DOI: 10.15514/ISPRAS-2025-37(3)-4

Моделирование сценариев деструктивного воздействия на целостность моделей машинного обучения

А.Б. Менисов, ORCID: 0000-0002-9955-2694 <men.arty@yandex.ru>
А.Г. Ломако, ORCID: 0000-0002-1764-1942 <vka@mil.ru>
Военно-космическая академия имени А.Ф.Можайского,
Россия, 197198, г. Санкт-Петербург, ул. Ждановская, д. 13.

Аннотация. Статья посвящена разработке моделей деструктивного воздействия на целостность моделей машинного обучения на основе SIR-прогнозирования масштаба угроз и рисков при различных сценариях развития компьютерных атак. В статье представлена оригинальная модель угроз информационной безопасности техническим компонентам искусственного интеллекта в условиях разнородно массовых компьютерных атак, отображающая уязвимые места и способы возможных действий злоумышленников. Авторами разработана методология адаптации модернизированных SIR-моделей природных эпидемий для выявления подобия и аналогов в характере распространения деструктивных сбоев в системах ИИ, вызванных разнородно-массовыми и таргетированными воздействиями. Выявленные закономерности позволили оценить риски возможного ущерба целостности и разработать эффективные стратегии предотвращения и исправления искажений моделей машинного обучения.

Ключевые слова: искусственный интеллект; целостность моделей машинного обучения; доверие; информационная безопасность; диагностическое тестирование; тестовая среда.

Для цитирования: Менисов А.Б., Ломако А.Г. Моделирование сценариев деструктивного воздействия на целостность моделей машинного обучения. Труды ИСП РАН, том 37, вып.3, 2025 г., стр. 59–68. DOI: 10.15514/ISPRAS–2025–37(3)–4.

Modeling Scenarios of Destructive Impact on the Integrity of Machine Learning Models

A.B. Menisov, ORCID: 0000-0002-9955-2694 <men.arty@yandex.ru>
A.G. Lomako, ORCID: 0000-0002-1764-1942 <vka@mil.ru>

Mozhaisky Military *Space Academy,*
13, Zhdanovskaya st., St. Petersburg, 197198, Russia.

Abstract. The article is devoted to the development of models of destructive impact on the integrity of machine learning models based on SIR forecasting of the scale of threats and risks of losses under various scenarios of computer attacks. The article presents an original model of information security threats to technical components of artificial intelligence in the context of heterogeneous mass computer attacks, displaying vulnerabilities and methods of possible enemy actions. The authors have developed a methodology for adapting modernized SIR models of natural epidemics to identify similarities and analogues in the nature of the spread of destructive failures in AI systems caused by heterogeneous mass and targeted impacts. The identified patterns made it possible to assess the risks of possible damage to integrity and develop effective strategies for preventing and correcting distortions of machine learning models.

Keywords: artificial intelligence; integrity of machine learning models; trust; information security; diagnostic testing; test environment.

For citation: Menisov A.B., Lomako A.G. Modeling scenarios of destructive impact on the integrity of machine learning models. Trudy ISP RAN/Proc. ISP RAS, vol. 37, issue 3, 2025, pp. 59-68 (in Russian). DOI: 10.15514/ISPRAS-2025-37(3)-4.

1. Введение

Несмотря на существенный рост производительности и качества решений прикладных задач с использованием технологий искусственного интеллекта (ИИ) [1], следует понимать, что это создает опасные риски потери работоспособности, обусловленные использованием открытых наборов данных, сторонних предобученных моделей и чужой инфраструктуры машинного обучения [2]. Необходимость построения системы защиты компонентам ИИ вызвана исключительной актуальностью этой проблематики в условиях современных вызовов и угроз [3]. Сбои и отказы систем ИИ могут быть вызваны новыми типами атак, когда злоумышленник получает доступ и манипулирует незащищёнными данными и моделями машинного обучения [4].

