ВЕСТНИК РОССИЙСКОЙ АКАДЕМИИ НАУК, 2024, том 94, № 3, с. 200–209

НАУЧНАЯ СЕССИЯ ОБЩЕГО СОБРАНИЯ ЧЛЕНОВ РАН

ДОВЕРЕННЫЙ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

© 2024 г. А.И. Аветисян[a,*]

[a] *Институт системного программирования им. В.П. Иванникова РАН, Москва, Россия*

**E-mail: arut@ispras.ru*

Поступила в редакцию 31.01.2024 г.
После доработки 02.02.2024 г.
Принята к публикации 06.03.2024 г.

В статье рассматривается проблема создания доверенных технологий искусственного интеллекта. Современный искусственный интеллект основан на машинном обучении и нейронных сетях и уязвим для ошибок. Попытки установить стандарты разработки доверенных технологий искусственного интеллекта пока не увенчались успехом. Для обеспечения его безопасности требуются соответствующая научно-технологическая база, новые инструменты и методики противодействия атакам. Автор приводит результаты, полученные Исследовательским центром доверенного искусственного интеллекта Института системного программирования им. В.П. Иванникова РАН, а также предлагает модель работы, которая может способствовать достижению технологической независимости и обеспечить долгосрочное устойчивое развитие в этой области.
Статья подготовлена на основе доклада, заслушанного на научной сессии Общего собрания членов РАН 12 декабря 2023 г.

Ключевые слова: доверенный искусственный интеллект, кибербезопасность, уязвимости, регулирование, репозиторий доверенного ПО, слабый искусственный интеллект, нейросети, федеративное обучение, машинное обучение.

DOI: 10.31857/S0869587324030039, **EDN:** GHFCAB

Термин “доверенный искусственный интеллект” вошёл в оборот относительно недавно и в последние несколько лет применяется в отношении защищённых и безопасных технологий искусственного интеллекта (ИИ). Чтобы понять, как сделать их именно такими, обратимся вначале к истории информационно-коммуникационных технологий (ИКТ) в целом, поскольку методики и инструменты обеспечения их безопасности появились задолго до систем искусственного интеллекта.

Активное развитие таких технологий началось в 1940-е годы. В СССР точкой отсчёта стал 1948 год: 29 июня был основан Институт точной механики и вычислительной техники (ИТМиВТ) Академии наук СССР; 4 декабря член-корреспондент АН СССР И.С. Брук и инженер-конструктор Б.И. Рамеев подали заявку на изобретение автоматической цифровой вычислительной машины; 19 декабря учреждено Специальное конструкторское бюро № 245 (впоследствии Научно-исследовательский центр электронной вычислительной техники). В 1949 г. основана кафедра вычислительной математики механико-математического факультета МГУ имени М.В. Ломоносова. Началось производство первых электронно-вычислительных машин: МЭСМ (1951), “Стрела” (1953), М-100 (1958), “Днепр” (1961). В 1967 г. была создана БЭСМ-6 – первый суперкомпьютер в СССР. Главным конструктором выступил академик С.А. Лебедев (ИТМиВТ); в том же институте была предложена и операционная система для БЭСМ-6 – Д-68 (разработчики – Л.Н. Королёв, А.Н. Томилин, В.П. Иванников и другие).

АВЕТИСЯН Арутюн Ишханович – академик РАН, заместитель президента РАН, директор ИСП РАН.

Научно-технологическая основа информационно-коммуникационных технологий, заложенная в те годы как в нашей стране, так и за рубежом, позволила обеспечить их бурное развитие в последующие десятилетия. К настоящему времени раз-

меры программных систем значительно выросли, многократно усложнились их разработка и сборка. Например, размер дистрибутива операционной системы Debian на базе Linux в 2019 г. превысил отметку в миллиард строк кода, а сейчас составляет почти 2 млрд. Размер популярных фреймворков машинного обучения PyTorch и TensorFlow превышает 7 млн и 10 млн строк соответственно. В мире постоянно растёт объём больших данных: в 2022 г. он составлял 97 зеттабайт (1 зеттабайт = 1 миллиард терабайт), а в 2025 г., по некоторым прогнозам [1], достигнет 180 зеттабайт. Число проектов на веб-сервере GitHub в 2023 г. составило 420 млн, а число разработчиков превысило 100 млн человек. В этом сложном мире становится всё труднее обеспечивать необходимые качества современного программного обеспечения (ПО): эффективность, продуктивность и доверенность, последняя включает в себя и безопасность.

Основной источник функциональных, архитектурных и других уязвимостей в ПО и аппаратуре – ошибки, которые очень трудно отличить от закладок и недокументированных возможностей. Информацией об уязвимостях могут воспользоваться хакеры (просто из интереса), преступники и представители спецслужб, причём эксплуатация некоторых уязвимостей не составляет труда для более или менее опытного человека. Известный пример – уязвимость Heartbleed: за счёт переполнения буфера атакующий может получить доступ к участку оперативной памяти с незашифрованными логинами и паролями. Версия библиотеки OpenSSL с этой уязвимостью вышла в марте 2012 г., а саму уязвимость удалось обнаружить только спустя два года. На момент обнаружения уязвимыми были полмиллиона веб-сайтов.

