СПЕЦИАЛЬНАЯ МОДЕЛЬ БЕЗОПАСНОСТИ СОЗДАНИЯ И ПРИМЕНЕНИЯ СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Гарбук С. В.[1]

DOI: 10.21681/2311-3456-2024-1-15-23

Цель: *обоснование характера и структуры угроз, проявляющихся при создании и применении систем искусственного интеллекта на базе алгоритмов машинного обучения, в зависимости от видов нарушений требований, предъявляемых к информационным компонентам этих систем.*

Методы: *исследования проводились с использованием методов формальной логики, многокритериальной оценки, инженерии программных систем.*

Результат: *показано, что угрозы безопасности создания и применения систем искусственного интеллекта обусловлены нарушением требований в области целостности и доступности, предъявляемых к функциональным характеристикам систем и предусмотренным условиям их эксплуатации, к используемым эталонным архитектурам моделей машинного обучения, а также к обучающим, тестовым и входным наборам данных. При этом угрозы безопасности могут проявляться в виде деградации и повышения погрешности оценки функциональных характеристик (нарушение функциональности), а также компрометации чувствительных сведений о самих системах и о третьих лицах. Нарушение функциональности систем, в свою очередь, может приводить к реализации угроз физической, информационной и экономической безопасности. В статье приведены логические зависимости, позволяющие оценивать структуру потенциальных угроз безопасности в зависимости от степени соответствия комплексу предъявляемых требований в области целостности и конфиденциальности информационных компонент систем.*

Научная новизна: *полученные результаты могут быть использованы при обосновании требований к процессам жизненного цикла систем искусственного интеллекта на основе алгоритмов машинного обучения, а также при оценивании потенциальных рисков при создании и применении таких систем.*

Ключевые слова: *качество систем искусственного интеллекта, функциональная корректность систем искусственного интеллекта, риски создания и применения систем искусственного интеллекта.*

A SPECIAL SECURITY MODEL FOR THE CREATION AND APPLICATION OF ARTIFICIAL INTELLIGENCE SYSTEMS

Garbuk S. V.[2]

The goal of the investigation: *to substantiate the nature and structure of threats manifested in the creation and application of artificial intelligence systems based on machine learning algorithms, from the types of violations of the requirements imposed on the information components of these systems.*

Methods: *The research was conducted using methods of formal logic, multi-criteria evaluation, and software systems engineering.*

Result: *It is shown that the threats to the security of the creation and application of artificial intelligence systems are caused by violation of the requirements in the field of integrity and accessibility imposed on the functional characteristics of systems and the provided conditions of their operation, to the used reference architectures of machine learning models, as well as to training and test datasets. At the same time, security threats can manifest themselves in the form of degradation and increased error in evaluating functional characteristics (violation of functionality), as well as compromising sensitive information about the systems themselves and about third parties. Violation of the functionality of systems, in turn, can lead to the realization of threats to physical, information and economic security. The article presents logical dependencies that allow us to assess the composition of potential security threats depending on the degree of compliance with the set of requirements in the field of integrity and confidentiality of information components of systems.*

1 Гарбук Сергей Владимирович, кандидат технических наук, старший научный сотрудник, НИУ ВШЭ, г. Москва, Россия. ORCHID: 0000-0001-5385-3961, E mail: sgarbuk@hse.ru .

2 Sergey V. Garbuk, Ph.D., Senior Research Fellow, National Research University Higher School of Economics, Moscow, Russia. ORCHID: 0000-0001-5385-3961, E mail: sgarbuk@hse.ru.

Scientific novelty: *The results obtained can be used to substantiate the requirements for the life cycle processes of artificial intelligence systems based on machine learning algorithms, as well as to assess potential risks in the creation and application of such systems.*

Keywords: *Security of artificial intelligence systems, quality of artificial intelligence systems, functional correctness of artificial intelligence systems, risks of creation and application of artificial intelligence systems.*

Введение

Впечатляющие результаты в области технологий искусственного интеллекта (ИИ) на современном этапе обусловлены, прежде всего, развитием одного из направлений ИИ – алгоритмов машинного обучения (МО), обеспечивающих эффективное решение различных задач в области автоматизированной обработки данных в условиях отсутствия основанных на знаниях моделей наблюдаемых объектов и процессов. Наряду с универсальностью применения алгоритмы МО обладают также такими особенностями, как отсутствие полной интерпретируемости, возможность дообучения алгоритмов МО в процессе эксплуатации систем ИИ (СИИ), высокая актуальность вопросов социальной приемлемости применения СИИ, необходимость сравнения функциональных возможностей СИИ и человека-оператора и др. [1].

