ВЕСТНИК РОССИЙСКОЙ АКАДЕМИИ НАУК, 2022, том 92, № 12, с. 1119–1123

С КАФЕДРЫ ПРЕЗИДИУМА РАН

КИБЕРБЕЗОПАСНОСТЬ В КОНТЕКСТЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

© 2022 г. А. И. Аветисян

Институт системного программирования им. В.П. Иванникова РАН, Москва, Россия

E-mail: arut@ispras.ru

Поступила в редакцию 20.04.2022 г.
После доработки 24.04.2022 г.
Принята к публикации 06.09.2022 г.

Широкое внедрение технологий искусственного интеллекта сопровождается новыми вызовами в сфере кибербезопасности. В статье рассматриваются возникающие угрозы такого рода, подчёркивается необходимость создания сквозных технологий, обеспечивающих весь жизненный цикл разработки и эксплуатации систем искусственного интеллекта с заданным уровнем доверия. Уделяется внимание и организационным моделям развития соответствующих технологий на примере Центра доверенного искусственного интеллекта, созданного в Институте системного программирования им. В.П. Иванникова РАН в 2021 г.

Статья подготовлена на основе доклада, заслушанного на одном из заседаний президиума РАН.

Ключевые слова: доверенный искусственный интеллект, кибербезопасность, машинное обучение, нейросетевые модели, датацентричность, методы атаки, программные инструменты, доверенные фреймворки.

DOI: 10.31857/S0869587322120039

В связи со стремительным развитием новых компьютерных технологий значительно возросла значимость проблем кибербезопасности, возникла необходимость рассматривать её в качестве отдельной области научного знания. В 2018 г. президиум РАН принял решение о формировании научного направления “Анализ, трансформация программ и кибербезопасность”. 24 февраля 2021 г. новая научная специальность (наименование – “кибербезопасность”, шифр 1.2.4), по которой присуждаются учёные степени в области физико-математических наук, была утверждена приказом № 118 Минобрнауки России. Специальность охватывает такие направления исследований, как анализ и систематизация уязвимостей, моделирование политик информационной безопасности, угроз и атак, масштабируемые средства интеллектуального анализа данных и процессов в распределённых системах и др.

АВЕТИСЯН Арутюн Ишханович – академик РАН, директор ИСП РАН.

Инструменты поддержания необходимого уровня кибербезопасности специфичны и зависят от разрабатываемых систем (традиционных или встроенных), требуя адаптации под конкретные цели и задачи. Эта проблематика настолько обширна, что подробное её обсуждение не представляется возможным в данной статье. Важно отметить следующее: появление и широкое внедрение технологий искусственного интеллекта (ИИ) одновременно сопровождалось появлением принципиально новых вызовов в области кибербезопасности, ответы на которые требуют проведения соответствующих фундаментальных исследований и инструментов [1].

Прежде чем обратиться к этим вызовам, необходимо кратко представить историю возникновения технологий искусственного интеллекта. Сам термин появился в 1956 г., однако особенно активно технологии ИИ начали развиваться и внедряться с 1990-х годов. В 1997 г. шахматный супер-

компьютер Deep Blue, разработанный компанией IBM, выиграл матч из шести партий у чемпиона мира по шахматам Г. Каспарова. В 2002 г. был выпущен первый робот-пылесос. В 2010 г. началось создание базы данных аннотированных изображений ImageNet [2], предназначенной для обработки и тестирования методов распознавания образов и машинного зрения. К настоящему времени в базе насчитывается более 14 млн изображений 21 тыс. категорий. В 2011 г. компьютер Watson компании IBM одержал победу в телевизионной игре-викторине “Jeopardy!” (на российском телевидении её версия получила название “Своя игра”). В том же году был разработан персональный виртуальный цифровой помощник Siri для использования в смартфоне. В 2016 г. сервис Google Translate начал использовать нейронный машинный перевод с восьми языков. В 2021 г. в Китае была представлена языковая модель WuDao2.0 [3], в которой используется 1.75 трлн параметров.