Повсеместное внедрение информационно-коммуникационных технологий и наличие серьёзных проблем с их безопасностью привели к необходимости законодательного регулирования процессов разработки. Стало ясно, что недостаточно использовать классические методы защиты (защита по периметру, проверка доступа, антивирусы и др.), так как уязвимости всё равно остаются внутри программного обеспечения. Необходим поиск новых моделей, методов и технологий в области анализа и трансформации программ, направленных на устранение максимального количества уязвимостей в исполняемом коде на этапе разработки ПО, а также поддержание устойчивости готового программного обеспечения, затруднение негативных воздействий существующих уязвимостей или смягчение последствий их воздействий.

Первые стандарты разработки безопасного ПО были предложены в США. В 1999 г. в Национальном институте стандартов и технологий (NIST) началась работа над стандартами Common Criteria. В 2004 г. компания Microsoft сделала достоянием гласности примерный регламент цикла разработки безопасного программного обеспечения (рис. 1). В России в 2016 г. был принят соответствующий ГОСТ Р 56939-2016 [2]; завершается подготовка ГОСТов по отдельным процессам и инструментам (например, по статическому анализу и по безопасному компилятору, оба вводятся в действие в апреле 2024 г.). В Евросоюзе в 2019 г. был принят Закон о кибербезопасности [3] – система сертификации ПО, сервисов и процессов. В Китае в 2023 г. утверждены 19 стандартов кибербезопасности, предложенных Техническим комитетом 260, отвечающим за национальную безопасность в этой сфере.

В связи с появлением в последние два десятилетия нормативной базы создаётся и соответствующий инструментарий. В Институте системного программирования РАН поиск в этом направлении ведётся с 2002 г. и к настоящему моменту создан полный стек (от англ. stack – стопка) технологий – от безопасного компилятора до анализа бинарного кода, обеспечивающий технологическую независимость в этой области.

Создание инструментов РБПО возможно только на основе результатов фундаментальных исследований, поэтому в конце 2018 г. президиумом РАН было принято решение о необходимости развития нового научного направления – кибербезопасности [4]. В 2021 г. в номенклатуру специальностей, по которым присуждается учёная степень кандидата или доктора физико-математических наук, приказом Минобрнауки России включена “кибербезопасность” [5]. Она объединяет такие направления исследований, как анализ и систематизация уязвимостей, моделирование политики информационной безопасности, угроз и атак, методы, алгоритмы и средства пострелизно-

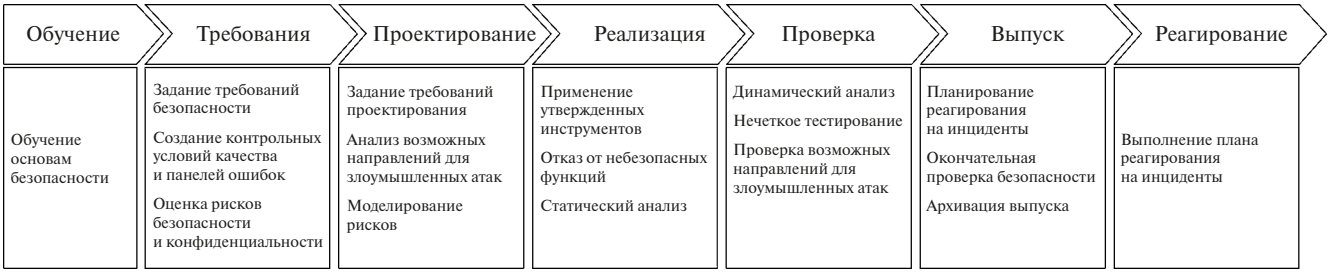


Рис. 1. Жизненный цикл разработки безопасного программного обеспечения (Microsoft)

го глубокого анализа защищённости программного обеспечения и многое другое.

Отметим, что так развивалась кибербезопасность программных систем, не использующих искусственный интеллект. С появлением и повсеместным внедрением ИИ многое изменилось.

История искусственного интеллекта восходит к 1950-м годам. Сам термин появился в 1956 г.; он был предложен американским математиком Дж. Маккарти на семинаре в Дартмутском университете (США) [6]. Цель исследований в этой области – позволить компьютеру выполнять интеллектуальные действия, ранее доступные только человеку (например, вести простой диалог или играть в шахматы). Изначально задачу видели в построении модели объекта автоматизации, а алгоритм решения основывался на правилах работы с моделью. В конце XX в. поиски в этом направлении активизировались, и основным стал метод решения задачи по аналогии: на этапе обучения алгоритм получал примеры вопросов и правильных ответов, а на этапе применения – получал вопросы и выдавал ответы. Машинное обучение довольно быстро совершенствовалось. В 1997 г. компьютер Deep Blue американской компании IBM выиграл шахматный матч из шести партий с Г. Каспаровым (тогда действующим чемпионом мира по версии Профессиональной шахматной ассоциации). В 2002 г. разработчики представили первый робот-пылесос. В 2010 г. была сформирована база данных ImageNet, объединившая 14 млн изображений 20 тыс. категорий. В 2011 г. суперкомпьютер Watson компании IBM одержал победу в интеллектуальном телешоу Jeopardy! В том же году в составе программного обеспечения смартфонов компании Apple появился первый виртуальный голосовой помощник Siri. В 2016 г. интернет-сервис Google Translate начал использовать нейронный машинный перевод для восьми языков, а с 2022 г. стали появляться и совершенствоваться “большие языковые модели” – Open AI ChatGPT, в их числе YandexGPT2, RuGPT3 (Сбер) и другие. Таким образом, можно утверждать, что системы современного искусственного интеллекта базируются на различных видах машинного обучения и нейросетях.