Кроме того, использование технологий ИИ предоставляет новые возможности злоумышленникам для реализации угроз информационной безопасности, таких как обучение нежелательному поведению (модификация модели машинного обучения путем искажения («отравления») данных) [5], хищение обучающих данных [6], введение модели в заблуждение (состязательные атаки) [7], подмена модели машинного обучения [8] и другие [9]. Опасность реализации таких угроз определяется тем, что алгоритмы обработки данных в машинном обучении представляют собой наборы обученных коэффициентов (большие массивы чисел). Поэтому несанкционированные изменения в структуре модели машинного обучения не могут быть выявлены стандартными средствами контроля и анализа программного обеспечения, которые сейчас применяются. При этом сами нежелательные изменения в структуре и поведении модели машинного обучения могут осуществляться различными способами, включая, например, манипуляции с обучающей выборкой, либо процессом обучения, что также невозможно проконтролировать стандартными средствами.

Таким образом, формируется потребность в разработке специальных программных средств защиты информации для борьбы с компьютерными атаками и снижением уязвимости алгоритмов, основанных на технологиях ИИ. Данные средства должны обеспечить необходимый базис для формирования новых подходов к тестированию алгоритмов ИИ, а

также позволить определить потенциальные угрозы методов машинного обучения и сформировать рекомендации по коррекции специального программного обеспечения. Систематическое применение подобного подхода позволит сформировать ландшафт потенциальных специфических угроз информационной безопасности, а также выработать лучшие практики по возможным контрмерам, позволяющим минимизировать риски при использовании алгоритмов, основанных на машинном обучении в практических приложениях.

На рис. 1 показаны потенциально-возможные сценарии злоумышленников на компоненты ИИ. Анализ фактографии реализации угроз на системы ИИ позволил определить основные негативные действия злоумышленников по намеренному вызову сбоев и отказов. Они представлены в сценариях компьютерных атак на системы, содержащие компоненты машинного обучения и демонстрируют расширение арсенала технологий деструктивного воздействия.

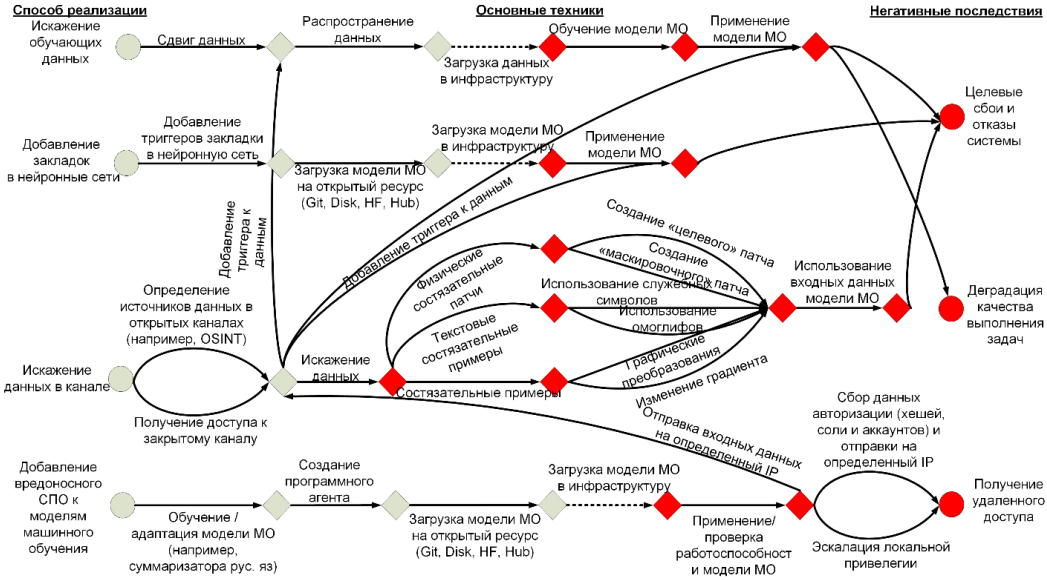

Рис. 1. Типизация сценариев деструктивных действий злоумышленников на системы, содержащие компоненты машинного обучения.
Fig. 1. Typification of scenarios of destructive actions of attackers on systems containing machine learning components.

