

БЕЗОПАСНОСТЬ ИИ: ТРЕБУЕМОЕ И НЕОБХОДИМОЕ

Сергей Канивец
Руководитель группы экспертного аудита



ДиалогНаука
СИСТЕМНЫЙ ИНТЕГРАТОР
ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ

Национальная стратегия развития искусственного интеллекта на период до 2030 года
(в редакции Указа Президента Российской Федерации от 15.02.2024 № 124)

Федеральный закон

Постановления Правительства

НПА уполномоченных органов

Национальные стандарты и проекты (>160)

Март 2026:

Проект ФЗ – Об основах государственного регулирования сфер применения технологий искусственного интеллекта в Российской Федерации

- ❏ Риск-ориентированный подход к регулированию применения ИИ
- ❏ Технологическая независимость
- ❏ Обеспечивается поддержка внедрения суверенных и национальных больших фундаментальных моделей ИИ
- ❏ Реестр доверенных моделей ИИ. Требования к безопасности. Требования к качеству

Март 2026:

Проект ФЗ – Об основах государственного регулирования сфер применения технологий искусственного интеллекта в Российской Федерации

- 🕒 Права граждан
- 🕒 Маркировка контента
- 🕒 Защита интеллектуальной собственности
- 🕒 Перечень ЦОД (суперкомпьютеров)

Июнь 2026:

Проект ФЗ – О поддержке развития технологий искусственного интеллекта в Российской Федерации

- ❏ проект не про технологии ИИ в целом, а направлен на аспекты разработки, внедрения и применения больших фундаментальных моделей (БФМ: не менее 1 миллиарда параметров)
- ❏ убрали «доверенные модели ИИ» и соответственно требования по безопасности к ним, их реестр и применение в ГИС/КИИ
- ❏ убрали ЦОД и меры их поддержки (поддержка разработчиков)
- ❏ убрали положения, связанные с мониторингом и реагированием на риски и угрозы применения ИИ

Июнь 2026:

Проект ФЗ – О поддержке развития технологий искусственного интеллекта в Российской Федерации

- ❏ пересмотрели подходы к маркировке: требования обязательной маркировки нет (как и требований к порядку и содержанию)
- ❏ раскрыли понятия «суверенных» и «национальных» БФМ:
 - они должны пройти подтверждение соответствия законодательству РФ и традиционным духовно-нравственным ценностям;
 - при разработке суверенной модели должна обеспечиваться полная воспроизводимость цикла разработки (в т.ч. обучение);
 - для разработки национальной модели могут использоваться компоненты, созданные за рубежом, в т.ч. open source

Июнь 2026:

Проект ФЗ – О поддержке развития технологий искусственного интеллекта в Российской Федерации

Правительство Российской Федерации:

- ❶ Устанавливает порядок учета и присвоения статуса суверенных и национальных БФМ;
- ❷ Устанавливает порядок подтверждения БФМ соответствию законодательству РФ и традиционным духовно-нравственным ценностям
- ❸ Устанавливает порядок предоставления доступа к составам данных в ГИС в целях обучения БФМ

Июнь 2026:

Проект ФЗ – О поддержке развития технологий искусственного интеллекта в Российской Федерации

Правительство Российской Федерации:

- 🕒 Устанавливает порядок учета и присвоения статуса суверенных и национальных БФМ;
- 🕒 Устанавливает порядок подтверждения БФМ соответствию законодательству РФ и традиционным духовно-нравственным ценностям
- 🕒 Устанавливает порядок предоставления доступа к составам данных в ГИС в целях обучения БФМ

Федеральный закон от 26.07.2026 № 243-ФЗ «О поддержке развития технологий искусственного интеллекта в Российской Федерации»

Вступил в силу с 01.09.26, часть мер с 01.03.2027: обязанности Правительства, требования к национальной и суверенной БФМ, обязанности разработчиков, маркировка материала

Август 2026:

Перечень поручений по итогам заседания Совета по науке и образованию

Правительству Российской Федерации:

- обеспечить централизованный сбор, обработку и предоставление научно-технической информации для обучения суверенных и/или национальных БФМ ИИ;
- выработать механизм предоставления научным работникам доступа к ПО, функционирующему с применением российских технологий ИИ, определив источники финансового обеспечения такого доступа

до 01.03.2027

Октябрь 2026:

Проект ФЗ – О безопасном использовании искусственного интеллекта

- Защита прав и свобод (запрещены дипфейки)
- Использование ИИ в социальной сфере, в Госах. Использование ИИ ФЛ с его согласия
- Доверенные системы ИИ – используемые в социальной сфере и в Госах
- Производители систем ИИ обязаны технически реализовать возможность внешней оценки корректности, надежности и законности результатов функционирования систем ИИ

Октябрь 2026:

Проект ФЗ – О безопасном использовании искусственного интеллекта

- ❶ Запрещенные системы ИИ: использование в целях нарушения законодательства РФ, наличие средств обхода закона, иное, установленное Правительством
- ❶ Реестр доверенных систем ИИ: компоненты из реестра российских программ, проведенная Минцифрой экспертиза
- ❶ Предусматривается возможность принятия юр.значимых решений системой ИИ и механизмы оспаривания
- ❶ Фиксация всех действий системы и пользователей (для доверенной ИИ)
- ❶ Вступление в силу – 01.2027, за исключением маркировки и согласия ФЛ (01.2028)

Октябрь 2026:

Проект ФЗ – О безопасном использовании искусственного интеллекта

Требования безопасности (доверенных ИИ)

Производитель:

- обеспечивает отказоустойчивость, противодействие атакам, НСД + ФСБ и ФСТЭК России;
- определение угроз, оценка рисков;
- документирование ЖЦ;
- функция рубильника.