Столь впечатляющие успехи ИИ за последние 10–15 лет связаны с ускоренным развитием суперкомпьютерных мощностей и появлением огромных массивов данных (рис. 2). Основу этих успехов заложили представители классической математики, в том числе отечественные, что признано на мировом уровне. Так, исследования академика АН СССР А.Н. Колмогорова [7] помогли обосновать решения задач, возникающих в процессе машинного обучения, а труды академика АН СССР и РАН А.Н. Тихонова [8] заложили основу методов вычислительной математики, позволив существенно ограничить суперкомпьютерные ресурсы, необходимые для решения задач искусственного интеллекта.

Научные подходы к ИИ, методы и алгоритмы активно развиваются, однако не следует забывать, что мы живём в мире так называемого “слабого искусственного интеллекта” (*англ.* Weak AI, или Narrow AI). Он основан на машинном обучении и нейронных сетях, способен извлекать информацию из ограниченного набора данных, а в случае их искажения может выдать необъективный (неэтичный, дискриминационный) результат. “Слабый ИИ” уязвим для предвзятости, необъективности и ошибок. “Сильный искусственный интеллект” (*англ.* Strong AI, General AI), способный решать задачи на уровне интеллектуальных возможностей человека, строить стратегии и функционировать в условиях неопределённости, пока существует только в теории, так как отсутствуют методы и алгоритмы, на основе которых он мог бы функционировать. Существующие направления исследований развиваются в рамках “слабого ИИ”. Устройства с “сильным ИИ”, можно предположить, удастся реализовать лишь в очень отдалённой перспективе.

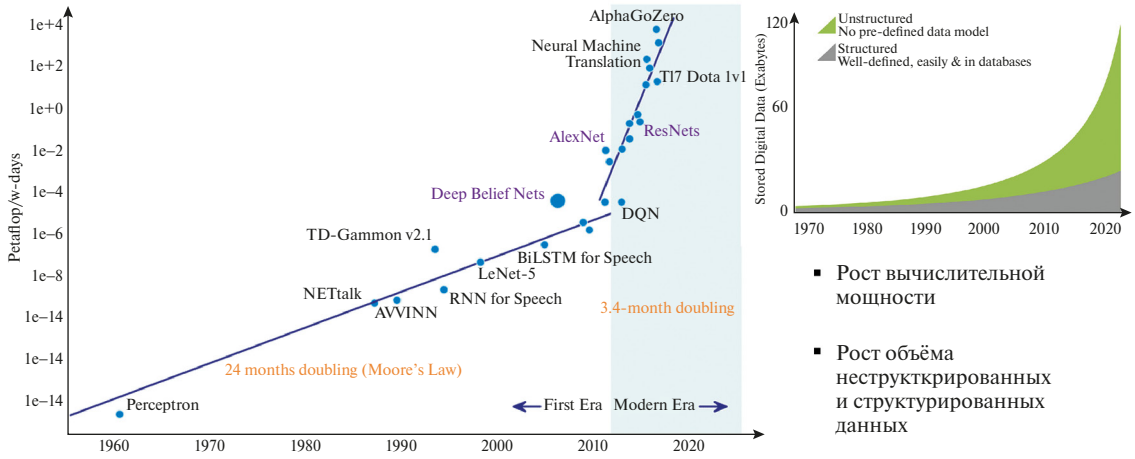


Рис. 2. Вычислительные ресурсы и большие данные – двигатели развития систем искусственного интеллекта

Наличие уязвимостей в системах искусственного интеллекта делает их работу небезопасной или неэтичной на бытовом уровне. Например, использование дискриминирующих алгоритмов привело к тому, что модель для выбора кандидатов на должности разработчиков в компании Amazon отдавала предпочтение мужчинам (модель была обучена на данных за десятилетний период, на протяжении которого основная часть резюме претендентов на эти вакансии поступала от мужчин) [9]. Ряд серьёзных проблем связан с практикой эксплуатации беспилотных автомобилей. В частности, они подвержены так называемым состязательным атакам: если нанести на дорожный знак невидимый для глаза рисунок, то компьютерное зрение может не распознать этот знак, что чревато аварией. Возможна и неадекватная реакция беспилотного автомобиля на дорожно-транспортное происшествие. В 2023 г. в США автомобиль Cruise наехал на пешехода, протащил его 6 м и остановился, не съезжая с пострадавшего, который в результате получил серьёзные травмы. После этого инцидента 950 беспилотных машин Cruise компании-изготовителю пришлось отозвать для обновления программного обеспечения [10]. Отсутствие обязательных цифровых меток на сгенерированном искусственным интеллектом контенте также может привести к печальным последствиям. Например, в 2023 г. в продажу поступили сгенерированные ИИ книги о грибах, но использовать их как практическое пособие не рекомендуется, так как в них обнаружены серьёзные содержательные ошибки [11].