Исследование путей обеспечения защищенности комплексов ИИ в условиях негативных сценариев, приводящих к нарушениям целостности данных и моделей машинного обучения, показало следующее:

- несовершенство методов обнаружения и нейтрализации для администрирования безопасности систем ИИ в условиях деструктивного влияния на качество решения прикладных задач, так как сравнительный анализ современных систем и мер обеспечения информационной безопасности систем ИИ показал не способность выявить признаки атак на компоненты ИИ;
- явное преимущество проактивной схемы обеспечения защиты в связи с преимуществами по сравнению с реактивной и постфактной;
- необходимости обеспечения своевременного выявления нетиповых аномалий функционирования систем ИИ с использованием идеологии обучаемого динамического детектирования.

Таким образом, существует необходимость разработки теоретических основ и специальных программных средств проактивной защиты информации для борьбы с компьютерными атаками и своевременной нейтрализацией уязвимостей алгоритмов, используемых в технологиях ИИ. Данные средства должны в итоге обеспечить необходимый и достаточный (в развитии) арсенал обеспечения защищенности систем ИИ в условиях угроз компьютерных атак на модели машинного обучения.

2. Формализованная постановка проблемы

Формализованная постановка проблемы исследования представлена следующим образом: пусть система защиты имеет исходные данные о предыдущих состояниях систем ИИ, а для изменения состояния можно использовать следующую модель, в которой выделены параметры, влияющие на функционирование систем ИИ:

$$\overline{f_k} = \overline{f_k}\left\{s_{i_1}^a; s_{i_2}^a; s_{i_3}^p; s_{i_4}^p; g_{\omega_k}; t_k; d_j; l_s; z_g; r_k\right\}, \quad (1)$$

где k – номер конкретного комплекса, k=1, …, K; i – номер признака, i=1, …, N, т.е. результата действия злоумышленников; $s_{i_1}^a$, $s_{i_2}^a$ – данные наблюдения предыдущих состояний систем ИИ, i1=1, …, m1, i2=m1+1, m2; $s_{i_3}^p$, $s_{i_4}^p$ – данные текущего контроля, i3=m2+1, m3, i4=m3+1, m4=m; ~ – знак, указывающий на признаки, которые подвержены влиянию данной совокупности внешних условий; g_{ω_k} – совокупность внешних условий, ωk=1, …, Ωk; tk – условная координата времени, показывающая полноту информации о k-м комплексе; dj – класс атаки на систему ИИ, j=1, …, N; ls – методы защиты, s=1, …, S; zg – последующие состояния k-й системы, g=1, …, G; rk – помехи, искажающие действительное состояние k-го комплекса.

Пусть известна модель защищаемого комплекса $\overline{f_k}$. Пусть l_s – защитные меры, которые могут системы ИИ $\overline{f_k}$ из состояния z_{g_i} перевести вероятность такого перехода. Обозначим через e_{s_j} меру результативности ls-го мероприятия защиты.

Тогда задачу нахождения оптимальной совокупности защитных мер можно сформулировать следующим образом. Необходимо найти такую совокупность защитных мер l_s^*, чтобы их результативность была достаточной. В этом случае результативность защиты:

$$e_{s_j}(z_{g_j} \to z_{g'})_{l_s} = p(d_j)_{z_g}\, p(z_{g_j} \to z_{g'})_{l_s}, \quad (2)$$

где $p(d_j)_{z_g}$ – вероятность проведения компьютерных атак, а максимальное значение меры результативности, или защита систем ИИ для $(z_{g_j} \to z_{g'})_{l_s}$, достигается при

$$e_{s_j}^* = \min_{l_s} e_{s_j}(z_{g_j} \to z_{g'})_{l_s} \quad (3)$$

Это определение действует для тех пар $(z_{g_j} \to z_{g'})_{l_s}$, для которых справедливо утверждение о том, что состояние zg' лучше, чем z_{g_j}.

Общая результативность защитных мер e_j является аддитивной функцией, состоящей из e_{s_j}.