Модель жизненного цикла и требования к информационным компонентам систем искусственного интеллекта

Модель жизненного цикла (ЖЦ) для типовой системы ИИ, учитывающая приведенные выше особенности алгоритмов МО, представлена на рис. 1 [2]. На разных этапах ЖЦ СИИ используются различные информационные компоненты (данные, формализованные описания, модели), необходимые для успешной реализации этих этапов:

- ✓ функциональные требования к системам (ФТ), которые могут быть представлены в виде вектора $F = \{F_1, F_2, \ldots F_N\}$ [3];
- ✓ описание предусмотренных условий эксплуатации (ПУЭ), в общем случае заданное многомерной плотностью распределения $W(E)$ существенных факторов эксплуатации (СФЭ) СИИ $E = \{e_1, e_2 \ldots e_K\}$ [3], где K – количество СФЭ;
- ✓ эталонные архитектуры программного обеспечения, реализующего алгоритмы МО (требования к этой информационной компоненте могут быть представлены в виде набора R_R);
- ✓ обучающие и дообучающие НД (R_{L1} *и* R_{L2}, соответственно);
- ✓ тестовые НД ($R_{T1} = R_{T2} = R_T$);
- ✓ входные данные (R_U).

Отметим, что подобный набор информационных компонент специфичен для СИИ, так что особенности систем обработки данных на основе алгоритмов МО полностью обуславливаются этим набором как минимально достаточным. При этом в составе СИИ могут присутствовать иные компоненты (например сенсоры, средства передачи, обработки, хранения и отображения информации, исполнительные устройства и т.п), также определяющие качество и безопасность работы системы, но не являющихся специфическими для систем обработки данных на основе алгоритмов МО.

Таким образом, полный набор требований к СИИ может быть представлен в виде объединения: $R_C \cup F \cup W(E) \cup R_R \cup R_{L1} \cup R_T \cup R_U \cup R_{L2}$, где R_C – набор общих требований к СИИ, не зависящих от особенностей используемых алгоритмов МО; R – множество требований, специфичных для СИИ, которое для удобства можно переписать в виде: $R = R_1 \cup R_2 \cup R_3 \cup R_4 \cup R_5 \cup R_6 \cup R_7$ (индексы требований соответствуют обозначениям информационных компонент на рис. 1).

Тогда, если система S обладает множеством специфичных для алгоритмов МО характеристик $r(S) = \{r_n(S)\}_{n=1..7}$, то под специальной моделью безопасности СИИ будем понимать совокупность зависимостей, определяющих влияние несоответствий характеристик r системы S требованиям из набора R на характер угроз, проявляющихся при создании и применении этой системы.

Будем считать также, что требования к информационным компонентам из набора R могут быть двух видов:

- ✓ требования целостности (используется индекс in, где n – номер информационной компоненты, $n = 1..7$, рис. 1), заключающиеся в корректности формирования информационной компоненты и в предотвращении её умышленных и непреднамеренных искажений;
- ✓ требования конфиденциальности (cn), заключающиеся в предотвращении компрометации соответствующих данных.

Далее будут рассмотрены наиболее типичные примеры, иллюстрирующие негативные последствия, вызванные отклонением от выполнения тех или иных требований. При этом будет приниматься во внимание, что среди характеристик СИИ $r(S)$ следует выделять как целевые характеристики (прежде всего, это ФХ СИИ $r_1(S) = f(S)$, подтвержденные в конкретных условиях эксплуатации $r_2 = w(E)$, а также характеристики конфиденциальности самой СИИ и прочих субъектов информационной безопасности),