В целом можно констатировать, что в настоящее время искусственный интеллект представляет собой сложный технологический комплекс с отсутствующей субъектностью. Исходя из этого представления в мире формируется юридическая основа его регулирования (как ранее формировались нормы регулирования других технологий). Так, в 2022 г. в США был опубликован проект “Билля о правах” ИИ [12], предлагаемого компаниями, общественными организациями и экспертными группами. В нём формулируются пять принципов создания и использования искусственного интеллекта, в числе которых разработка безопасных и эффективных систем; отсутствие алгоритмической дискриминации; обеспечение конфиденциальности данных и др. В 2023 г. в США на государственном уровне был одобрен ещё один важный документ – Executive Order on Safe, Secure, and Trustworthy AI (Указ о безопасном, защищённом и доверенном искусственном интеллекте) [13], который устанавливает новые стандарты в сфере безопасного развития ИИ и содержит поручения для ведомств и разработчиков. Например, разработчики ряда значимых систем обязаны делиться с правительством результатами тестов на безопасность продуктов; кроме того, сгенерированный ИИ контент должен маркироваться специальными цифровыми метками. Последнюю инициативу разделяют и ведущие компании в области ИИ (OpenAI, Meta.Platforms, Alphabet и др.), которые уже обязались реализовать систему цифровых водяных знаков для всех форм синтезированного контента. В том же году Агентство национальной безопасности США объявило о создании Центра безопасности искусственного интеллекта, а Национальный научный фонд – о создании семи исследовательских институтов ИИ (один из них – Institute for Trustworthy AI in Law & Society, Институт доверенного искусственного интеллекта в юридических и общественных науках).

В Евросоюзе в 2020 г. появился документ “Whitepaper on AI: a European approach to excellence and trust” (“Положение по ИИ: европейский подход к качеству и доверию”) [14], который объясняет важность ИИ и призывает к его оптимизации и развитию экосистемы, в частности, инициирует работу над нормативной базой и определяет ключевые требования к ней. В их числе – безопасные обучающие данные без дискриминации; надёжность и воспроизводимость; контроль человека над ИИ; защита биометрических данных. В 2023 г. был предварительно одобрен EU AI Act (Закон Евросоюза об искусственном интеллекте) [15], который предлагает разделить все системы с ИИ на три категории: с неприемлемыми рисками, высокими и низкими. Первые подлежат запрету, вторые должны быть приведены в соответствие с “определёнными юридическими требованиями”, третьи регулированию не подлежат. В число запрещённых систем попадёт сбор биометрии в общественных местах, а также социальный скоринг.[1]

Если же говорить об общемировых подходах, то 19 мая 2023 г. на саммите глав государств “Большой семёрки” в Хиросиме был принят специальный документ для содействия развитию передовых систем искусственного интеллекта на глобальном уровне. 30 октября 2023 г. те же лидеры поддержали “Международный кодекс поведения” и “Руководящие принципы для организаций, разрабатывающих передовые системы ИИ”. Например, в кодексе заявлено, что таким организациям следует “присоединиться к процессам разработки, продвижения и принятия, где это необходимо, общих стандартов, инструментов, механизмов и лучших практик для обеспечения безопасности, надёжности и достоверности передовых систем ИИ”. А разработчикам следует “добиваться полной прозрачности – документировать используемые наборы данных, процессы и решения, принятые в ходе разработки системы” [16].

Регуляторика активно развивается и в нашей стране. Россия здесь – в числе лидеров. В 2019 г.

[1] Социальный скоринг – вид оценки платёжеспособности заёмщика банка по его социальным характеристикам и прогнозирования поведения клиента с помощью анализа его присутствия в социальных сетях.

Указом Президента РФ № 490 была принята Национальная стратегия развития искусственного интеллекта на период до 2030 года [17]. В 2021 г. десятки компаний подписали Кодекс этики в сфере искусственного интеллекта [18], разработанный при участии Минэкономразвития России, Аналитического центра при Правительстве РФ, а также около 500 экспертов академического и бизнес-сообщества. Кодекс подчёркивает приоритет прав человека, ответственность человека за действия ИИ, потребность в безопасности и защищённости данных, а также необходимость разработки безопасных технологий. Появляются первые ГОСТы, например, ГОСТ 59921 “Системы ИИ в клинической медицине” (принят в 2022 г.; устанавливает требования к клиническому тестированию ИИ-систем на основе глубоких нейронных сетей) [19].

Таким образом, в мире предпринимаются попытки ограничить внедрение небезопасных технологий и установить стандарты разработки безопасных, как это было ранее сделано для обычного программного обеспечения. Однако на практике общепринятых подходов и инструментов разработки безопасных технологий искусственного интеллекта пока не существует. Те наработки, которые появились за последние десятилетия в области обычного ПО, только частично применимы к ИИ. Это связано с датацентричностью искусственного интеллекта – основной источник его уязвимостей связан с данными (рис. 3).

Исходный код инфраструктур машинного обучения можно анализировать с помощью традиционных средств, но с данными и моделями машинного обучения всё гораздо сложнее. Например, данные могут быть отравленными, а предобученные модели – содержать закладки или вредоносное программное обеспечение. Полностью проверить модель на уязвимости традиционными способами, на уровне обычного ПО, невозможно, потому что основная информация содержится в параметрах модели, то есть в данных, а не в исполняемом коде. Нужны новые технологии.

