

# СИСТЕМОТЕХНИЧЕСКИЕ ПОДХОДЫ К ОБЕСПЕЧЕНИЮ БЕЗОПАСНОСТИ СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА С УЧЕТОМ КАЧЕСТВА ДАННЫХ ДЛЯ МАШИННОГО ОБУЧЕНИЯ

Зацаринный А.А., Гаврилов В.Е.

*Федеральный исследовательский центр «Информатика и управление» Российской академии наук, Москва, Россия, Вавилова 44/2, e-mail: frccsc@frccsc.ru.*

---

В статье рассмотрены некоторые проблемы обеспечения безопасности систем искусственного интеллекта, связанные с качеством данных для машинного обучения. Дан краткий анализ действующего стандарта, определяющего процессы на всех стадиях жизненного цикла данных. Сформулированы некоторые рекомендации по обеспечению безопасности систем искусственного интеллекта (СИИ) при реализации процессов жизненного цикла данных.

---

Ключевые слова: информационная безопасность, комплексная безопасность, функциональная безопасность, системы с искусственным интеллектом, техническая защита информации.

## SYSTEM ENGINEERING APPROACHES TO SECURING THE SAFETY OF ARTIFICIAL INTELLIGENCE SYSTEMS, TAKING INTO ACCOUNT THE QUALITY OF DATA FOR MACHINE LEARNING

Zatsarinny Alexander A., Gavrilov Victor E.

*Federal Research Center «Computer Science and Control» of the Russian Academy of Sciences, Moscow, Russia  
e-mail: frccsc@frccsc.ru*

---

The article discusses some of the challenges of ensuring the security of artificial intelligence systems related to the quality of data for machine learning. It provides a brief analysis of the current standard that defines processes at all stages of the data lifecycle. The article also offers some recommendations for ensuring the security of artificial intelligence systems (AIS) during the implementation of data lifecycle processes.

---

Keywords: information security, comprehensive security, functional security, artificial intelligence systems, technical information protection.

### Введение

В условиях развития цифровой трансформации особое место во всех сферах деятельности занимают информационно-аналитические системы, развитие которых все чаще связывают с широким применением технологий искусственного интеллекта (ТИИ), которое предусмотрено в рамках реализации национального проекта «Экономика данных» [1]. ТИИ сегодня являются одним из ключевых направлений мирового технологического прогресса, т.к. с каждым годом все глубже проникают в различные сферы развития общества и определяют будущее многих отраслей. На необходимость развития этого направления в России постоянно обращает внимание Президент России В.В. Путин. Достаточно упомянуть международную конференцию «Путешествие в мир искусственного интеллекта» [2], конференцию «Сбербанка» AI Journey-2025 [3], совещание по вопросам развития автономных систем на площадке Большой кольцевой линии Московского метрополитена [4]. В рамках этих мероприятий В.В. Путин отмечает стратегическую значимость ТИИ, которые должны стать важнейшим ресурсом для достижения национальных целей развития, укрепления обороноспособности страны, качественного развития экономики, социальных отраслей и госуправления. Даны конкретные поручения Правительству Российской Федерации, направленные на ускорение развития ТИИ (переход государственных органов на использование систем ИИ с применением платформенного подхода, увеличение числа бюджетных мест в вузах для подготовки специалистов по специальностям ИИ, реализация комплекса мер по увеличению вычислительных мощностей суперкомпьютеров в России и другие).

Вместе с тем, целенаправленное и эффективное выполнение приведенных поручений в такой огромной и ресурсоемкой стране как Россия невозможно без принятия соответствующих мер по обеспечению безопасности систем с искусственным интеллектом (СИИ), включая нормативное регулирование. В [5] были предложены подходы к развитию существующей нормативной базы в области обеспечения комплексной безопасности СИИ.

Вместе с тем, учитывая, что наиболее применимыми являются технологии машинного обучения особую значимость, приобретает проблема качества обучающих данных. Эта проблема является одной из ключевых при создании доверенных СИИ [6-9].

В настоящей статье рассматриваются некоторые аспекты обеспечения безопасности СИИ применительно к различным стадиям их создания, связанные с качеством данных для машинного обучения.