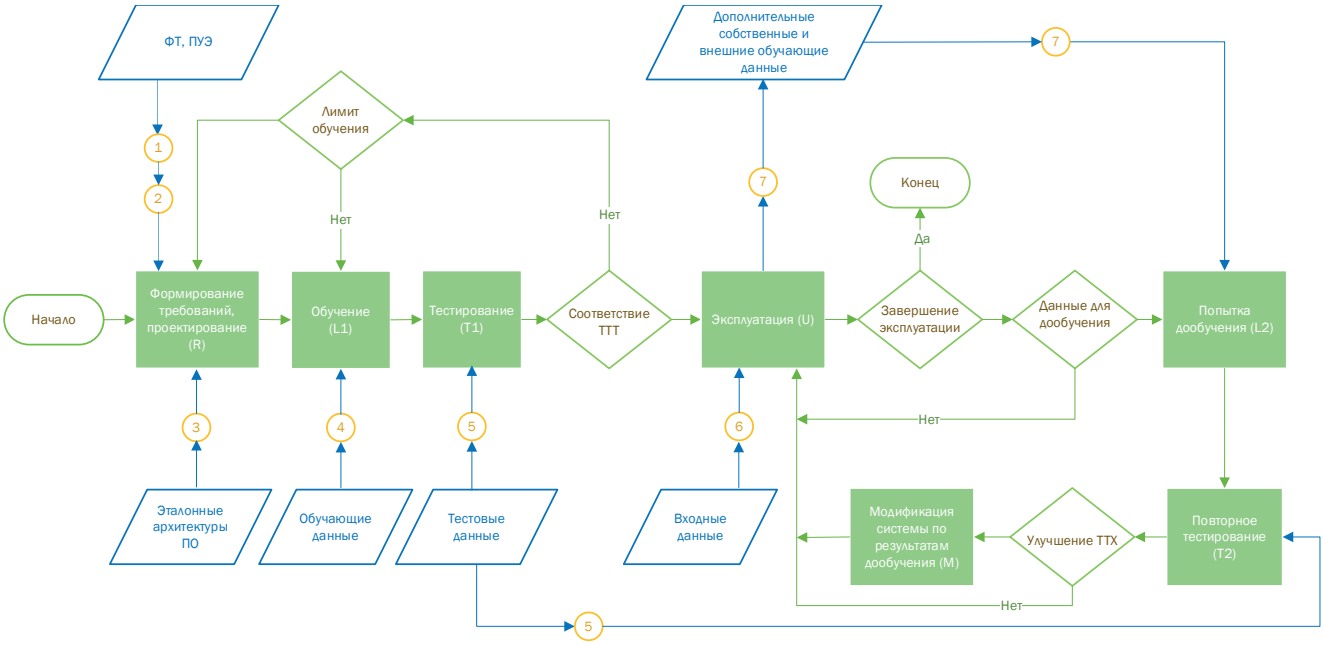

Рис.1. Модель ЖЦ СИИ. Цифрами на рисунке обозначены точки возникновения несоответствия требованиям, установленным к различным информационным компонентам

так и «процессные» характеристики, важные не сами по себе, а лишь в силу влияния этих характеристик на целевые (например, характеристики целостности и конфиденциальности используемых НД).

Угрозы безопасности при создании и применении систем искусственного интеллекта

Нарушение целостности ФТ (нарушение требований вида $i1$, то есть $r_{i1} \neq R_{i1}$) может быть обусловлено недостаточной представительностью выбранного набора существенных характеристик и показателей качества СИИ[3], некорректными весовыми коэффициентами частных показателей качества, влияющих на точность определения интегрального показателя, неправильно выбранными пороговыми значениями показателей качества. В свою очередь, нарушение конфиденциальности ФТ (нарушение требований $c1$) является нежелательным при наличии активного злоумышленника.

Выполнение требований целостности к предусмотренным условиям эксплуатации (ПУЭ) СИИ (требования вида $i2$) обеспечивает возможность формирования репрезентативных обучающих НД и выбора оптимальных архитектур СИИ, проведения репрезентативных испытаний системы, а также возможность принятия обоснованных решений по применению СИИ. Таким образом, нарушение выполнения требований $i2$ будет приводить как к непосредственной деградации функциональных характеристик (ФХ) СИИ, так и к нарушению целостности выбранных архитектур систем и обучающих НД, что может быть записано в виде:

$$(r_{i2} \neq R_{i2}) \rightarrow (r_{i3} \neq R_{i3}); (r_{i2} \neq R_{i2}) \rightarrow (r_{i4} \neq R_{i4}). \quad (1)$$

Наиболее опасным является нарушение требований к целостности входных данных, обрабатываемых в СИИ (требования $i6$). Создаваемые системы ИИ, в частности – реализованные на искусственных нейронных сетях (ИНС), оказываются неустойчивы к небольшим искажениям входных данных. Недостаточное выполнение требований i6 приводит к завышению оценок ФХ СИИ, так как в процессе испытания состязательные атаки как фактор, снижающий качество работы системы, учитывается недостаточно.

Невыполнение требований к конфиденциальности сведений об использованных архитектурах СИИ (несоответствие требованиям $c3$), обучающих ($c4$), тестовых ($c5$) и дообучающих ($c6$) НД создает дополнительные риски невыполнения требования к целостности входных данных [4, 5]:

$$(r_{c3} \neq R_{c3}) + (r_{c4} \neq R_{c4}) + (r_{c5} \neq R_{c5}) + (r_{c7} \neq R_{c7}) \rightarrow \rightarrow (r_{i6} \neq R_{i6}). \quad (2)$$

В процессе работы СИИ уровень конфиденциальности данных, накапливаемых и обрабатываемых в системе, может возрастать, что потенциально может привести к неправильному определению (занижению) требований в области ИБ на стадии проектирования СИИ. Кроме того, учитывая, что входные данные с высоким уровнем конфиденциальности используются в данном случае и для дообучения

3 ГОСТ Р 59898–2021 Оценка качества систем искусственного интеллекта. Общие положения. Введен 2022 03-01. М.: Российский институт стандартизации. 2021, 20 с.