Именно поэтому в 2021 г. при поддержке Минэкономразвития России в Институте системного программирования им. В.П. Иванникова РАН учреждён Исследовательский центр доверенного искусственного интеллекта, цель которого – создать платформу, которая объединяла бы программные инструменты и методики для противодействия принципиально новым угрозам, возникающим на всех этапах жизненного цикла технологий ИИ. Здесь уже получены результаты по ряду актуальных направлений. Это касается, например, противодействия отравлению данных. С помощью добавления на изображения незаметных триггеров можно добиться того, что компьютерное зрение будет распознавать эти изображения с недопустимыми искажениями (рис. 4) [20, 21]. Существуют разные подходы к решению этой проблемы. Предлагается, в частности, чистка зловредных нейронов (по отравленным тренировочным данным можно найти редко активируемые нейроны; скорее всего, именно они кодируют триггер) [22] или же обнаружение и удаление триггера с помощью объяснимого ИИ, а затем применение генеративной модели для восстановления изображения [23]. В Исследовательском центре доверенного ИИ ИСП РАН разрабатываются защитные алгоритмы от незаметного отравления изображений. Кроме того, ведутся исследования по противодействию “состязательным атакам”, в ходе их разработана защищённая система компьютерного зрения с дообученной моделью, которая позволя-

Рис. 3. Появление уязвимостей в системах с ИИ

$$\theta' = \arg\min_{\theta} E_{(x, y) \sim \mathcal{D}}[(1 - \lambda) \cdot \mathcal{L}(F(x; \theta), y) + \lambda \cdot \mathcal{L}(F(x + t; \theta), y_T)]$$

Атака №1: Заставить модель присвоить заведомо ложную целевую метку при наличии триггера во входных признаках https://arxiv.org.abs.1708.06733

Свойства успешных атак через закладки:
- хорошая точность на чистых тестовых данных
- сбой модели на отравленных данных
- атака незаметна: малая доля отравленных данных, малый размер триггера

Атака №2: Заставить модель присвоить выбранную целевую метку (рыба) на заданном множестве целей из другого класса (собака) https://arxiv.org.abs.1804.00792

$$\mathbf{p} = \arg\min_{\mathbf{x}} \|f(\mathbf{x}) - f(\mathbf{t})\|_2^2 + \beta\|\mathbf{x} - f\mathbf{b}\|_2^2$$

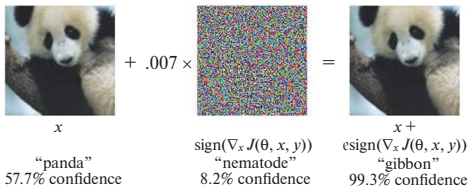
Рис. 4. Отравленные данные: вставка закладок

ет системе распознавать человека даже в футболке со зловредным стикером (рис. 5).

Здесь важно отметить, что атакам подвержены и закрытые системы (так называемые атаки “чёрного ящика”). Это связано с тем, что не исключена возможность перенесения состязательных примеров на другие модели и наборы данных, поэтому идею безусловной защищённости закрытых систем нельзя признать верной. Для обеспечения полноценной безопасности необходимо располагать соответствующей научно-технологической базой и созданными на её основе программными инструментами и методиками противодействия атакам.

Важность этой задачи осознаётся повсеместно. С 2017 г. реализуется множество проектов по развитию доверенного ИИ. Например, такие проекты консорциума “The Linux Foundation”, как Adversarial Robustness Toolbox, AI Explainability 360 AI Fairness

Незаметная градиентная атака на классификаторы изображений: максимизация ошибки через изменение входа

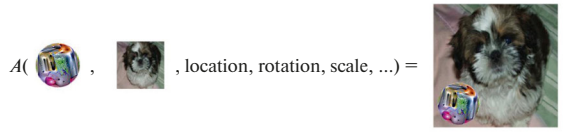


$$x' = \mathrm{Proj}_{[0, 1]^d}(x + \varepsilon\mathrm{sign}(\nabla_x J(\theta, x, y))) \text{ без целевого класса}$$

$$x' = \mathrm{Proj}_{[0, 1]^d}(x - \varepsilon\mathrm{sign}(\nabla_x J(\theta, x, y'))) \text{ направленная (на класс } y')$$

$$x' = \mathrm{Proj}_{B_\varepsilon \cap [0, 1]^d}(x_k + \alpha\mathrm{sign}(\nabla_x J(\theta, x_k, y))) \text{ итерационная (сильная)}$$

Состязательные патчи: выучить универсальный зловредный стикер



$$\hat{p} = \arg\min_{p} E_{x \sim X, t \sim T, l \sim L}[(\log \Pr(\hat{y}|A(p, x, l, t)]$$

Изменение алгоритма

Состязательные футболки (справа реализована защита)

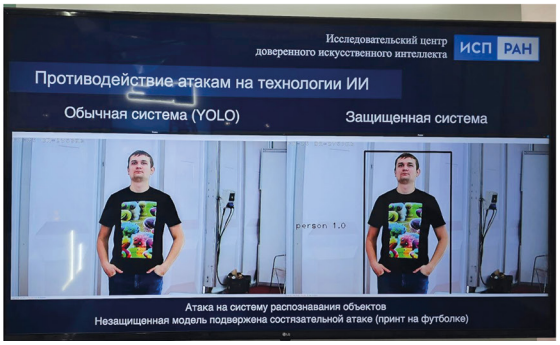


Рис. 5. Противодействие состязательным атакам

360 и другие. Консорциум поддерживает также проекты других компаний и организаций, нацеленные на поиск уязвимостей моделей и повышение безопасности их использования (NextAttack, University of Virginia; Foolbox, University of Tuebingen; CleverHans, CleverHans Project), определение смещения модели (Aequitas, Chicago University; Fairlearn, Microsoft) и др. Однако несмотря на масштабные исследования, до сих пор не существует типового сценария разработки, обеспечивающего и возможность непрерывного выпуска оттестированных новых версий продукта (CI/CD – непрерывная интеграция и непрерывная поставка), и применение инструментов противодействия атакам. Такие сценарии уже много лет применяются при разработке обычных программ без ИИ.