В современном мире искусственный интеллект широко применяется в интернет-помощниках (помимо Google Translate, это Google Photos, Google Assistant), на транспорте (беспилотные автомобили и летательные аппараты), в области финансов (PayPal, поиск подозрительной активности в транзакциях), торговле (товарные рекомендации в ритейле и роботизация складского бизнеса), медицине (компьютерная диагностика, подбор методов лечения, фитнес-браслеты, глюкометры и другие гаджеты), системах безопасности (распознавание лиц с помощью компьютерного зрения), космических исследованиях (робот Curiosity, разработанный NASA, перемещался по поверхности Марса в отсутствие связи с Землёй), в промышленности (роботизация производства, сокращение штата сотрудников). Объём глобального рынка технологий искусственного интеллекта постоянно растёт и, по некоторым данным, в 2024 г. превысит 500 млрд долл.

Искусственный интеллект внедряется повсеместно, однако с технологической точки зрения остаётся слабым и плохо защищённым. Слабый ИИ – общепринятый термин, отражающий состояние современных технологий искусственного интеллекта, основанных на методах машинного обучения, глубокого обучения и нейронных сетей. Слабый ИИ извлекает информацию из ограниченного набора данных и может решать только те задачи, на которые он запрограммирован. Вместе с тем он обрабатывает информацию быстрее человека, таким образом избавляя нас от рутинных задач. В противоположность слабому ИИ в перспективе должен появиться сильный, способный делать интеллектуальные выводы, использовать стратегии, функционировать в условиях неопределённости, общаться на естественном языке и планировать действия, то есть решать задачи на уровне интеллекта человека [4]. Однако неизвестно, когда такой искусственный интеллект удастся создать, поскольку в настоящее время отсутствуют даже методы его разработки.

Отдельную проблему в развитии технологий ИИ представляет собой экспоненциальный рост вычислительных ресурсов [5].

В основе современного искусственного интеллекта лежат модели машинного обучения. Именно они становятся слабым местом с точки зрения кибербезопасности. Появляются новые классы атак – на обученные нейросетевые модели, причём уязвимости могут возникать либо внедряться на всех этапах жизненного цикла модели [6]. Это атаки уклонения, кражи конфиденциальных данных и самих моделей (рис. 1).

Рассмотрим несколько примеров подробнее. Одним из распространённых методов атаки служит отравление данных и моделей [7]. В небольшое количество обучающих примеров добавляется триггер – специально подготовленный фрагмент изображения. В результате обучения на таком наборе данных модель становится отравленной. Триггер приводит её к заведомо ошибочному предсказанию на этапе эксплуатации (в том числе к предсказанию заведомо известного нарушителю результата). Предобученные отравленные модели могут распространяться через Интернет и нести в себе угрозу при переносе знаний.

Ещё один распространённый способ разрушающего воздействия на искусственный интеллект – атаки уклонения [8], в том числе неразличимые. Результат такой атаки представлен на рисунке 2. Для достижения разрушающего результата в некоторых случаях достаточно изменить один пиксель, в итоге ИИ утку идентифицирует с лошадью.

Существуют также атаки “белого ящика”, когда нарушителю становится доступна полная информация о модели машинного обучения, и атаки “чёрного ящика”, когда нарушитель получает доступ к предсказаниям модели на основе произвольных входных данных (метки либо вероятности классов).

Возможны кражи данных и моделей из облачных сред [9]. Значимую роль играют и уязвимости в исходном коде фреймворков[1] машинного обучения [10]. К примеру, фреймворк машинного обучения TensorFlow достаточно обширен и содержит около трёх миллионов строк кода, а также несколько десятков библиотек-зависимостей (NumPy и др.). Классические уязвимости (CVE) в исходном коде фреймворков и библиотек расши-

[1] Фре́ймворк (от англ. framework – остов, каркас) – программная платформа, определяющая структуру программной системы; программное обеспечение, облегчающее разработку и объединение разных компонентов большого программного проекта.

