Перспективные подходы повышения защищенности систем искусственного интеллекта

А. Б. Менисов, А. Г. Ломако
Военно-космическая академия имени А.Ф. Можайского
E-mail: vka@mil.ru

***Аннотация*. Роль кибербезопасности в создании надежных и заслуживающих доверия систем искусственного интеллекта особенно важна, так как злоумышленники проводят все более разнообразные и масштабные компьютерные атаки на системы искусственного интеллекта (СИИ) и постоянно ищут новые уязвимости. В статье представлен анализ угроз СИИ и способов защиты моделей машинного обучения. Для выполнения требования адаптации под угрозы безопасности авторы предлагают подход адаптивной кибериммунной защиты СИИ.**

Ключевые слова: искусственный интеллект; информационная безопасность; кибериммунные системы; адаптивная защита; кибериммунная защита

I. Введение

В эпоху цифровых технологий все коммерческие и государственные организации сталкиваются с широким спектром автоматизированных и быстро распространяющихся угроз безопасности информации – от кражи конфиденциальных данных и манипулирования ими до огромных убытков, вызванных прерыванием (полной остановкой) различных процессов. Это вызвано не только растущей сложностью, разнообразием и масштабом цифровизации, но и развитием угроз. В недавнем прошлом, когда субъекты угроз (хакеры, преступники, злоумышленники) были менее компетентны, а цифровая активность была более предсказуемой, традиционный подход к безопасности часто был достаточным для предотвращения угроз безопасности информации. Тем не менее, значимость рисков новых внешних и внутренних угроз, а также усложнение структуры объектов информационной инфраструктуры катастрофически снижают их защищенность. Традиционные средства защиты не позволяют обнаружить новые тактики, техники и способы злоумышленников, которые теперь могут маскироваться под сетевой шум и проникать в сложные информационные инфраструктуры.

Чаще всего злоумышленники при осуществлении атак используют вредоносное программное обеспечение, методы социальной инженерии и эксплуатации веб-уязвимостей, а также способ отслеживания активности пользователей. Объектами атак, как правило, являются компьютеры, серверы и сетевое оборудование, а также программное обеспечение информационных систем, информационно-телекоммуникационных сетей и автоматизированных систем управления.

Новые информационные технологии, такие как системы искусственного интеллекта (СИИ), внедряются во все сферы общества и государства. Российский регулятор (ФСТЭК) определил 5 угроз безопасности информации, связанных с технологиями искусственного интеллекта [1]: раскрытия информации о модели машинного обучения; хищения обучающих данных; нарушения функционирования («обхода») средств, реализующих технологии искусственного интеллекта; модификации модели машинного обучения путем искажения («отравления») обучающих данных; и подмены модели машинного обучения, Стоит отметить, что проблема создания доверенного (безопасного) искусственного интеллекта давно встала перед обществом, и использовать эти технологии как «черный ящик» небезопасно.

II. Угрозы системам искусственного интеллекта

СИИ могут быть атакованы путем эксплуатации уязвимостей и дефектов программного обеспечения, а также новых уязвимостей различных компонентов машинного обучения, процессом машинного обучения, когда злоумышленники получают доступ и манипулируют незащищенным программным кодом и данными. Например, фреймворки машинного обучения содержат дефекты и уязвимости, которые могут привести к отказу в обслуживании, раскрытию информации или удаленному выполнению кода. В табл. 1 представлены результаты проведения статического анализа с помощью средства Coverity [2].

ТАБЛИЦА I. Дефекты и уязвимости фреймворков машинного обучения

Фреймворки (версия)	Дефекты					
	Уровень опасности					
	Высокий	Средний	Низкий	Активность обнаружения (за последние 5 лет)	Top-25	Всего
CNTK 2.7	12	41	4	0	1	57
Dlib 19.24	52	112	21	0	5	185
Keras 2.10.0	0	16	38	3	3	54
MXNet 1.9.1	465	918	197	2	5	1580
Sklearn 1.1.2	93	323	451	2	2	867
PyTorch 1.12.1	1468	1150	241	0	5	2859
Tensorflow 2.10	43	140	8	323	3	191

Факторы снижения качества моделей машинного обучения (МО), которые могут быть связаны с естественными (непреднамеренное снижение качества)

Работа выполнена при финансовой поддержке гранта Президента Российской Федерации для государственной поддержки молодых российских ученых - кандидатов наук, МК-2485.2022.4.