### **1. Основные подходы к обеспечению комплексной безопасности СИИ**

В отличие от традиционных подходов к обеспечению информационной безопасности автоматизированных систем (АС), в которых основной упор сделан на обеспечение технической защиты информации, применительно к АС, использующих интеллектуальные технологии, прежде всего ТИИ, эту проблему необходимо рассматривать в контексте комплексной безопасности, которая определяется во взаимосвязке трех составляющих [10]:

- технической безопасности, направленной на обеспечение конфиденциальности, целостности и доступности обрабатываемой и хранимой информации;
- функциональной безопасности, направленной на обеспечение устойчивого и корректного выполнения заданных функций;
- системной безопасности, связанной с корректностью, точностью и полнотой постановки задачи.

### **2. Качество обучающих данных как фактор обеспечения безопасности систем искусственного интеллекта**

Качество СИИ на основе технологии машинного обучения (МО) в значительной мере определяется качеством обучающих данных. В серии стандартов [12] – [16] представлен комплексный подход к оценке и повышению качества данных для аналитики и машинного обучения, включая основные определения [12], описание модели качества данных и перечень показателей для аналитики и машинного обучения [13], требования к управлению качеством данных [14], содержит организационные подходы к обеспечению качества данных [15], структуру стратегического управления качеством данных [16].

В стандарте процессного характера определен порядок сбора, подготовки и использования данных на всех стадиях жизненного цикла данных [17]. Однако, в нем не отражены в явном виде некоторые проблемы обеспечения безопасности информации СИИ, в т.ч. в части оценки качества данных, которые являются важнейшими для обеспечения функциональной безопасности СИИ [10]. Отчасти проблема обеспечения и поддержания качества данных, как составной части общей системы безопасности СИИ, рассматривается без привязки к стадиям жизненного цикла данных в проекте требований [11]. При этом стандарт [18] определяет процедуры и порядок оценки качества данных, в т.ч. в привязке к стадиям жизненного цикла данных [17]. В связи с этим, проблемы обеспечения безопасности информации СИИ далее рассмотрены применительно к этим стадиям.

### **3. Проблемы обеспечения безопасности информации СИИ на стадиях жизненного цикла данных**

#### **Стадия 1. Замысел**

Стандарт [17] не предусматривает каких-либо действий, связанных с данными на этой стадии. Однако такой подход является не вполне корректным. Представляется, что на этой стадии необходимо определить:

- какие данные потребуются для решения заданной функциональной задачи;
- перечень источников, из которых они могут быть получены;
- какие источники являются доверенными;
- какие организационно-технические решения необходимы для получения требуемых данных.

Без решения приведенных вопросов постановка работ по созданию СИИ может привести к отрицательным результатам и финансовым потерям.

В случае серьезных затруднений решения этих вопросов целесообразно рассмотреть необходимость использования технологии искусственного интеллекта в планируемой к созданию автоматизированной системе.

#### **Стадия 2. Формирование деловых требований**

На этой стадии в отношении собственно используемых данных стандарт предусматривает лишь определение спецификации требований к данным, исходя из целей и потребностей конечных пользователей. В рамках этой работы было бы целесообразно сформировать требования к формату данных, определить структуру данных, область их возможных значений и распределения. Кроме того, оценить возможность получения данных в ходе эксплуатации СИИ, стабильность их основных характеристик (возможен ли их дрейф со временем, от чего это зависит).

На этой же стадии стандартом предусматривается формирование требований по защите персональных данных, что представляется совершенно недостаточным. Если к СИИ предъявляются требования функциональной безопасности [7] как способности к устойчивому и корректному выполнению заявленных

функций, то и требования технической безопасности [7] хотя бы в части обеспечения целостности и доступности должны выполняться независимо от наличия или отсутствия информации ограниченного распространения. Действительно, наличие незаблокированных уязвимостей в системе технической защиты информации позволяет потенциальному нарушителю изменять как входные данные СИИ, так и закон функционирования модели ИНС, включая так называемую «атаку на решающий бит».

### **Стадия 3. Планирование работы с данными**

Стадия планирования работы с данными включает решение о составе наборов данных, которые нужны для ответов на вопросы, сформулированные на стадии формирования деловых требований.