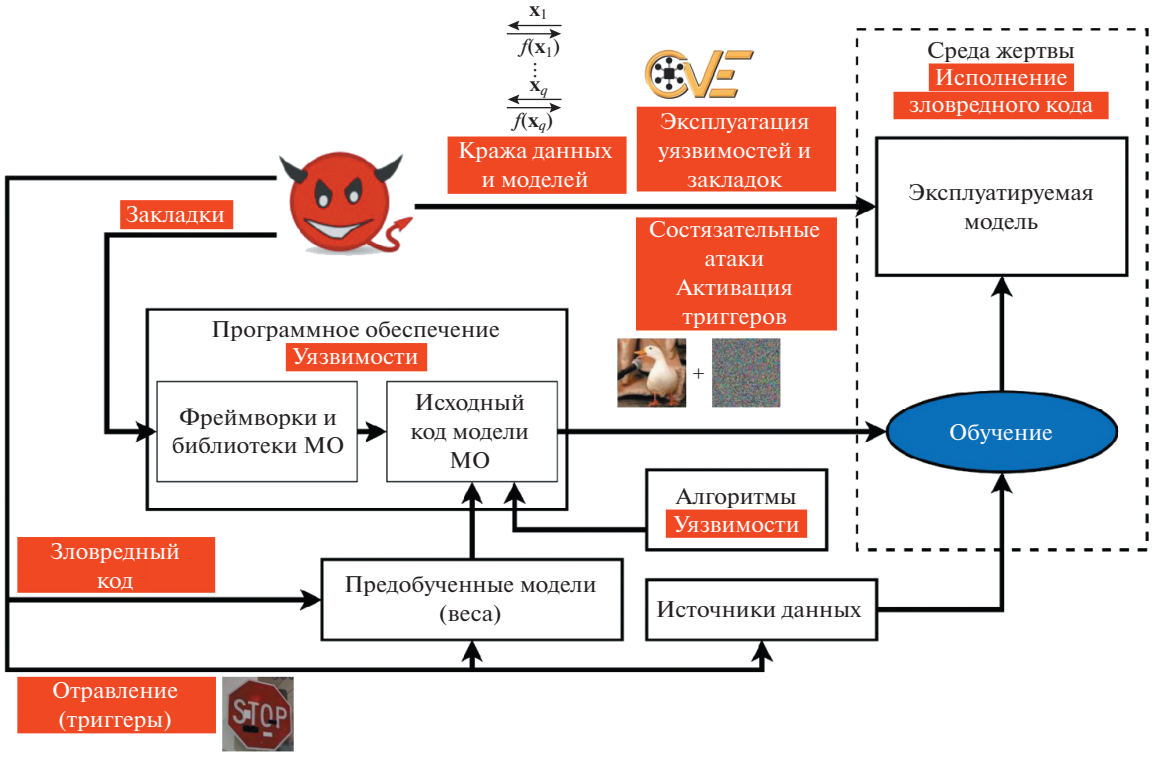

Рис. 1. Атаки на модели машинного обучения (общая схема)

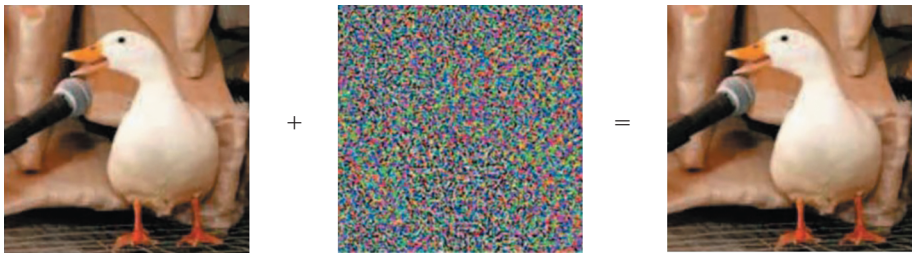

Рис. 2. Результат атаки уклонения

ряют поверхность атаки на эксплуатируемые модели (примеры: переполнение буфера в библиотеке OpenCV, атака с помощью специально подготовленного BMP-изображения). Чтобы избежать проблем, возникающих из-за таких уязвимостей, необходимо использовать статический анализ исходного кода фреймворков и создавать их доверенные версии.