Тогда для определения совокупности защитных воздействий можно сформулировать ограничение исследования: оптимальная совокупность защитных действий обладает тем свойством, что, каково бы не было первоначальное l_s при состоянии комплекса z_{g_j}, последующие защитные действия должны быть оптимальны относительно первоначального. Исходя из этого, искомую результативность защитных действий можно получить в следующем виде:

$$e_j = \max_{l_s} \sum | e_{s_j}(z_{g_j} \to z_{g'})_{l_s} + e_{s_j}(z_{g'} \to z_{g''})_{l_s} | . \quad (4)$$

Процесс обеспечения защищенности систем ИИ является сложным многоуровневым циклическим процессом, включающим в себя сбор и обработку данных различных систем, определение их состояния, выбор стратегии защиты и проведение защитных мер (рис. 2 и 3). В каждом цикле проводится оценка и коррекция управляющего воздействия. После того как установлены специфические особенности функционирования, можно переходить к управлению защиты системы ИИ.

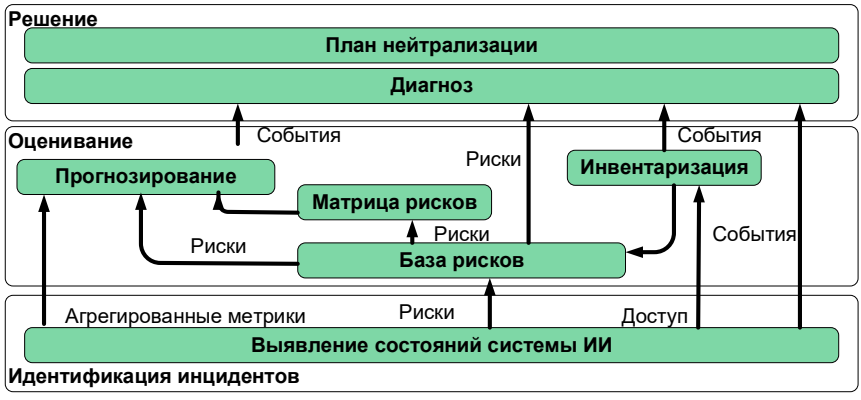

Рис. 2. Принципиальная схема обеспечения защиты систем ИИ.
Fig. 2. Schematic diagram of the protection of AI systems.

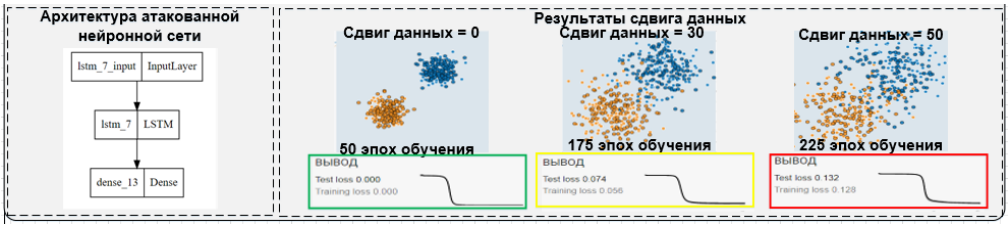

Рис. 3. Демонстрация результативности сдвига данных для 3-х слойной нейронной сети.
Fig. 3. Demonstration of the effectiveness of data shifting for a 3-layer neural network.

На каждом этапе защиты обрабатываются данные детектирования, потоков инцидентов, отслеживаемых за контролируемый период времени. При этом компонент прогнозирования определяет не только отдельные ситуации нарушений, но и всю возможную картину рисков реализации сценариев деградации систем ИИ.

Таким образом, выбор совокупности защитных мер начинается с выявления признаков нарушения функционирования (компьютерных атак) систем ИИ, последующей выработкой решений и нейтрализацией угроз целостности моделей машинного обучения.

3. Описание процесса моделирования

Для формирования семейства моделей оценивания опасностей деструктивного воздействия и потенциального ущерба на компоненты систем ИИ была выдвинута гипотеза о пригодности эпидемиологического моделирования ущерба штатному функционированию компонентов систем ИИ. Это связано с подобием структур и сходством параметров моделей машинного обучения и информационных систем (типы взаимосвязей представлены на рис. 4).

Результаты действий злоумышленника по деградации таких систем подобны ущербу в биологической популяции при распространении инфекционных заболеваний. Подобные модели уже использовались в области информационной безопасности, например, при моделировании распространения компьютерных вирусов и функционирование ботнетов.