Стандарт [17] достаточно подробно описывает рабочие процессы с данными на этой стадии, однако, с точки зрения обеспечения безопасности СИИ некоторые позиции целесообразно уточнить. Очень важен вопрос о доверии к используемым данным, а также к их источнику. При этом следует учитывать, что владелец массива данных (при использовании заимствованных данных) мог проводить их верификацию с точки зрения актуальных для его целей критериев, которые могут не отражать потребности заказчика разрабатываемой/модернизируемой СИИ. Одновременно уже на этой стадии необходимо предусмотреть меры по защите массива данных в соответствии с [19] в части обеспечения их целостности и доступности на всех этапах жизненного цикла СИИ, а также их конфиденциальности/секретности в случаях, предусмотренных законодательством Российской Федерации (не только персональных данных). Решение об использовании синтетических данных требует серьезной проработки, т.к. в некоторых случаях это приводит к деградации модели ИИ [20]. Целесообразно также на этой стадии ранжировать по степени значимости для конкретных целей заказчика характеристики качества данных, на соответствие которым в дальнейшем будет проводиться оценка качества данных [13]. Это облегчит выполнение работ по сбору/подбору из имеющихся массива данных на последующих стадиях. При определении способов и сроков хранения данных необходимо также определиться и технологией их гарантированного уничтожения в зависимости от степени их значимости и/или конфиденциальности (секретности).

### **Стадия 4. Комплектование наборов данных**

На стадии комплектования наборов данных необходимо убедиться в доверии к источнику [21], особенно при разработке критически важных СИИ. Необходимо с самых первых шагов комплектования наборов данных обеспечивать контроль их целостности, даже в отношении общедоступных массивов данных. Это позволит избежать в дальнейшем целевой атаки отравления данных. По возможности следует избегать использования синтетических данных, в ряде случаев это может приводить к деградации модели ИИ. Собственно сбор/отбор данных должен производиться с учетом определенных на предыдущей стадии приоритетных характеристик качества данных. Как правило, к таким характеристикам в первую очередь относятся релевантность, репрезентативность и актуальность. В зависимости от целевого назначения СИИ приоритетными могут стать и другие характеристики данных, например, своевременность, связанная со скоростью передачи данных, для систем реального времени или согласованность для аналитических систем.

### **Стадия 5. Подготовка наборов данных**

Эта стадия вносит, пожалуй, наибольший вклад в обеспечение качества, а, следовательно, и функциональной безопасности создаваемой/модернизируемой СИИ. Стадия включает в себя преобразование данных физического мира в цифровой формат (кодирование). Очень важное место в подготовке данных занимает признаковое описание объекта. От качества выполнения этой процедуры, полноты учета всех приоритетных характеристик качества объектов в значительной мере зависит и качество машинного обучения в целом. Именно на этом этапе высок риск утраты причинно-следственных связей в предметной области – один из труднопреодолимых недостатков ТИИ. Действительно, все дальнейшие процессы уже связаны лишь с обработкой цифровых данных и оптимизацией параметров статистических критериев. Физическая смысл данных при этом уже не учитывается. В связи с этим представляется, что оценку качества данных экспертным методом в соответствии с [13] целесообразно производить для набора данных, полученных после восстановления из оцифрованных данных, т.к. именно они в дальнейшем используются в математических моделях машинного обучения и функционирования СИИ. Действительно, многие характеристики качества данных, такие, в частности, как точность, полнота, согласованность, доступны для восприятия и оценки экспертами именно в формате первичных данных.

На этом этапе надо окончательно определить приоритетные характеристики качества данных и показатели качества данных для последующей их оценки в зависимости от целевого назначения СИИ. Заметим, что для подготовки данных большого объема зачастую приходится использовать предобученную ИНС, что порождает проблему снижения качества подготовки. Действительно, обучение этой служебной ИНС происходит на предварительно подготовленных вручную наборах данных, содержащих некоторый процент ошибок, дополнительные ошибки вносит собственно ТИИ. В итоге качество подготовленных таким образом данных

может уступать аналогичному при ручной подготовке. Для последующей оценки качества данных необходимо определить приоритетные характеристики качества, подлежащие оцениванию в зависимости от целевого назначения СИИ, и пороговые значения показателей качества. Эта работа должна выполняться экспертами как в предметной области, разрабатываемой СИИ, так и в области технологий искусственного интеллекта. При этом при получении интегральных оценок экспертов целесообразно учитывать степень их согласованности. При необходимости целесообразно исключать крайние мнения, прибегать к средневзвешенным оценкам в зависимости от квалификации экспертов, а в некоторых случаях использовать медианные оценки.