В числе изощрённых методов кибератак – встраивание зловредного кода в модели машинного обучения. Встраивание такого кода (его размер может достигать нескольких мегабайт) в вещественные параметры нейросетевых моделей происходит без существенной потери их точности и не обнаруживается антивирусным программным обеспечением [11]. Возможна также компрометация устройства жертвы при использовании предобученных моделей со зловредным кодом, распространяемых через Интернет (GitHub и другие ресурсы).

В связи с этими угрозами необходима разработка методов их предотвращения, а также создание комплекса инструментов, обеспечивающего безопасность на протяжении всего жизненного цикла систем ИИ с заданным уровнем доверия – от проектирования, сбора и подготовки данных до эксплуатации. Проблемы создания таких доверенных систем сейчас активно обсуждаются научным сообществом разных стран. Например, Национальный институт стандартов и технологий США (NIST) занимается созданием NIST AI Risk Management Framework. В Германии создаёт-

ся DIN DKE German Standardization Map on Artificial Intelligence. Известны также MITRE ATLAS, Adversarial Threat Landscape for Artificial-Intelligence Systems (США) и Google Responsible AI practices (США). Растёт число научных публикаций, посвящённых атакам на системы ИИ, однако вопрос применимости предлагаемых методов защиты к реальным системам искусственного интеллекта пока остаётся без ответа и требует исследований [12]. Например, наиболее изученная модель нарушителя в атаках с использованием состязательных примеров редко бывает эффективной на практике. Необходима разработка реалистичных моделей нарушителя, а также методов противодействия, в том числе основанных на интерпретируемости моделей машинного обучения. На практике задача классификации изображений служит лишь вспомогательной; методы защиты от угроз в реальных задачах (детекция объектов и др.) отличаются и исследованы в меньшей степени.

Из изложенного можно сделать следующий вывод: разработка методов и технологий создания систем доверенного ИИ возможна только с привлечением сообщества учёных, специализирующихся в этой научной области (такое сообщество ещё предстоит сформировать), а также индустриальных партнёров, нуждающихся в решении своих прикладных задач. В целях эффективной разработки доверенных систем ИИ желательно создать облачную платформу, которая должна объединить:

- методы и методики разработки и оценки доверенных систем;
- программные инструменты анализа и выявления угроз, специфичных для ИИ, а также с целью противодействия им;
- доверенные фреймворки машинного обучения.

Требования к доверенным системам с ИИ должны охватывать весь их жизненный цикл, включая:

- анализ и проектирование (например, формирование требований к устойчивости к атакам уклонения, отравления и извлечения информации);
- разработку (методы противодействия атакам на модели машинного обучения);
- тестирование (технологическая тестовая база оценки безопасности систем с ИИ);
- эксплуатацию (экспертиза обучающих выборок и моделей).

Долгосрочное развитие данной области науки, создание соответствующих методик и инструментов возможно только при объединении усилий академического сообщества, индустрии и государственных ведомств. Именно в таком ключе развивается Центр доверенного искусственного интеллекта в Институте системного программирования им. В.П. Иванникова РАН. Он был создан в 2021 г. по итогам победы в конкурсе, проведённом в рамках федерального проекта “Искусственный интеллект”, курируемого Минэкономразвития России. В числе партнёров Центра – НИУ “Московский физико-технический институт”, Сколковский институт науки и технологий, Механико-математический факультет и Медицинский научно-образовательный центр МГУ им. М.В. Ломоносова, Университет Иннополис, Национальный исследовательский Нижегородский государственный университет им. Н.И. Лобачевского, Институт психологии РАН, Межведомственный суперкомпьютерный центр РАН, АО “Лаборатория Касперского”, ЗАО “ЕС-Лизинг”, компания “Интерпроком”, ООО НПК “ТехноПром”. Индустриальные партнёры осуществляют подготовку и передачу наборов данных, тестирование доверенных фреймворков машинного обучения, опытную эксплуатацию. Программа Центра предусматривает создание методик и соответствующих программных и аппаратно-программных платформ для разработки и верификации технологий ИИ с требуемым уровнем доверия. Ключевые направления программы:

- классификация угроз и разработка программных инструментов для анализа, выявления и противодействия угрозам, специфичным для систем с ИИ (атаки уклонения, атаки с внедрением закладок и зловредного кода, кражи моделей и данных);
- повышение интерпретируемости моделей;
- создание методик и бенчмарков[2] на основе реальных приложений (медицина, социология, информационная безопасность);
- формирование доверенных сред разработки моделей машинного обучения;
- создание отчуждаемой облачной платформы для разработки доверенных систем, использующих ИИ.