Существуют и другие проблемы, связанные с датацентричностью. Например, не во всех областях возможен обмен данными согласно существующему законодательству (это особенно актуально для медицины). В таких случаях возможно применение метода так называемого федеративного (распределённого машинного) обучения, позволяющего обучать модели на нескольких устройствах без обмена образцами данных. Федеративное обучение делится на горизонтальное (например, необходимость анализа ЭКГ большой группы людей) и вертикальное, с учётом разных сфер социальной активности одного и того же человека (медицина, финансы и др.). Ещё одна задача – упрощение коммуникаций, поскольку они могут занимать до 95% времени обучения ИИ. Её можно решить несколькими способами, в том числе сжатием посылок, пропуском раундов общения, временным упрощением модели и разделением её на составляющие, что позволяет кратно уменьшить коммуникационные затраты и время обучения.

В федеративном обучении существуют две основные проблемы: жёсткие требования к стоимости коммуникаций (то есть к объёму передаваемых данных) и конфиденциальности информации. В их решении, а также для оптимальной квантизации сигналов широко используются результаты из многомерной геометрии, полученные в 1970-х годах академиком РАН Б.С. Кашиным. Продемонстрировали свою эффективность и так называемые жадные алгоритмы [24]. В Исследовательском центре доверенного ИИ ИСП РАН актуальные задачи федеративного обучения решает коллектив под руководством доктора физико-математических наук А.В. Гасникова [25–27]. Полученные результаты позволяют снизить коммуникационные затраты максимум в 15 раз, а процесс обучения сократить в 5 раз [28, 29].

Важна также задача, касающаяся совершенствования технологий маркирования контента цифровыми водяными знаками. Она актуальна как для контента, созданного человеком (в целях защиты авторских прав и отслеживания распространения интеллектуального продукта) [30], так и для синтезированного (обнаружение сгенерированного текста [31], дипфейков в аудио- и видеопотоках и изображениях [32]), а также для глубоких нейронных сетей, защиты от анонимной кражи обучающих наборов данных и обученных моделей [33–35]). В США уже готовится законодательная база, обязывающая маркировать ИИ-контент водяными знаками. Вероятнее всего по этому пути пойдут и другие страны, в том числе Россия, поэтому столь актуально развитие соответствующих отечественных технологий.

Необходимо отметить, что разработка даже очень качественных технологий доверенного ИИ далеко не гарантирует обеспечение технологической независимости и долгосрочного устойчивого развития в этой области. Нужны модели совместной работы, которые в условиях коллаборативной экономики способствуют кратному увеличению производительности труда. Целесообразно сформировать экосистему, которая позволит отдельным компаниям перейти от конкуренции к кооперации, совместно создавать и контролировать репозиторий доверенного программного обеспечения. Такая модель работы, представленная на рисунке 6, уже используется ИСП РАН и партнёрами в Центре доверенного системного ПО. Вокруг него сформирован консорциум; более 50 компаний (Лаборатория Касперского, Postgres Professional и др.) и вузов (МГТУ им. Н.Э. Баумана, ЧГУ и др.) совместно ведут исследования по повышению безопасности ОС Linux и системных компонентов (например, OpenSSL, QEMU, libvirt). Около 300 исправлений уже признаны международным сообществом и внесены в основные ветки проектов.

Реализация данной модели работы показала, что Академия наук является ключевой структурой, которая может формировать такие сообщества, объединяя государство, бизнес и научные институты. Создание репозитория доверенных решений было поддержано на стратегической сессии “Развитие искусственного интеллекта” под председательством главы правительства РФ М.В. Мишустина в сентябре 2023 г.

В настоящее время существующую экосистему необходимо масштабировать в целях создания технологий доверенного ИИ. Для этого целесообразно:

• развернуть комплексную научно-техническую программу в целях поиска перспективных подходов к обеспечению кибербезопасности, создания интеллектуальных технологий и инструментальных средств, обеспечивающих минимизацию угроз безопасности, связанных с ошибками, включая новые виды уязвимостей и рисков, которые характерны для технологий ИИ;

• определить, что важным механизмом реализации указанной программы служит создание репозиториев доверенных средств ИИ и инструментов обеспечения доверия;

• развивать нормативное регулирование ИИ в Российской Федерации, которое в зависимости