Для тестирования целесообразно использовать набор тестовых данных, подготовленный независимыми экспертами по тем же критериям, что и обучающие данные, в том числе с помощью независимой вспомогательной ИНС. Это позволит избежать некоторой возможной предвзятости экспертов и обеспечить определенную независимость проверки.

#### **Стадия 6. Построение модели ИИ**

До построения модели целесообразно провести оценку качества данных в соответствии с [13] и не только при создании СИИ, связанных с общественной безопасностью, как это предусмотрено [17]. Действительно, использование при построении модели ИИ набора данных, не отвечающих определенным на предыдущих стадиях требованиям (характеристикам и показателям качества), может привести к финансовым и временным потерям.

#### **Стадия 7. Развертывание системы ИИ**

На этой стадии целесообразно провести проверку выполнения требований по технической защите информации (аттестацию или самоаттестацию в зависимости от характера решаемых СИИ задач и степени секретности/конфиденциальности обрабатываемой информации), проверку функциональной безопасности в соответствии с разработанной программой и методикой, включая проверку корректности форматов представления входных данных (соответствия использованным в процессе обучения) от различных датчиков и других устройств съема информации или их корректного преобразования, верификацию и валидацию соответствующего ПО. При этом оценка качества ПО, реализующего функционирование обученной модели ИНС, традиционными методами может вызывать большие затруднения в связи с эффектом необъяснимости ИИ.

#### **Стадия 8. Эксплуатация системы ИИ**

На этой стадии должен проводиться постоянный мониторинг функционирования СИИ, чтобы своевременно выявить момент снижения эффективности ниже заданного уровня и необходимость дообучения. Критерии эффективности функционирования СИИ целесообразно определить уже на стадии построения модели ИИ с учетом выбранных показателей качества данных и пороговых значений характеристик качества СИИ в соответствии с [18], [22]. В ходе мониторинга также необходимо отслеживать изменения в формате и характере входных данных, чтобы своевременно внести коррективы в массив обучающих и тестовых данных при необходимости дообучения.

#### **Стадия 9. Вывод данных из эксплуатации**

Если принято решение о сохранении данных для целей аудита или последующего использования при создании новых СИИ, вместе с массивом данных необходимо сохранять вспомогательную служебную информацию об источнике данных, формате преобразования исходных данных, методах и результатах разметки, дополненных и синтетических данных и др. кроме того было бы полезно сохранять информацию о результатах и условиях использования данных в конкретных СИИ. Массив архивированных данных целесообразно подписать электронной подписью, приняв меры по поддержанию ее в актуальном состоянии на весь период хранения. Методы уничтожения данных должны выбираться в соответствии со степенью их секретности (конфиденциальности) и распространяться на все элементы памяти вычислительной среды СИИ, которые были задействованы в ходе ее эксплуатации.

#### **Стадия 10. Вывод системы ИИ из эксплуатации**

При выводе системы из эксплуатации необходимо принять меры по очистке всех элементов памяти СВТ, которые могли потенциально содержать массивы данных с учетом их степени секретности (конфиденциальности), что представляет собой нетривиальную задачу особенно для высокопроизводительных вычислительных систем.

Заметим также, что функциональная безопасность автоматизированных систем, включая СИИ, особенно киберфизических, решающим образом зависит от безопасности сопутствующей инфраструктуры. Например, даже хорошо защищенный и обученный с подтвержденным качеством беспилотный транспорт нельзя считать безопасным, если не обеспечена на том же уровне безопасность поддерживающей дорожной инфраструктуры.