СИИ, нарушение требований вида $c1$ создаст предпосылки для успешной реализации атаки на конфиденциальность обучающих данных СИИ:

$$(r_{c1} \neq R_{c1}) \rightarrow (r_{c4} \neq R_{c4}), (r_{c1} \neq R_{c1}) \rightarrow (r_{c7} \neq R_{c7}). \quad (3)$$

Нарушение требований к целостности архитектур СИИ и обучающих НД в дальнейшем повышает эффективность реализации атак на целостность входных данных:

$$(r_{i3} \neq R_{i3}) + (r_{i4} \neq R_{i4}) \rightarrow (r_{i6} \neq R_{i6}). \quad (4)$$

Предотвращение атак на целостность архитектуры СИИ и обучающих НД заключается, прежде всего, в использовании эталонных моделей МО из доверенных источников [6], а также в реализации разработчиком и эксплуатантом (в случае дообучения СИИ в процессе эксплуатации) необходимого комплекса мер информационной безопасности.

Невыполнение требований к целостности тестовых, как и обучающих, НД (требования вида $i5$ может приводить к нарушению их репрезентативности, точности, достоверности и к реализации других факторов снижения качества. При этом функциональные характеристики систем ИИ не изменяются, но оценка этих характеристик, получаемая при испытании систем с использованием модифицированных тестовых НД, может содержать недопустимо высокую погрешность. Следует различать две составляющие погрешности оценки ФХ СИИ:

- ✓ смещение характеристик, вызванное нарушением баланса тестовых примеров различной сложности. Относительная нехватка сложных примеров приводит к неоправданному завышению (чрезмерно оптимистичной оценке) ФХ, избыток сложных примеров – к недооценке ФХ;
- ✓ возрастание случайной составляющей погрешности, обусловленной недостаточным объёмом тестового НД.

Отметим, что возникновение предпосылок к завышению функциональных возможностей испытуемых алгоритмов МО, может объясняться как естественными причинами (например, недостаточной квалификацией персонала, подготавливающего тестовые НД), так и умышленными действиями злоумышленника. В обоих случаях наличие существенных погрешностей в оценке функциональных характеристик может привести к некорректному применению СИИ по назначению, а в случае умышленных искажений – еще и создать дополнительные предпосылки для реализации противником эффективных атак на входные данные:

$$(r_{i5} \neq R_{i5}) \rightarrow (r_{i6} \neq R_{i6}). \quad (5)$$

Таким образом, анализ рисков, проявляющихся при невыполнении требований типа in и cn, показал, что возможными негативными последствиями, обусловленными несоответствием требованиям, специфичных для СИИ на основе алгоритмов МО, являются:

1) существенное возрастание ошибки оценивания ФХ при тестировании (испытаниях) СИИ за счет смещения (как правило – в сторону завышения характеристик) и возрастания случайной составляющей погрешности оценок при снижении вариативности тестовых НД. Неточное понимание функциональности систем существенно усложняет или даже делает невозможным принятие эксплуатирующей стороной рациональных решений об использовании СИИ на практике;
2) деградация ФХ, ограничивающая возможность применения систем в реальных условиях эксплуатации. Причины такой деградации заключаются либо во внесении преднамеренных искажений в обучающие НД и архитектуру СИИ, в результате чего ФХ ухудшаются в предусмотренных условиях эксплуатации, либо в создании злоумышленниками в ходе реального применения СИИ условий применения, существенно отличающихся от предусмотренных разработчиками системы (ПУЭ). Во втором случае ФХ систем сохраняют гарантированные разработчиком значения в предусмотренных условиях эксплуатации, однако деградируют в реальных условиях, выходящих за рамки ПУЭ;
3) нежелательное нарушение конфиденциальности сведений о тактико-технических характеристиках и особенностях применения СИИ, приводящее, например, к повышению эффективности деструктивных информационных воздействий на СИИ злоумышленниками, в том числе – за счет оптимизации способов искажения входных данных СИИ и т.п.;
4) компрометация сведений о физических и юридических лицах, интересы которых так или иначе затрагиваются при реализации процессов ЖЦ СИИ (заинтересованные лица СИИ).

Зависимости возможных негативных последствий, обусловленных нарушением специальных требований при создании и применении СИИ, представлены в табл. 1. В таблице использованы обозначения inm и cnm для зависимостей, характеризующих влияние нарушения соответственно целостности и конфиденциальности n-й информационной компоненты на m-е негативное последствие, $m = 1..4$.