Для нейронных сетей выявлена зависимость развития эпидемии от механизма распространения среди базовых элементов – нейронов.

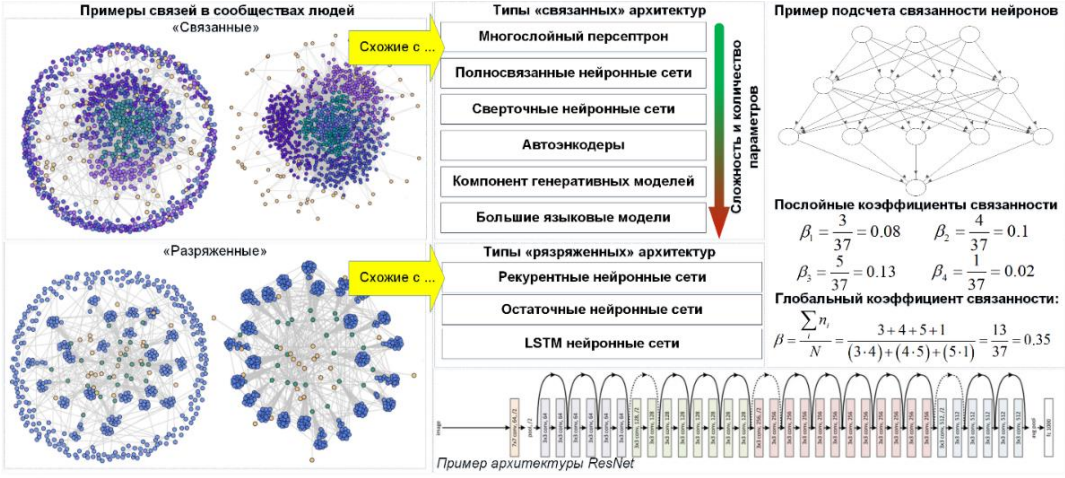

Рис. 4. Подобие структур и сходства параметров SIR-моделирования в нейронных сетях.
Fig. 4. Similarity of structures and similarities of parameters of SIR modeling in neural networks.

В связи с этим произведена адаптация SIR-модели:

$$\begin{aligned} \frac{dS_i}{dt} &= -\beta\rho S_i \sum_i A_{ij} I_j \\ \frac{dI_i}{dt} &= \beta\rho S_i \sum_i A_{ij} I_j - \gamma I_i \,, \\ \frac{dR_i}{dt} &= \gamma I_i \end{aligned} \tag{5}$$

где введены следующие параметры: β - коэффициент связанности нейронной сети; A_{ij} – матрица смежности, γ – коэффициент защищенности нейронной сети, ρ – степень инфицирования.

Решение системы дифференциальных уравнений следующее:

$$\frac{dS}{dt} = -\beta\rho S \frac{dR}{dt}\frac{1}{\gamma}, S = S_0 e^{-\frac{\beta\rho}{\gamma}R}$$
$$\frac{dI_i(t)}{dt} = \beta\rho(1-I_i)\sum_i A_{ij}x_j - \gamma I_i \quad (6)$$

которое формирует критерий опасности и выражается следующими условиями:

$$\beta\rho > \gamma, \quad R(\infty) = const > 0$$
$$\beta\rho < \gamma, \quad R(\infty) \to 0 \quad (7)$$

Интерпретацию параметров процесса деградации моделей машинного обучения начнем с описания базового элемента нейронных сетей. В основе всех нейронных сетей лежит базовый элемент – нейрон, который математически полностью аналогичен логистической регрессии (в случае если используется функция активации сигмоида). Это обусловлено тем, что данный элемент содержится во всех архитектурах нейронных сетей (рис. 5).