[3] или искусственными (преднамеренное снижение качества) причинами [4], могут возникать на всех этапах жизненного цикла таких систем, в том числе:

- на этапе сбора данных для обучения моделей машинного обучения существуют возможность как «отравления» неразмеченных данных, поступающих экспертам в прикладных областях для разметки, но также «отравление» может осуществляться незаметной для эксперта модификацией данных. В результате обучения на «отравленных» данных модель может быть легко атакована злоумышленником, производившим «отравление», при помощи аналогичных модификаций;
- на этапе обучения и формирования гиперпараметров модели машинного обучения возможно встраивание зловредного кода, при этом модель МО показывает хорошую точность на тестовой выборке. Зловредный код при этом может быть извлечен с помощью «вспомогательных» программ, созданных злоумышленником как дополнение к поставляемой модели;
- на этапе эксплуатации моделей, в том числе их обновления. В последние годы разрабатываются атаки черного ящика на нейросетевые модели машинного обучения с помощью состязательных примеров, сопоставимые по эффективности с атаками белого ящика.

Перечисленные примеры угроз подразумевают наличие злоумышленника, воздействующего на СИИ на одном из этапов ее жизненного цикла, однако в общем случае угрозы безопасности могут возникать и без такого воздействия из-за следующих факторов (для типичных входных данных без атак на эти данные):

- дисбаланс, ложные закономерности в данных, изменение распределений поступающих данных;
- неадекватный выбор модели и гиперпараметров ее обучения.

Наиболее актуальными примерами преднамеренного снижения качества, специфичными для СИИ являются:

- на стадии создания системы – наличие преднамеренных искажений е обучающей выборке системы распознавания изображений, приводящих к ошибкам в работе системы распознавания, вызванным специальными, заранее определенными искажениями в исходных данных, включая состязательные атаки;
- на стадии эксплуатации системы – отсутствие достоверных и представительных оценок устойчивости системы распознавания изображений к воздействию преднамеренных состязательных атак, приводящее к неустойчивой работе системы в процессе ее эксплуатации.

III. Анализ существующих подходов к обеспечению защищенности СИИ

В области обеспечения безопасности СИИ работает относительно достаточное количество организаций, большинство составляют китайские, а также американские компании и университеты: MIT, Стэнфорд, Пекинский, Тяньцзиньский, Национальный университет Чунг Син. Стоит выделить организации в области безопасности СИИ: Enisa, MITRE и Microsoft, которые создали матрицу угроз для каталогизации угроз СИИ [5]. Контент матрицы включает в себя документацию о случаях компьютерных атак на коммерческие системы искусственного интеллекта. Для специалистов по безопасности Microsoft предлагает программный продукт Counterfit с открытым исходным кодом для оценки состояния защищенности СИИ.

В табл. 2 представлен анализ различных подходов к обеспечению защищенности систем искусственного интеллекта [6]. Различные подходы анализировались по сложности и оправдываемости.

ТАБЛИЦА II. Анализ подходов к обеспечению защищенности СИИ

Тип	**Подтип**	**Сложность вычислений**	**Оправды ваемость**
Предобработка данных	Маскировка градиента	Ср	Ср
	Сжатие признаков	Б	Ср
	Градиентная регуляризация	Ср	Ср
Фильтрация данных	Дополнительная модель машинного обучения	Б	В
	Ансамблевая защита	Б	В
Диагностирование модели машинного обучения		М	В
Оптимизация модели машинного обучения	Реконструкция модели машинного обучения	М	Ср
	Состязательное обучение	Ср	В
	Обрезка нейронов	М	В
Комплексная защита		М	Н

Сложность вычислений определяется вычислительными затратами каждого подхода. Быстрое реагирование на всевозможные виды воздействия может поддерживать требуемый уровень надежности СИИ. Три категории «Б», «Ср» и «М» соответствуют быстрым (менее 1 секунды), средним (от нескольких секунд до нескольких минут) и медленным (от нескольких минут до нескольких часов или нескольких дней) вычислительным затратам.

Критерий оправдываемости подходов измеряет уровень развития каждого подхода – на количестве соответствующих публикаций и патентных заявок. Три категории «Н», «Ср» и «В» соответственно соответствуют низкому (менее 5 публикаций, нет приложений), среднему (5–20 публикаций, прототипы приложений) и высокому уровню (более 20 публикаций и успешно применяются в реальных СИИ).

IV. Перспективные подходы к обеспечению защищенности СИИ

Несмотря на недостатки в обеспечении безопасности, СИИ становятся основой для выполнения производственных процессов. Ожидается, что в связи с необходимостью создания сильного ИИ (который способен решать множество задач) будут использоваться биоинспирированные подходы [7]. Кибериммунные системы (КИС) – одна из самых популярных категорий биотехнологических методов [8, 9]. Эти подходы

основаны на механизмах биологической иммунной системы (БИС).