Таблица 1.

Матрица специальных требований к информационным компонентам СИИ и возможных негативных последствий, обусловленных невыполнением этих требований

Объект управления качеством и безопасностью (информационная компонента СИИ)	Аспект управления качеством и безопасностью (вид требований к информационной компоненте)	Возможные негативные последствия от несоответствия требованиям			
		1. Рост погрешности оценки ФХ при тестировании СИИ	2. Деградация ФХ СИИ в ПУЭ	3. Компрометация ТТХ и вариантов применения СИИ	4. Компрометация данных о заинтересованных лицах СИИ
1. Функциональные требования (ФТ) к СИИ	$i1$ – полнота набора и обоснованность критериальных значений (в том числе – путем оценки возможностей человека-оператора) ФТ, обоснованность метрик для сравнения функциональных характеристик (ФХ) с ФТ	$i11$ – смещение оценки интегрального показателя качества	$i12 = 0$	$i13 = 0$	$i14 = 0$
	$c1$ – конфиденциальность ФТ	$c11 = 0$	$c12 = 0$	$c13$ – прямая компрометация ТТХ	$c14 = 0$
2. Описание предусмотренных условий эксплуатации (ПУЭ)	$i2$ – полнота набора существенных факторов эксплуатации (СФЭ), соответствие законов распределения СФЭ в ПУЭ и в реальных условиях применения СИИ	$i21$ – смещение и возрастание случайной составляющей погрешности оценки ФХ[41]	$i22$ – ухудшение ФХ вследствие неверного выбора обучающих НД и архитектур СИИ	$i23 = 0$	$i24 = 0$
	$c2$ – конфиденциальность ПУЭ, предотвращение компрометации сведений, которые могут быть использованы злоумышленником для целенаправленного создания условий, выходящих за пределы ПУЭ	$c21$ – смещение оценки ФХ	$c22$ – ухудшение ФХ при создания злоумышленником условий, выходящих за ПУЭ	$c23$ – компрометация ТТХ в части ПУЭ	$c24 = 0$
3. Архитектура СИИ	$i3$ – отсутствие программных закладок, обеспечивающих злоумышленнику повышенные по сравнению с предполагавшимися при тестировании возможности по реализации информационных атак на входные данные (например, по подбору эффективных состязательных атак)	$i31$ – смещение оценки ФХ	$i32$ – ухудшение ФХ	$i33 = 0$	$i34 = 0$
	$c3$ – конфиденциальность сведений об архитектуре СИИ, которые могут быть использованы для реализации эффективных атак на входные данные	$c31$ – смещение оценки ФХ	$c32 = 0$	$c33$ – раскрытие уязвимостей к атакам на входные данные	$c34 = 0$

4 Смещение оценки функциональных характеристик, вызванное несоответствием установленным к СИИ требованиям, как правило приводит к завышению (получению излишне оптимистичных) оценок.

4. Обучающие НД	$i4$ – предотвращение целенаправленных искажений («отравления») обучающих НД, приводящих к повышенным по сравнению с предполагавшимися при тестировании возможностям злоумышленника по реализации информационных атак на входные данные (например, по подбору эффективных состязательных атак)	$i41$ – смещение оценки ФХ	$i42$ – ухудшение ФХ	$i43 = 0$	$i44 = 0$
	$c4$ – конфиденциальность обучающих НД, предотвращение компрометации сведений, облегчающих последующую реализацию эффективных атак на входные данные	$c41$ – смещение оценки ФХ	$c42 = 0$	$c43$ – раскрытие ПУЭ (вариативность обучающего НД) и уязвимостей к атакам на входные данные	$c44$ – раскрытие конфиденциальных данных по объектам, на которых происходит обучение
5. Тестовые НД	$i5$ – предотвращение искажений, влияющих на репрезентативность тестового НД – смещение (нарушение баланса тестовых сценариев разной сложности) и снижение вариативности (сокращение объема выборки, большое число дублирующих тестовых сценариев)	$i51$ – смещение и возрастание случайной составляющей погрешности оценки ФХ	$i52 = 0$	$i53 = 0$	$i54 = 0$
	$c5$ – предотвращение компрометации тестовых НД, способной привести к снижению представительности тестирования (доступ к тестовым НД разработчиков приводит к переобучению СИИ, доступ злоумышленников – повышению возможностей по реализации информационных атак на входные НД	$c51$ – смещение оценки ФХ	$c52 = 0$	$c53$ – раскрытие ПУЭ (вариативность тестового НД) и уязвимостей к атакам на входные данные	$c54$ – раскрытие конфиденциальных данных по объектам, на которых сформирован тестовый НД
6. Входные данные СИИ	$i6$ – защита от искажений, в том числе – умышленных (например, состязательные атаки) и естественных	$i61$ – смещение оценки ФХ	$i62 = 0$	$i63 = 0$	$i64 = 0$
	$c6$ – конфиденциальность входных данных, в том числе, с учетом возрастания уровня их конфиденциальности при накоплении	$c61 = 0$	$c62 = 0$	$c63$ – раскрытие сценариев применения СИИ	$c64 = 0$