## Заключение

Таким образом, на основании приведенного в статье анализа положения с нормативной базой, определяющей обеспечение комплексной безопасности систем с искусственным интеллектом, показано, что развитие нормативно-технической базы в части качества данных для машинного обучения должно стать одним из действенных механизмов снижения рисков при применении СИИ. Выделим следующие предложения:

1. При планировании работ с данными в ходе создания СИИ необходимо учитывать проблемы обеспечения функциональной и технической безопасности с ориентацией на существующие нормы и требования в области обеспечения безопасности информации даже в случаях, не предусмотренных действующим законодательством. Это позволит в ряде случаев избежать неэффективного расходования средств и противостоять некоторым специфическим СИИ атакам.

2. На всех стадиях жизненного цикла данных, начиная с замысла, необходимо учитывать требования стандартов в части оценки качества данных и последующей оценки качества СИИ в целом.

## Список литературы

1. Национальный проект «Экономика данных», – Режим доступа: <http://government.ru/rugovclassifier/909/events/>
2. Международная конференция «Путешествие в мир искусственного интеллекта» (Москва, 11-13 декабря 2024). URL: <http://kremlin.ru/events/president/news/75830>
3. Путин поручил широко использовать технологии ИИ по всей России <https://clck.ru/3SpWDZ>
4. Сопещение по вопросам развития автономных систем <https://clck.ru/3SpWF6>
5. Зацаринный А.А., Гаврилов В.Е. Некоторые подходы к развитию технического регулирования в области информационной безопасности систем с искусственным интеллектом //ИТ-Стандарт. 2025. № 3 (44). С. 11-18.
6. Гаврилов В.Е., Зацаринный А.А. Проблемы и угрозы внедрения некоторых новых цифровых технологий. // Системы и средства информатики. 2022. Т. 32. № 3. С. 15-25.
7. Гаврилов В.Е., Зацаринный А.А. Особенности обеспечения функциональной безопасности автоматизированных систем с применением технологии искусственной интеллекта. //Системы и средства информатики. 2024. Т. 34. № 3. С. 23-34.
8. Намиот Д.Е. Искусственный Интеллект в Кибербезопасности. Хроника. Выпуск 4. //International Journal of Open Information Technologies. ISSN: 2307-8162 vol. 14, no. 1, 2026
9. Зацаринный А.А., Иванов К.В. Ключевые вопросы внедрения технологий искусственного интеллекта в контексте обеспечения военной безопасности государства //Известия РАРАН. 2025. №3 (138). С. 13-19.
10. А.А. Зацаринный, В.Е. Гаврилов Проблемы нормативно-правового и технического регулирования обеспечения информационной безопасности при создании автоматизированных систем военного назначения. Материалы 6-й Международной межведомственной научно-практической конференции научного отделения № 10 Российской академии ракетных и артиллерийских наук, Москва, 18 марта 2021 года, Т.2 с.69-75.
11. Предложения в проект требований по обеспечению информационной безопасности в системах, реализующих ИИ. Академия криптографии РФ. 2025. <https://cryptoacademy.gov.ru/upload/medialibrary/b41/w7iu2yb vazyn6v6j0o163jm6wul60fk.pdf>
12. ГОСТ Р 71484.1-2024 (ИСО/МЭК 5259-1:2024) Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 1. Обзор, термины и примеры.
13. ГОСТ Р 71484.2-2024 (ИСО/МЭК 5259-2:2023) Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 2. Показатели качества данных.
14. ГОСТ Р 71484.3-2024 (ИСО/МЭК 5259-3:2024) Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 3. Требования и рекомендации по управлению качеством данных.
15. ГОСТ Р 71484.4-2024 (ИСО/МЭК 5259-4:2023) Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 4. Структура процесса управления качеством данных.
16. Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 5. Структура стратегического управления качеством данных (ISO/IEC 5259-5, MOD) (1.11.164-1.348.24) (проект, окончательная редакция).
17. ГОСТ Р 70889-2023 (ИСО/МЭК 8183:2023) Информационные технологии. Искусственный интеллект. Структура жизненного цикла данных.
18. ГОСТ Р 59898 – 2021 искусственный интеллект. Оценка качества систем искусственного интеллекта. Общие положения.
19. ПНСТ 848-2023 Искусственный Интеллект. Большие данные. Обзор и требования по обеспечению

сохранности данных.