Программа базируется на промышленных технологиях Института системного программирования им. Н.И. Иванникова РАН (Talisman, Asperitas и др.), долгосрочных партнёрских отношениях с индустрией и академическим сообществом. Уже выполнен ряд научно-исследовательских работ, в том числе с Академией криптографии РФ, реализуется долгосрочный контракт с компанией “Samsung Electronics” по тематике доверенного ИИ (интерпретируемость моделей машинного обучения). При этом следует учитывать, что реа-

[2] Бенчмарк в вычислительной технике – процесс запуска компьютерной программы, набора программ или других операций с целью оценки относительной производительности объекта, обычно путём выполнения ряда стандартных тестов и испытаний.

лизации одной такой программы для решения всего круга сложных задач явно недостаточно. Необходимо создать специализированное научное сообщество при руководящей роли РАН, инициировать другие программы, активизировать международное сотрудничество (в первую очередь, в рамках ЕврАзЭС). Имеющийся технологический задел и научные школы позволяют сформировать центр компетенций в области доверенного искусственного интеллекта, что будет способствовать обеспечению технологической независимости и безопасности нашей страны.

ЛИТЕРАТУРА

1. *Li J.* Cyber security meets artificial intelligence: a survey // Frontiers Inf. Technol. Electronic Eng. 2018. V. 19. P. 1462–1474. https://doi.org/10.1631/FITEE.1800573
2. *Deng J., Dong W., Socher R. et al.* ImageNet: A large-scale hierarchical image database // IEEE Conference on Computer Vision and Pattern Recognition. 2009. P. 248–255. https://doi.org/10.1109/CVPR.2009.5206848
3. Wu Dao 2.0: China's Improved Version of GPT-3 (электронный ресурс). https://research.aimultiple.com/wu-dao/
4. *Goertzel B., Pennachin C.* (Eds). Artificial general intelligence. Part 2. New York: Springer, 2007.
5. AI and Compute (электронный ресурс). https://openai.com/blog/ai-and-compute/
6. *Chakraborty A., Alam M., Dey V. et al.* Adversarial Attacks and Defences: A Survey. 2018. ArXiv, abs/1810.00069
7. *Gu T., Dolan-Gavitt B., Garg S.* BadNets: Identifying vulnerabilities in the machine learning model supply chain. 2017. arXiv preprint arXiv:1708.06733
8. *Biggio B., Corona I., Maiorca D. et al.* Evasion attacks against machine learning at test time // Joint European conference on machine learning and knowledge discovery in databases. Berlin, Heidelberg: Springer, 2013. P. 387–402.
9. *Tramèr F., Zhang F., Juels A. et al.* Stealing machine learning models via prediction APIs // Proceedings of the 25th USENIX Conference on Security Symposium (SEC'16). 2016. USENIX Association, USA. P. 601–618.
10. *Xiao Q., Li K., Zhang D., Xu W.* Security Risks in Deep Learning Implementations // 2018 IEEE Symposium on Security and Privacy Workshops. P. 123–128. https://doi.org/10.1109/SPW.2018.00027
11. *Wang Z., Liu Ch., Cui X.* EvilModel 2.0: Hiding Malware Inside of Neural Network Models. 2021 IEEE Symposium on Computers and Communications (ISCC). https://doi.org/10.1109/ISCC53001.2021.9631425
12. *Tramèr F.* Does Adversarial Machine Learning Research Matter? 2021. https://floriantramer.com/docs/slides/advml21award.pdf