7. Дообучающие НД	$i7$ – достоверность и информативность дообучающих НД, предотвращение статистического смещения дообучающего НД, вызванного существенным отклонением условий эксплуатации СИИ от предусмотренных	$i71$ – смещение оценки ФХ для ПУЭ	$i72$ – ухудшение ФХ в ПУЭ	$i73 = 0$	$i75 = 0$
	$c7$ – конфиденциальность дообучающих НД, предотвращение компрометации сведений, облегчающих последующую реализацию эффективных атак на входные данные	$c71$ – смещение оценки ФХ	$c72 = 0$	$c73$ – раскрытие реализованных сценариев применения СИИ	$c74$ – раскрытие данных по объектам дообучения СИИ

Проиллюстрированные в табл. 1 зависимости характера потенциального ущерба от вида нереализуемых требований к информационным компонентам СИИ могут быть использованы для прогнозирования возможных негативных последствий, специфичных для алгоритмов МО. Для этого зависимости из табл. 1 должны быть представлены в виде матрицы чувствительности S, состоящей из векторов-столбцов s_{nm}, вида:

$$S = \{s_{nm}\} = \{\begin{matrix} inm \\ cnm \end{matrix}\}, \; n = 1..7, \; m = 1..4. \qquad (6)$$

Тогда, если вектор соответствия требованиям будет записан в виде вектора-строки:

$$A = \{i1, \; c1, \; i2, \; c2... \; i7, \; c7\}, \qquad (7)$$

где in и cn – бинарные коэффициенты, отражающие несоответствие (1) или соответствие (0) требованиям к целостности и конфиденциальности n-й информационной компоненты СИИ (табл. 1), то выражение для оценки параметров, характеризующих возможные негативные последствия, примет вид:

$$D = \{d_1, \; d_2... \; d_4\} = AS, \qquad (8)$$

а значение интегрального показателя уровня рисков для вектора соответствия требованиям A рассчитывается путем свертки параметров d_i с соответствующими весовыми коэффициентами:

$$d_0 = \Sigma_{i=1}^{4} \; \nu_i \, d_i \, . \qquad (9)$$

Для вычисления вектора параметров D и интегрального показателя D_0 значения элементов матрицы s_{nm} должны быть заданы на шкале отношений, допускающей операцию сложения, с учетом разнородного характера негативных последствий, обусловленных несоответствием различным требованиям.

Анализ рисков нарушения конфиденциальности данных должен проводиться с учетом специфики конкретной СИИ. В то же время, для рисков нарушения функциональной корректности могут быть сформулированы некоторые общие закономерности. Структура рисков, обусловленных нарушением функциональной корректности, с учетом перечисленных выше особенностей СИИ и в разрезе интересов и приоритетов различных заинтересованных сторон представлена в табл. 2.

Наиболее опасные угрозы и, соответственно, наиболее важные характеристики и требования связаны с обеспечением безопасности жизни и здоровья людей, а также с предотвращением крупных инцидентов экологической безопасности [7, 8]. Так или иначе, для систем повышенной опасности (п. 1 в табл. 2) необходимо иметь гарантии того, что уровень формируемых ими угроз не превышает уровень, демонстрируемый квалифицированными людьми-операторами, выполняющими соответствующие задачи управления и обработки данных в ручном режиме.

Особый вид требований связан с предотвращением угроз информационной безопасности (ИБ) в отношении заинтересованных лиц, вызванных нарушением функциональной корректности СИИ (п. 2 в табл. 2). Если некорректная работа систем может привести к реализации деструктивных информационно-психологических воздействий на общество (дезинформация, злонамеренное нарушение социальной стабильности), то в формировании требований ИБ к таким СИИ заинтересованы общество в целом и соответствующие государственные регуляторы (п. 2.2 в табл. 2).

Для многих прикладных СИИ специфичны угрозы этического характера и другие угрозы, предотвращение которых достигается реализацией мер т.н. «мягкого» права (п. 3 в табл. 2). К таким СИИ относятся, например, системы в кредитно-финансовой сфере и в области образования, поисково-справочные, маркетинговые и иные информационные системы, использующие методы персонализации на основе ИИ [9, 10].