20. I. Shumailov, Z. Shumaylov, Y. Zhao, Y. Gal, N. Papernot, R. Anderson The Curse of Recursion: Training on Generated Data Makes Models Forget, <https://doi.org/10.48550/arXiv.2305.17493>, [Submitted on 27 May 2023]

21. ПНСТ 847-2023 Искусственный интеллект. Большие данные. Функциональные требования к происхождению данных

22. Оценка качества систем искусственного интеллекта. Методы оценки (1.11.164-1.161.22), проект

## References

1. National Project "Data Economy", – Mode of access: <http://government.ru/rugovclassifier/909/events/>
2. International Conference "Journey to the World of Artificial Intelligence" (Moscow, December 11-13, 2024). URL: <http://kremlin.ru/events/president/news/75830>
3. Putin instructed to widely use AI technologies throughout Russia <https://clck.ru/3SpWDZ>
4. Meeting on the development of autonomous systems <https://clck.ru/3SpWF6>
5. Zatsarinny A.A., Gavrilov V.E. Some approaches to the development of technical regulation in the field of information security of systems with artificial intelligence. 2025. № 3 (44). Pp. 11-18.
6. Gavrilov V.E., Zatsarinny A.A. Problems and threats of the introduction of some new digital technologies. Systems and means of informatics. 2022. T. 32. № 3. Pp. 15-25.
7. Gavrilov V.E., Zatsarinny A.A. Features of ensuring the functional safety of automated systems using artificial intelligence technology. Systems and means of informatics. 2024. T. 34. № 3. Pp. 23-34.
8. Namiot D.E. Artificial Intelligence in Cybersecurity. Chronicle. Issue 4. //International Journal of Open Information Technologies. ISSN: 2307-8162 vol. 14, no. 1, 2026
9. Zatsarinny A.A., Ivanov K.V. Key issues of the introduction of artificial intelligence technologies in the context of ensuring the military security of the state // Izvestiya RARAN. 2025. №3 (138). Pp. 13-19.
10. A.A. Zatsarinny, V.E. Gavrilov Problems of Regulatory and Technical Regulation of Information Security in the Creation of Automated Military Systems. Proceedings of the 6th International Interdepartmental Scientific and Practical Conference of the Scientific Department No 10 of the Russian Academy of Rocket and Artillery Sciences, Moscow, March 18, 2021, Vol. 2, pp. 69-75.
11. Proposals for the draft requirements for ensuring information security in systems implementing AI. Academy of Cryptography of the Russian Federation. 2025. <https://cryptoacademy.gov.ru/upload/medialibrary/b41/w7iu2yb vazyn6v6jo0o163jm6wul60fk.pdf>
12. GOST R 71484.1-2024 (ISO/IEC 5259-1:2024) Artificial intelligence. Data quality for analytics and machine learning. Part 1. Overview, terms and examples.
13. GOST R 71484.2-2024 (ISO/IEC 5259-2:2023) Artificial intelligence. Data quality for analytics and machine learning. Part 2. Data quality indicators.
14. GOST R 71484.3-2024 (ISO/IEC 5259-3:2024) Artificial intelligence. Data quality for analytics and machine learning. Part 3. Data quality management requirements and best practices.
15. GOST R 71484.4-2024 (ISO/IEC 5259-4:2023) Artificial intelligence. Data quality for analytics and machine learning. Part 4. Structure of the data quality management process.
16. Artificial intelligence. Data quality for analytics and machine learning. Part 5. Strategic Data Quality Management Framework (ISO/IEC 5259-5, MOD) (1.11.164-1.348.24) (draft, final revision).
17. GOST R 70889-2023 (ISO/IEC 8183:2023) Information Technology. Artificial intelligence. Data lifecycle structure.
18. GOST R 59898 – 2021 Artificial Intelligence. Assessment of the quality of artificial intelligence systems. General provisions.
19. PNST 848-2023 Artificial Intelligence. Big data. Overview and data security requirements.
20. I. Shumailov, Z. Shumaylov, Y. Zhao, Y. Gal, N. Papernot, R. Anderson The Curse of Recursion: Training on Generated Data Makes Models Forget, <https://doi.org/10.48550/arXiv.2305.17493>, [Submitted on 27 May 2023]
21. PNST 847-2023 Artificial Intelligence. Big data. Functional requirements for the origin of data
22. Assessment of the quality of artificial intelligence systems. Evaluation methods (1.11.164-1.161.22), draft