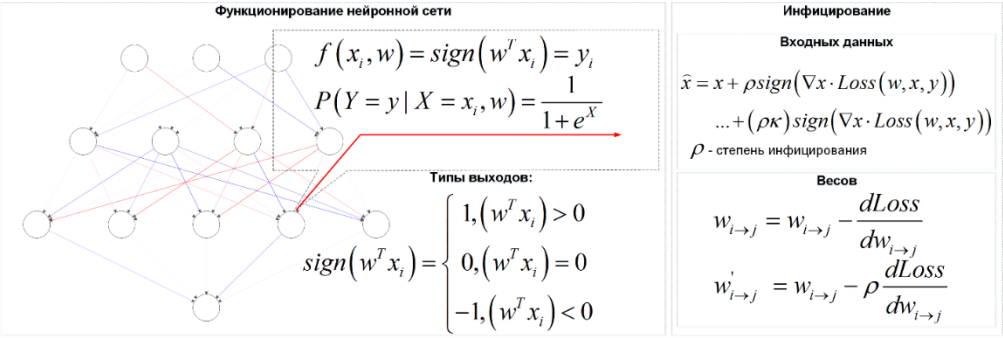

Рис. 5. Интерпретация процессов деградации моделей машинного обучения при деструктивных возмущениях.
Fig. 5. Interpretation of the degradation processes of machine learning models under destructive disturbances.

Параметрами любой нейронной сети являются - входные данные, веса, функция преобразования, функция активация. То есть, функционирование нейрона в обычных условиях происходит как $f: X \to Y$, где $\{y_1, ..., y_i\} \subset Y$ – пространство выходов, а $x = \{x_1, ..., x_j\} \subset X$ – пространство входных данных.

Основная цель компьютерных атак на системы ИИ – нарушить заявленное качество ее функционирования небольшими изменениями входных данных и весов моделей машинного обучения) так, чтобы модифицированные данные $\hat{x}$ вызовут сбои и ошибки $f(\hat{x}) \neq y$.

Дополнительно возникает необходимость маскировки таких воздействий, т. е. способность скрывать модифицированные данные.

Определим показатели эффективности атак на компоненты машинного обучения:

- метрика результативности, измеряющая несоответствие между обычными и модифицированными данными машинного обучения:

$$M_{res}(x_j, \hat{x}_j) = 1 - \frac{\left|f(x_j) \cap f(\hat{x}_j)\right|}{\left|f(x_j)\right|}, \quad (8)$$

другими словами, метрика определена в интервале [0, 1] и показывает, что при $M_{res}=1$ все модифицированные данные привели к неверному результату, и наоборот, при $M_{res}=0$ атака не достигла успеха;

- метрика скрытности, измеряющая возможность обнаружения модификации данных:

$$M_{hid}(y_j, \hat{x}_j) = \begin{cases} 0, \textit{если } f(\hat{x}_j) \approx y_j \\ 1, \textit{в другом случае} \end{cases}. \quad (9)$$

4. Экспериментальное исследование

Обоснование адекватности разработанных моделей и механизмов прогнозирования рисков потерь качества функционирования нейронных сетей при нарушениях целостности моделей машинного обучения с различной степенью деградации продемонстрировано двумя экспериментами: 1) сдвигом данных на простой нейронной сети, 2) при состязательных атаках на языковые модели машинного обучения.

Зависимость параметров сдвига данных представлено на примере целенаправленного сдвига данных для трехслойной нейронной сети (как видно на рис. 3 функция ошибки увеличивается при увеличении сдвига данных).

Обобщение результатов экспериментов по проведению состязательных атак на языковые модели машинного обучения (лингвистические модели) представлены на рис. 6 и показали сходимость результатов моделирования и экспериментов по воздействию на модели машинного обучения, представленные в табл. 1.

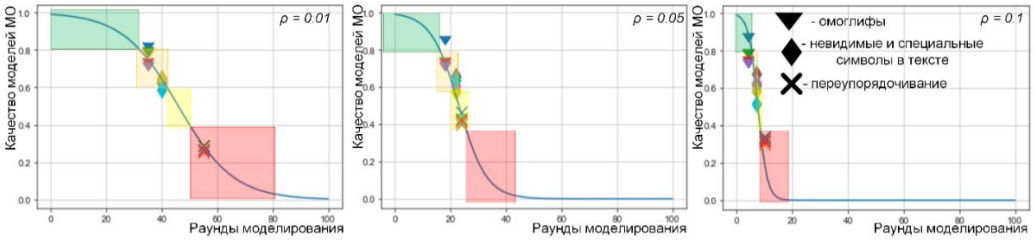

Рис. 6. Расчет уровня деградации для лингвистических моделей.
Fig. 6. Calculation of the degradation level for linguistic models.