Таблица 2.

Приоритеты заинтересованных сторон в области предотвращения угроз, обусловленных несоответствием функциональных характеристик СИИ установленным требованиям

Вид угроз, обусловленных нарушением функциональной корректности СИИ	Категория заинтересованной стороны	
	Лица, непосредственно участвующие в создании и применении СИИ (акторы ИИ)	Третьи лица
1. Угрозы жизни и здоровью людей, экологические угрозы	1.1. Потребители, разработчики и поставщики (собственная безопасность, дополнительные требования гос. регуляторов)	1.2. Общество в целом и регуляторы (безопасность общества и окружающей среды)
2. Угрозы информационной безопасности в отношении заинтересованных сторон	Нет	2.2. Общество в целом и государственные регуляторы (защита персональных данных, предотвращение деструктивных информационно-психологических воздействий)
3. Нарушение этических и других норм «мягкого» права	Нет	3.2. Общество в целом (социальная приемлемость создания и применения СИИ)
4. Неопределенные потребительские свойства, не влияющие непосредственно на безопасность жизни и здоровья людей, экологическую безопасность	4.1. Потребители (функциональные характеристики, определяющие возможность применения СИИ по назначению), разработчики и поставщики (характеристики конкурентоспособности СИИ)	Нет

Для СИИ, не предназначенных непосредственно для решения задач в области безопасности и не представляющих угрозы для жизни, здоровья людей и окружающей природной среды (п. 4 в табл. 2), отклонение функциональных характеристик от установленных требований ограничивается ухудшением потребительских свойств систем и может интерпретироваться, как реализация угроз экономической безопасности акторов ИИ. В формировании требований и предотвращении соответствующих угроз в таком случае заинтересованы, прежде всего, разработчики, поставщики и потребители СИИ (п. 4.1 в табл. 2).