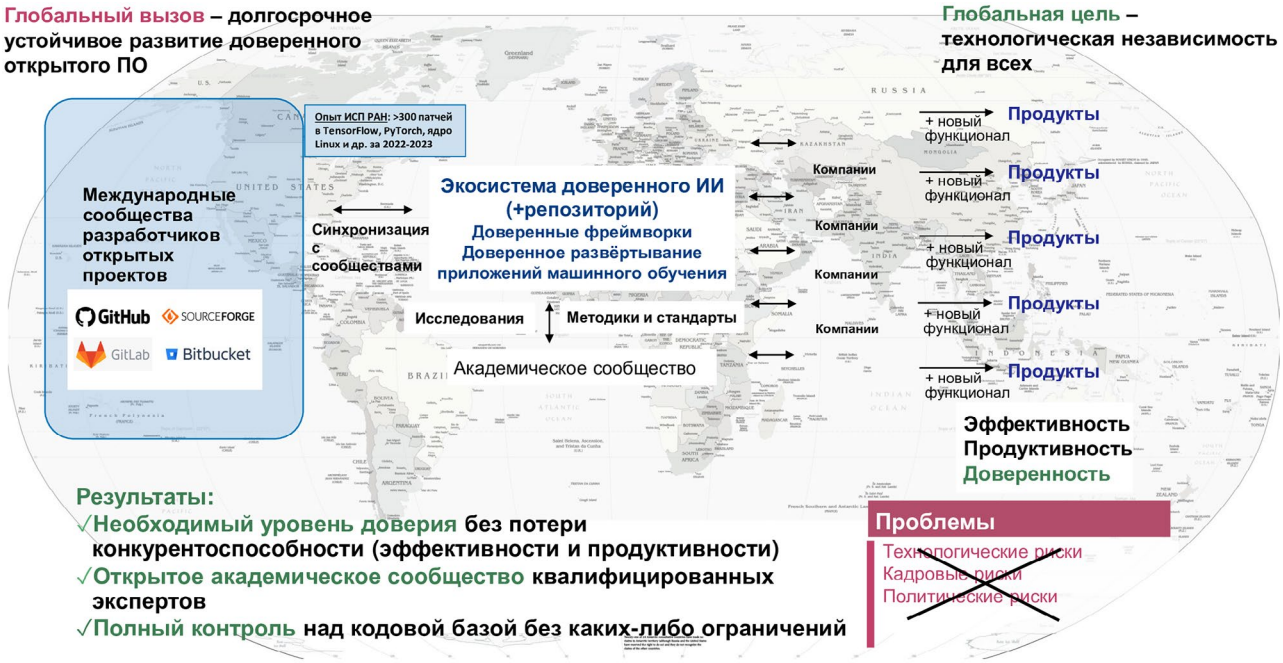


Рис. 6. Пример глобальной модели долгосрочного развития

от применения предусматривает как возможности саморегулирования, так и обязательную государственную сертификацию на основе высокотехнологичных программных средств;

• расширить подготовку специалистов высшей квалификации по специальности “Кибербезопасность”.

Подготовка высококвалифицированных профильных специалистов, создание нормативной базы и репозиториев доверенного ПО, а также разработка новых подходов и инструментов обеспечения доверия к ИИ будут способствовать достижению технологической независимости нашей страны, обеспечению её безопасного будущего.

ЛИТЕРАТУРА

1. Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025. https://www.statista.com/statistics/871513/worldwide-data-created/
2. ГОСТ Р 56939-2016 “Защита информации. Разработка безопасного программного обеспечения. Общие требования”.

   GOST R 56939-2016 “Information protection. Secure software development. General requirements”. https://docs.cntd.ru/document/1200135525
3. Regulation (EU) 2019/881 of the European Parliament and of the Council on ENISA (the European Union Agency for Cybersecurity) and on information and communications technology cybersecurity certification and repealing Regulation (EU) No 526/2013 (Cybersecurity Act). https://eur-lex.europa.eu/eli/reg/2019/881/oj
4. Постановление президиума Российской академии наук “О мерах по развитию системного программирования как ключевого направления противодействия киберугрозам”.

   Act of The Presidium of RAS. On supporting development of system programming as a key direction for countering cyberthreats. https://www.ras.ru/presidium/documents/directions.aspx?ID=1f5522e9-ff25-4af0-a49a-6a5675525597
5. Приказ Минобрнауки России № 118 от 24 февраля 2021 года.

   Order of the Ministry of Science and Education of Russia no. 118 dated February 24, 2021. https://vak.minobrnauki.gov.ru/uploader/loader?type=1&name=91506173002&f=7892
6. *Zhang Y., Nauman U.* Deep Learning Trends Driven by Temes: A Philosophical Perspective // IEEE Access. January 2020. V. 8. P. 96587–196599. http://dx.doi.org/10.1109/ACCESS.2020.3032143
7. *Колмогоров А.Н.* О представлении непрерывных функций нескольких переменных в виде суперпозиций непрерывных функций одного переменного и сложения // Докл. АН СССР. 1957. Т. 114. № 5. С. 953–956. *Kolmogorov A.N.* On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition // Dokl. Akad. Nauk SSSR.1957. V. 114. № 5. P. 953–956. (In Russ.)
8. *Tychonoff A.* Ein Fixpunktsatz // Mathematische Annalen. 1935. V. 111. P. 767–776. https://link.springer.com/article/10.1007/BF01472256