Табл. 1. Лингвистические модели, отобранные для проведения эксперимента.
Table 1. Linguistic models selected for the experiment.

Название модели	Прикладная задача	Архитектура	Набор обучающих данных	Качество
IE-Net	Вопросно-ответные системы	Двоичные нейросети	SQuAD (Stanford QA Dataset)	F-мера=0,932
CB-NTR	Рубрикация текста	BERT	Reuters-21578	F-мера=0,907
ACE	Выявления поименованных сущностей	LSTM, Transformer	CoNLL-2003	F-мера=0,946
AraBERTv1	Семантический поиск	BERT	Large-Scale Arabic Book Reviews	Точность=0,867
Bi-LSTM	Выявление фейков	Bi-LSTM	FakeNewsNet	Точность=0,822

Для расчета уровня деградации для различных значений параметра инфицирования моделей $\rho \in (0.01, 0.05, 0.1)$ было произведено 100 раундов моделирования. Уровни деградации имеют следующие значения: незначительный – до 0.8, умеренный – от 0.8 до 0.6, высокий – от 0.6 до 0.4 и критический – ниже 0.4.

Для формирования текстовых состязательных примеров использованы обучающие данные для каждой лингвистической модели. На каждую модель машинного обучения сформировано 100 состязательных примеров с тремя долями модификации текста. Результаты представлены в табл. 2 и показали сходимость результатов моделирования и экспериментов.

Разработанные модели противодействия для формирования риск-ориентированного администрирования безопасности комплексов интеллектуальной обработки данных с возможностями самоконтроля и самовосстановления применяемы по следующей схеме: получение информации о моделях машинного обучения, определение начальных параметров модели, расчет ущерба и опасности. Комплексное применение дает возможность сформировать матрицу рисков для моделей машинного обучения. На рис. 7 более подробно рассмотрен отдельный сегмент матрицы рисков нейронной сети ResNet.

Табл. 2. Результаты модификации для текстовых состязательных примеров.
Table 2. Modification results for text adversarial examples.

Модель	Качество моделей машинного обучения после модификации								
	при доле омографов в тексте, %			при доле невидимых символов в тексте, %			при доле символов переупорядочивания, %		
	1	5	10	1	5	10	1	5	10
IE-Net	0,894	0,645	0,342	0,844	0,676	0,289	0,932	0,876	0,765
CB-NTR	0,804	0,639	0,299	0,809	0,630	0,253	0,865	0,735	0,715
ACE	0,811	0,620	0,337	0,818	0,618	0,289	0,847	0,762	0,730
AraBERTv1	0,771	0,645	0,315	0,757	0,584	0,270	0,762	0,754	0,720
Bi-LSTM	0,754	0,589	0,323	0,735	0,515	0,264	0,745	0,734	0,701

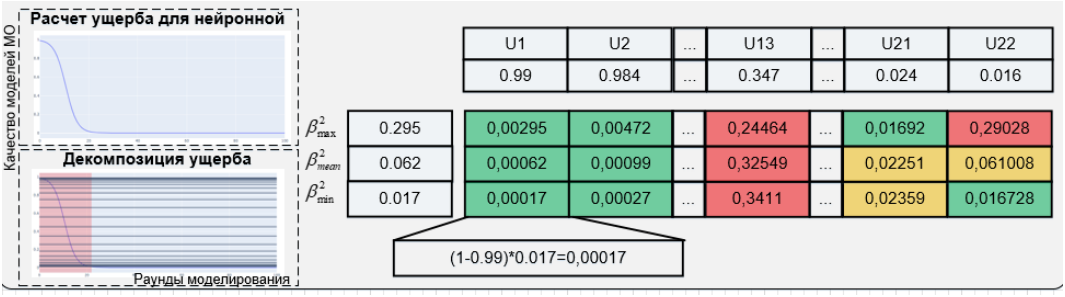

Рис. 7. Сегмент результатов моделирования рисков для нейронной сети типа ResNet.
Fig. 7. Segment of risk modeling results for a ResNet neural network.