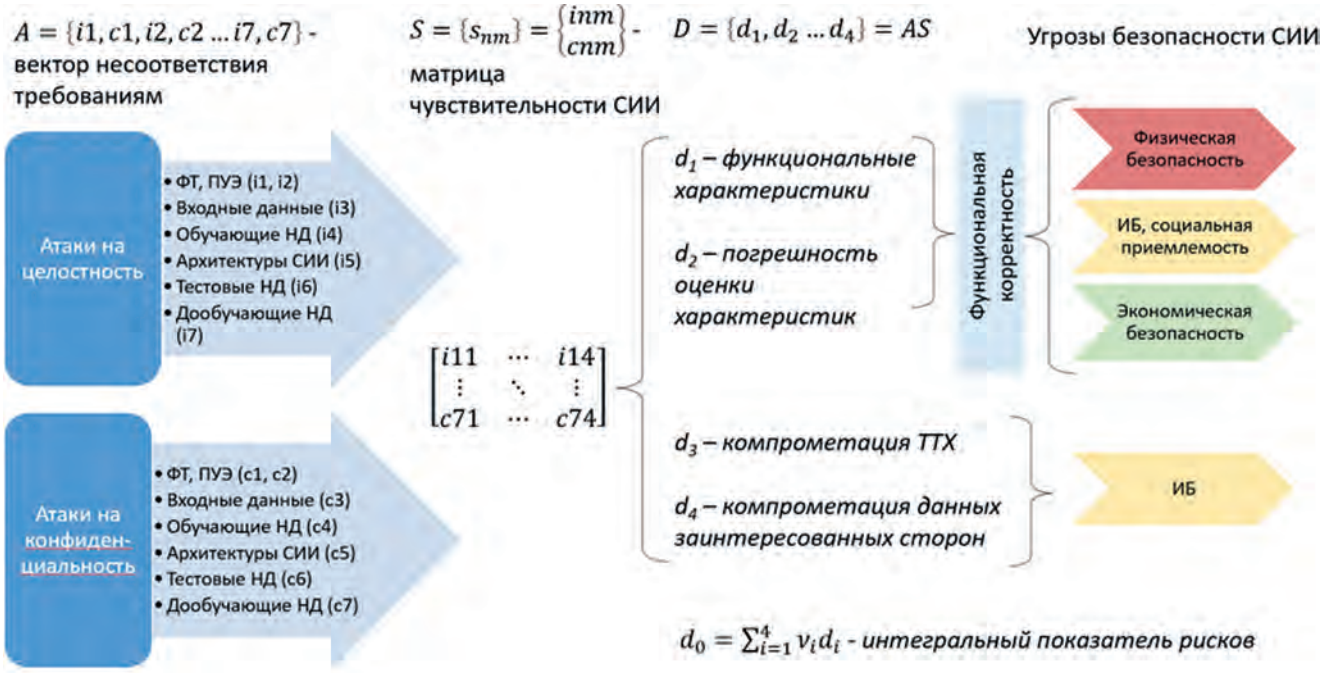

Рис.2. Специальная модель безопасности СИИ

Схема, поясняющая предложенную специальную модель безопасности СИИ, представлена на рис. 2. На основе данных о имеющихся и потенциальных несоответствиях требованиям в области целостности и конфиденциальности различных информационных компонент СИИ (A) и матрицы чувствительности СИИ к этим несоответствиям (S) оценивается вектор возможных негативных последствий (D), содержащий параметры функциональной корректности (d_1 и d_2) СИИ и конфиденциальности данных о СИИ и заинтересованных сторонах (d_3 и d_4, соответственно). Параметры функциональной корректности далее подвергаются дополнительному анализу с учетом прикладного значения и особенностей применения СИИ для оценивания рисков в области физической, ментальной (ИБ и социальная приемлемость) и экономической безопасности, после чего полученные оценки в сочетании с ранее полученными оценками рисков ИБ используются в качестве выходных данных модели. При необходимости может быть рассчитан показатель интегральных рисков, связанных с несоответствием СИИ установленным требованиям (d_0).

Предложенная модель построена с учетом общих особенностей систем обработки данных на основе алгоритмов машинного обучения, является качественной и иллюстрирует структуру зависимостей различного рода рисков, обусловленных несоответствием информационных компонентов СИИ различным требованиям. Количественное описание этих зависимостей представляет предмет дальнейших исследований с учетом отраслевой специфики прикладных систем ИИ.

Заключение

Таким образом, в статье рассмотрено влияние несоответствия различного рода требованиям, предъявляемым к информационным компонентам систем ИИ, на риски, возникающие при создании и применении этих систем. Анализ структуры зависимостей рисков с учетом общих особенностей систем обработки данных на основе алгоритмов машинного обучения показал, что эти риски сводятся к ухудшению функциональных характеристик СИИ, увеличению погрешности оценки этих характеристик эксплуатантом и компрометации данных, обрабатываемых в системе, причем часть рисков в области конфиденциальности данных представляет непосредственную угрозу, приводя к компрометации сведений об особенностях систем и о связанных с ними заинтересованных сторонах, а часть – создает дополнительные предпосылки для снижения функциональной корректности СИИ. Предложенная модель безопасности может быть использована для качественной оценки рисков, возникающих при создании и эксплуатации прикладных систем ИИ различного назначения, и организации мер по снижению этих рисков.

Литература

1. *Гарбук С. В., Губинский А. М. Искусственный интеллект в ведущих странах мира: стратегии развития и военное применение. – М.: Знание, 2020. 590 с.*
2. *Гарбук С. В. Метод оценки влияния параметров стандартизации на эффективность создания и применения систем искусственного интеллекта // Информационно-экономические аспекты стандартизации и технического регулирования. 2022. № 3. С. 4–14.*
3. *Garbuk S. V. Intellimetry as a way to ensure AI trustworthiness // The Proceedings of the 2018 International Conference on Artificial Intelligence Applications and Innovations (IC-AIAI). Limassol, Cyprus, 6-10.10.2018. pp. 27–30.*
4. *Войнов Д. М., Ковалев В. А. Экспериментальная оценка состязательных атак на глубокие нейронные сети при решении задач распознавания медицинских изображений // Информатика., 2019. Т. 16. №3. С. 14–22.*
5. *Gary McGraw, Richie Bonett, Harold Figueroa, Victor Shepardson. Security Engineering for Machine Learning. Computer. IEEE Computer Society, 2019, vol.52, no. 8. pp. 54-57.*
6. *Унифицированная программная платформа машинного обучения «Платформа-ГНС» [Электронный ресурс] // Сайт ГосНИИАС. URL: https://www.gosniias.ru/platform.html (дата обращения: 01.09.2023).*
7. *Patrick Hall, James Curtis, Parul Pandey. Machine Learning for High-Risk Applications. Approaches for Responsible AI. 2023 April. 470 p.*
8. *Rahman, M. M. Should I Be Scared of Artificial Intelligence? // Academia Letters, Article 2536. DOI: https://doi.org/10.20935/AL2536.*
9. *O'Keefe K., Daragh O Brien. Ethical Data and Information Management: Concepts, Tools and Methods, Kogan Page. 2018. pp. 46-47, 214–218, 262-263.*
10. *Карпов В. Э., Готовцев П. М., Ройзензон Г. В. К вопросу об этике и системах искусственного интеллекта // Философия и общество. 2018. №2. С.84–105.*