9. Insight – Amazon scraps secret AI recruiting toll that showed bias against women. https://www.reuters.com/article/idUSKCN1MK0AG/
10. Cruise robotaxi service hid severity of accident, California officials claim. https://www.theguardian.com/business/2023/dec/04/california-cruise-robotaxi-san-francisco-accident-severity
11. Mushroom pickers urged to avoid foraging books on Amazon that appear to be written by AI. https://www.theguardian.com/technology/2023/sep/01/mushroom-pickers-urged-to-avoid-foraging-books-on-amazon-that-appear-to-be-written-by-ai
12. Blueprint for an AI Bill of Rights. https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf
13. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/
14. White Paper on Artificial Intelligence: a European approach to excellence and trust. https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en
15. EU AI Act. https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf
16. Hiroshima Process International Code of Conduct for Advanced AI Systems. https://digital-strategy.ec.europa.eu/en/library/hiroshima-process-international-code-conduct-advanced-ai-systems
17. Указ Президента Российской Федерации “О развитии искусственного интеллекта в Российской Федерации”. Национальная стратегия развития искусственного интеллекта на период до 2030 года.
Decree of the President of Russia “On developing artificial intelligence in Russia”, National strategy of developing artificial intelligence for the period until 2030. http://static.kremlin.ru/media/events/files/ru/AH4x6HgKWANwVtMOfPDhcbRpvd1HCCsv.pdf
18. Кодекс этики в сфере искусственного интеллекта.
Ethics Codex on Artificial Intelligence. https://ethics.a-ai.ru/assets/ethics_files/2023/05/12/Кодекс_этики_20_10_1.pdf
19. ГОСТ Р 59921.2-2021 “Системы искусственного интеллекта в клинической медицине. Часть 2. Программа и методика технических испытаний”.
GOST R 59921.2-2021 “Artificial Intelligence Systems in Clinical Medicine. Part 2. Program and methodology of technical validation”. https://docs.cntd.ru/document/1200181991
20. Gu *T., Dolan-Gavitt B., Garg S.* BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. https://arxiv.org/abs/1708.06733
21. *Shafahi A., Huang W., Najibi M. et al.* Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks. https://arxiv.org/abs/1804.00792v2
22. *Liu K., Dolan-Gavitt B., Garg S.* Fine-Pruning: Defending Against Backdooring Attacks on Deep Neural Networks. https://arxiv.org/abs/1805.12185
23. *Bao Gia Doan, Abbasnejad E., Ranasinghe D.C.* Februus: Input Purification Defense Against Trojan Attacks on Deep Neural Network Systems. https://arxiv.org/abs/1908.03369
24. *Temlyakov V.* Greedy Approximation. Cambridge University Press, 2011.
25. *Beznosikov A., Horváth S., Richtárik P., Safaryan M.* On biased compression for distributed learning // Journal of Machine Learning Research. 2023. V. 24(276). P. 1–50.
26. *Gorbunov E., Rogozin A., Beznosikov A. et al.* Recent theoretical advances in decentralized distributed convex optimization // High-Dimensional Optimization and Probability: With a View Towards Data Science. Cham: Springer International Publishing, 2022. P. 253–325.
27. *Metelev D., Rogozin A., Kovalev D., Gasnikov A.* Is consensus acceleration possible in decentralized optimization over slowly time-varying networks? // International Conference on Machine Learning, 2023. PMLR. P. 24532–24554.
28. *Beznosikov A. et al.* Distributed methods with compressed communication for solving variational inequalities, with theoretical guarantees // Advances in Neural Information Processing Systems. 2022. V. 35. P. 14013–14029.
29. *Beznosikov A., Gasnikov A.* Similarity, Compression and Local Steps: Three Pillars of Efficient Communications for Distributed Variational Inequalities. https://arxiv.org/abs/2302.07615v1
30. *Yakushev A., Markin Yu., Obydenkov D. et al.* Docmarking: Real-Time Screen-Cam Robust Document Image Watermarking // 2022 Ivannikov Ispras Open Conference (ISPRAS), IEEE. P. 142–150.
31. *Sankar Sadasivan V., Kumar A., Balasubramanian S. et al.* Can AI-Generated Text be Reliably Detected? https://arxiv.org/pdf/2303.11156.pdf
32. *Yu N., Skripniuk V., Abdelnabi S., Fritz M.* Artificial Fingerprinting for Generative Models: Rooting Deepfake Attribution in Training Data. https://openaccess.thecvf.com/content/ICCV2021/papers/Yu_Artificial_Fingerprinting_for_Generative_Models_Rooting_Deepfake_Attribution_in_Training_ICCV_2021_paper.pdf
33. *Adi Y., Baum C., Cisse M. et al.* Turning Your Weakness Into a Strength: Watermarking Deep Neural Networks by Backdooring. Proceedings of the 27th USENIX Security Symposium. August 15–17, 2018. Baltimore, MD, USA. P. 1615–1631.

34. *Lia Y., Wangb H., Barni M.* A survey of deep neural network watermarking techniques. https://arxiv.org/pdf/2103.09274.pdf

35. *Li Y., Zhang Z., Bai J. et al.* Open-sourced Dataset Protection via Backdoor Watermarking. https://arxiv.org/abs/2010.05821

TRUSTED ARTIFICIAL INTELLIGENCE

A.I. Avetisyan[a],*

[a]*Ivannikov Institute for System Programming of the Russian Academy of Sciences, Moscow, Russia*
**E-mail: arut@ispras.ru*

In this paper we discuss the problem of creating trusted artificial intelligence (AI) technologies. Modern AI is based on machine learning and neural networks and is vulnerable to biases and errors. Efforts are made to establish standards for the development of trusted AI technologies, but they have not yet succeeded. AI technologies trust can only be achieved with the appropriate scientific and technological base and corresponding tools and techniques for countering attacks. We present the ISP RAS Trusted AI Research Center results and propose a work model that can ensure technological independence and long-term sustainable development in this area.
The paper expands the lecture given at the scientific session of the General Assembly of the RAS on December 12, 2023.

Keywords: trusted artificial intelligence, cybersecurity, vulnerabilities, regulation, trusted software repository, weak artificial intelligence, neural networks, federative learning, machine learning.