БИС – это надежная и самоорганизующаяся система, которая обрабатывает данные о состоянии тела и предпринимает соответствующие действия, чтобы поддерживать здоровое состояние [адаптировано из 10]. БИС представляет собой многоуровневую систему, в которой на каждом уровне активны различные типы защитных механизмов. Существуют три основные линии защиты: анатомический барьер, врожденный иммунитет и адаптивный иммунитет. Анатомический барьер состоит из физических препятствий, таких как кожа и слизистые оболочки, которые предотвращают попадание в организм патогенов, таких как бактерии и вирусы. Если патоген прорывает первую линию защиты, врожденный иммунитет обеспечивает немедленный, но неспецифический ответ, такой как воспалительный ответ и воздействие антимикробных белков. Если патогены успешно уклоняются от врожденного ответа, он активирует адаптивный иммунитет, который адаптирует свой ответ во время инфекции, чтобы улучшить свои навыки распознавания и убить патоген.

КИС уже доказали свою пригодность для решения реальных задач в различных областях, таких как распознавание образов и классификация, обнаружение аномалий, оптимизации и т. д. СИИ достигли такого уровня сложности, когда вмешательство человека в обеспечение безопасности [11] становится все более трудным. Все более актуально требование, что такие системы должны иметь возможность адаптироваться сами, исключая необходимость вмешательства человека. Автономная кибериммунная защита СИИ способна поддерживать, улучшать и восстанавливать функциональность СИИ без внешних воздействий и обладает свойствами самоорганизации в следующих аспектах:

- самонастройки;
- само-оптимизации;
- самовосстановления;
- самоадаптации.

Исходя из вышеперечисленных свойств автономная кибериммунная защита СИИ, должна обладать следующим функционалом:

- автоматически конфигурироваться и реконфигурироваться в зависимости от изменяющейся внешней среды;
- автоматически оптимизировать свою производительность для обеспечения наиболее эффективного вычислительного процесса;
- автоматически обнаруживать, идентифицировать и защищаться от различных типов атак для поддержания общей безопасности и целостности СИИ;
- автоматически адаптироваться к своей среде по мере ее изменения, взаимодействуя с соседними системами и устанавливая протоколы связи;
- поддерживать устойчивость функционирования в недоверенной среде.

V. Заключение

Растущие производительность и сложность систем искусственного интеллекта создали условия для появления нового типа киберугроз – компьютерных атак, направленных непосредственно на отдельные компоненты и СИИ в целом. Поскольку внимание сосредоточено на разработке и интеграции СИИ в приложения и рабочие процессы, их безопасность часто упускается из виду.

В частности, важно понимать соответствующие подходы к обеспечению защищенности для комплексного управления угрозами в многоуровневой экосистеме и разработки специальных средств контроля, обеспечивающих безопасность СИИ.

Любые подходы, обеспечивающие автономность защиты, являются перспективными, однако, наиболее перспективным и комплексным является кибериммунный подход.

Направлениями дальнейших перспектив исследования являются:

- разработка концепции проактивной защиты систем искусственного интеллекта на основе кибериммунного ответа;
- разработка архитектуры системы кибербезопасности компонентов искусственного интеллекта с кибериммунным ответом.

Список литературы

[1] Список угроз ФСТЭК // URL: https://bdu.fstec.ru/threat (дата обращения: 20.03.2023).

[2] Wu Y., Su J., Moran D.D., Near C.D. Automated Software Testing Starting from Static Analysis: Current State of the Art //arXiv preprint arXiv:2301.06215. – 2023.

[3] ГОСТ Р 59276-2020. Системы искусственного интеллекта. Способы обеспечения доверия. Общие положения.

[4] Ломако А.Г., Менисов А.Б. Ландшафт угроз безопасности информации технологий искусственного интеллекта // 31-й научно-технической конференции методы и технические средства обеспечения безопасности информации. 2022. С. 49.

[5] ATLAS Navigator. URL: https://mitre-atlas.github.io/atlas-navigator/ (дата обращения - 20.03.2023).

[6] Менисов А.Б. Технологии искусственного интеллекта и кибербезопасность: монография / Менисов А.Б.. Москва : Ай Пи Ар Медиа, 2022. 133 с.

[7] Pinto R., Gonçalves G. Application of artificial immune systems in advanced manufacturing // Array. 2022. С. 100238.

[8] Kaspersky Cyber Immunity // URL: https://os.kaspersky.com/technologies/cyber-immunity/ (дата обращения: 20.03.2023).

[9] Petrenko S. Developing a Cybersecurity Immune System for Industry 4.0. CRC Press, 2022.

[10] Гусейнов А.А., Зиновьев А.А. Большая российская энциклопедия. М.: БРЭ. 2008. Т. 10. С. 493-495.

[11] ГОСТ Р 59898-2021 «Оценка качества систем искусственного интеллекта. Общие положения».