]. Модель машинного обеспечения на этапе вывода — это обычная программа. И, соответственно, она должна сертифицироваться, как и любая другая программа для критических применений. И вероятностные оценки здесь уже вообще «не работают».

Каким же образом может гарантироваться работа систем машинного обучения? Гарантии для программного обеспечения (Software Assurance или SwA) — это критический процесс разработки, который обеспечивает надежность, безопасность и защищенность программных продуктов. Он включает в себя множество действий: анализ требований, анализ проекта, проверку кода, тестирование и формальную проверку.

Есть классическая V-модель разработки программного обеспечения [12]. Два направления проверки (две стороны буквы V):

— верификация: правильно ли мы строим продукт?

— валидация: построен ли правильный продукт?

На каждом промежуточном уровне каждого направления есть соответствующий набор тестов. Во время верификации проверяется, соответствует ли продукт требованиям: у него есть все функции для использования по назначению, как описано на этапе планирования после проверки с его потенциальными пользователями, и эти функции работают по назначению. Это подразумевает, во-первых, установление требований, а во-вторых, создание на их основе спецификации проекта системы. Затем разработка движется вглубь, при этом уточняются данные предыдущего шага. Во время валидации проверяется, описывают ли требования то, что действительно необходимо, правильно ли они учитывают цели заинтересованных сторон, соответствует ли полученное программное обеспечение модели применения.

Для систем машинного обучения (нейронных сетей) есть, очевидно, компоненты, которые могут быть проверены подобным образом. Например, анализ входных данных, мониторинг работы системы и т. п. Но ключевая функция (вывод) таким образом (построчно) проверена быть не может. Компания Daedalean и агентство EASA (European Union Aviation Safety Agency) предложили термин Learning Assurance (гарантии обучения) вместо Software Assurance и соответствующую W-модель. Это предложение было опубликовано как концепция обеспечения проектирования для нейронных сетей (Concepts of Design Assurance for Neural Networks или CoDANN) [13]. Она может стать основой для будущих нормативных требований.

EASA также опубликовало дорожную карту для своих проектов сертификации, в которой сертификация приложений первого уровня (ассистенты для человека) относится к 2025 г., а последнего третьего уровня (неотменяемые действия) — только к 2035–2050 гг.

В работе [14] авторы достаточно подробно останавливаются на принципиальной несовместимости процесса разработки ML-приложений и по-

ложений DO-178. Основные несоответствия могут быть представлены следующим образом:

• Детерминированный подход к сертификации программных систем против недетерминированных моделей машинного обучения.

• Покрытие кода. Упомянутая выше V-модель позволяет восстановить обоснование для произвольной строчки кода: зачем она нужна и для удовлетворения какой именно спецификации появилась. Это, очевидно, не так для моделей машинного обучения.

• Охват данных. Стандартный подход в ML — это точечная робастность. Сертификация моделей машинного обучения — это исследование устойчивости в некотором ограниченном диапазоне модификаций правильных данных, тогда как для сертификации программ должны исследоваться все возможные значения.

ЗАКЛЮЧЕНИЕ

В настоящее время сертификация систем машинного обучения, как это понимается для традиционного программного обеспечения, в общем случае невозможна. Работающим детерминистским подходом является формальная верификация моделей машинного обучения, но она имеет проблемы с масштабируемостью. Возможно, решением для сертификации систем машинного обучения будет изменение существующих стандартов. С практической точки зрения сертификация моделей машинного обучения — это сертификация робастности, когда гарантируются метрики при заданном бюджете (размере) модификации тренировочных данных.

Правовое регулирование ИИ не имеет отношения к доказательству работоспособности систем ИИ, а описывает только требования к конечному продукту.

Аудит систем ИИ является практичным и осуществимым шагом, который в идеале должен применяться ко всем промышленным системам. Как основу для аудита, а также возможных корпоративных, отраслевых или даже национальных стандартов можно рекомендовать концептуальный документ EASA: «First usable guidance for level 1 machine learning applications» [15].

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

СПИСОК ЛИТЕРАТУРЫ

1. Robust and Verified Deep Learning Group. https://deepmindsafetyresearch.medium.com/towards-robust-and-verified-ai-specification-testing-robust-training-and-formal-verification-69bd1bc48bda (accessed 24.07.2023).
2. Madry Lab. https://people.csail.mit.edu/madry/6.S979/files/lecture_4.pdf (accessed 24.07.2023).

3. Game Changers. https://www.cbinsights.com/research/report/game-changing-technologies-2022/ (accessed 24.07.2023).
4. An In-Depth Guide to Help You Start Auditing Your AI Models. https://censius.ai/blogs/ai-audit-guide (accessed 24.07.2023).
5. IEEE Standard for Software Reviews and Audits // IEEE Std 1028-2008. 2008. P. 1–53; doi: 10.1109/IEEESTD.2008.4601584.
6. The IIA's Artificial Intelligence Auditing Framework. https://www.theiia.org/en/content/articles/global-perspectives-and-insights/2017/the-iias-artificial-intelligence-auditing-framework-practical-applications-part-ii/ (accessed 24.07.2023).
7. Realize the Full Potential of Artificial Intelligence. https://www.coso.org/Shared%20Documents/Realize-the-Full-Potential-of-Artificial-Intelligence.pdf (accessed 24.07.2023).
8. AI TRiSM. https://www.gartner.com/en/information-technology/glossary/ ai-trism (accessed 24.07.2023).
9. *Намиот Д. Е., Ильюшин Е. А., Пилипенко О. Г.* Доверенные платформы искусственного интеллекта // Intern. J. Open Inf. Technol. 2022. V. 10, No. 7. P. 119–127.
10. *Намиот Д. Е.* Схемы атак на модели машинного обучения // Intern. J. Open Inf. Technol. 2023. V. 11, No. 5. P. 68–86.
11. Equipment, Systems, and Installations. https://www.law.cornell.edu/cfr/text/14/25.1309.
12. *Ruparelia N. B.* Software Development Lifecycle Models // ACM SIGSOFT Softw. Eng. Notes. 2010. V. 35, No. 3. P. 8–13.
13. EASA AI Task Force, Daedalean AG. Concepts of Design Assurance for Neural Networks (CoDANN). EASA, 2020.
14. *Dmitriev K., Schumann J., Holzapfel F.* Towards Design Assurance Level C for Machine-Learning Airborne Applications // 41st Digital Avionics Systems Conference (DASC), 2022 IEEE/AIAA, Portsmouth, VA, USA, 2022. P. 1–6; doi: 10.1109/DASC55683.2022.9925741.
15. First Usable Guidance for Level 1 Machine Learning Applications. https://www.easa.europa.eu/en/newsroom-and-events/news/easa-releases-its-concept-paper-first-usable-guidance-level-1-machine-0.