Заключение

Таким образом, обоснована идеология адаптации модернизированных SIR-моделей природных эпидемий для выявления сходства в характере распространения деструктивных сбоев в системах ИИ, вызванных разнородно-массовыми и таргетированными воздействиями. Проведено обоснование пригодности семейства SIR-моделей для исследования процессов компьютерных атак на комплексы интеллектуальной обработки данных с моделями машинного обучения и разработки решений по противодействию. Выявленные закономерности позволили оценить риски возможного ущерба целостности и разработать эффективные стратегии предотвращения и

исправления искажений моделей машинного обучения. Разработанные математические модели пригодны для любой архитектуры нейронных сетей, а также для других парадигм машинного обучения. Модели позволяют сформировать обобщенную систему знаний для прогнозирования рисков нарушения целостности моделей машинного обучения в комплексах интеллектуальной обработки данных и уточнить требования к возможностям восстановления и самовосстановления.

Список литературы / References

[1]. Qin Y. et al. Artificial intelligence and economic development: An evolutionary investigation and systematic review //Journal of the Knowledge Economy. – 2024. – vol. 15. – №. 1. – p. 1736-1770.
[2]. Менисов А. Б., Ломако А. Г., Сабиров Т. Р. Метод тестирования лингвистических моделей машинного обучения текстовыми состязательными примерами //Научно-технический вестник информационных технологий, механики и оптики. – 2023. – т. 23. – №. 5. – с. 946-954.
[3]. Papagianni A. et al. Frugal and Robust AI for Defence Advanced Intelligence //Paradigms on Technology Development for Security Practitioners. – Cham: Springer Nature Switzerland, 2024. – p. 427-437.
[4]. Weng Y., Wu J. Leveraging artificial intelligence to enhance data security and combat cyber attacks //Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023. – 2024. – vol. 5. – №. 1. – p. 392-399.
[5]. Nguyen T. T. et al. Manipulating recommender systems: A survey of poisoning attacks and countermeasures //ACM Computing Surveys. – 2024. – vol. 57. – №. 1. – p. 1-39.
[6]. Rosenblatt M. et al. Data leakage inflates prediction performance in connectome-based machine learning models //Nature Communications. – 2024. – vol. 15. – №. 1. – p. 1829.
[7]. Kim S. et al. Propile: Probing privacy leakage in large language models //Advances in Neural Information Processing Systems. – 2024. – vol. 36.
[8]. Менисов, А. Б. Ландшафт угроз систем искусственного интеллекта: монография / А. Б. Менисов. — Москва: Ай Пи Ар Медиа, 2023. — 126 с.
[9]. Костогрызов А. И., Нистратов А. А. Анализ угроз злоумышленной модификации модели машинного обучения для систем с искусственным интеллектом //Вопросы кибербезопасности. – 2023. – №. 5. – с. 9.

Информация об авторах / Information about authors

Артем Бакытжанович МЕНИСОВ – кандидат технических наук, старший преподаватель кафедры систем сбора и обработки информации Военно-космической академии имени А.Ф.Можайского. Сфера научных интересов: построение доверенных систем искусственного интеллекта, применение машинного обучения для задач обеспечения информационной безопасности.

Artem Bakytzhanovich MENISOV – Cand. Sci. (Tech.), Senior lecturer of the Department of Information Collection and Processing Systems at the Mozhaysky Military Space Academy. Research interests: building trusted artificial intelligence systems, using machine learning for information security tasks.

Александр Григорьевич ЛОМАКО – доктор технических наук, профессор, профессор кафедры систем сбора и обработки информации Военно-космической академии имени А.Ф.Можайского. Его научные интересы включают области теоретического и системного программирования, моделирования интеллектуального поведения кибернетических систем в приложении к задачам информационной безопасности.

Alexander Grigorievich LOMAKO – Dr. Sci. (Tech.), Professor, Professor of the Department of Information Collection and Processing Systems at the Mozhaysky Military Space Academy. His research interests include the areas of theoretical and system programming, modeling of intelligent behavior of cybernetic systems in application to information security problems.