Оператор:

- обеспечивает отказоустойчивость, противодействие атакам, НСД + ФСБ и ФСТЭК России;
- допускает к использованию только обученных лиц;
- информирование Минцифры и производителя о сбоях, атаках и НСД

Октябрь 2026:

Проект ФЗ – О безопасном использовании искусственного интеллекта

Минцифры:

- Ведет реестр доверенных систем ИИ (порядок определяет Правительство)
- Ведет перечень запрещенных систем ИИ (порядок определяет Правительство)
- Устанавливает правила по эксплуатации ОИИ и ЦОД (РФ), используемых для обеспечения функционирования систем ИИ (порядок определяет Правительство)
- Минцифры (с ФСБ) устанавливает требования к качеству и безопасности обучающим наборам данных (для доверенных ИИ)

Приказ № 117

Правки в Приказ № 21

Мероприятия и меры по ЗИ, содержащейся в ИС



Приказ № 117 (текущая редакция)

П.60. должна быть обеспечена возможность исключения:

- НСД к информации или воздействия на ИС;
- несанкционированного распространения и модификации информации;
- использования ИС не по их назначению за счет воздействия на наборы данных, применяемые модели ИИ и их параметры, процессы и сервисы по обработке данных и поиску решений

Приказ № 117 (проект № 170500)

Приказ № 117 (текущая редакция)

П.60. должна быть обеспечена возможность исключения:

- НСД к информации или воздействия на ИС;
- несанкционированного распространения и модификации информации;
- использования ИС не по их назначению за счет воздействия на ~~наборы данных~~, применяемые модели ИИ и их параметры, процессы и сервисы по обработке данных и поиску решений

Приказ № 117 (проект № 170500)

П.60. должна быть обеспечена возможность исключения НСД к информации или воздействия на ИС, несанкционированного распространения и модификации информации, а также использования ИС не по их назначению за счет воздействия на применяемые модели ИИ и их параметры, **агенты искусственного интеллекта**

Приказ № 117 (текущая редакция)

П.60. должна быть обеспечена возможность исключения:

- НСД к информации или воздействия на ИС;
- несанкционированного распространения и модификации информации;
- использования ИС не по их назначению за счет воздействия на ~~наборы данных~~, применяемые модели ИИ и их параметры, процессы и сервисы по обработке данных и поиску решений

Приказ № 117 (проект № 170500)

П.60. должна быть обеспечена возможность исключения НСД к информации или воздействия на ИС, несанкционированного распространения и модификации информации, а также использования ИС не по их назначению за счет воздействия на применяемые модели ИИ и их параметры, **агенты искусственного интеллекта**

Требования безопасности:

- выделение моделей ИИ в отдельный сегмент;
- доступ пользователей с усиленной MFA;
- доступ пользователей с минимальными правами;
- регистрация и анализ действий

Приказ № 117 (текущая редакция)

П.61. При взаимодействии пользователей в целях выполнения ими своих обязанностей (функций) с сервисами на основе ИИ посредством направления запроса и получения ответа должны быть...

В состав ИС должны включаться доверенные технологии ИИ или их компоненты

Приказ № 117 (проект № 170500)

Приказ № 117 (текущая редакция)

~~П.61. При взаимодействии пользователей в целях выполнения ими своих обязанностей (функций) с сервисами на основе ИИ посредством направления запроса и получения ответа должны быть...~~

~~В состав ИС должны включаться доверенные технологии ИИ или их компоненты~~

Приказ № 117 (проект № 170500)

П.61. При применении агентов ИИ (в т.ч. автономных), оператором ИС должны быть приняты меры ЗИ, направленные на:

- предотвращение несанкционированного воздействия на ресурсы ИС;
- контроль и управление правами доступа агентов

Приказ № 117

- ГИС/МИС
- ИС гос. органов
- ИС ФГУП/МУП
- ИС гос. учреждений
(образование, медицина,
ПФ, культура...)

Приказ № 117

- ГИС/МИС
- ИС гос. органов
- ИС ФГУП/МУП
- ИС гос. учреждений
(образование, медицина,
ПФ, культура...)

Мероприятия и меры по ЗИ, содержащейся в ИС

- все те же
- ЗОКИИ (приказ 239)
- ПДн (приказ 21)
- АСУ ТП (КВО) (приказ 31)
- ИС УП (ОПК) (приказ 31)

Приказ № 117

- ГИС/МИС
- ИС гос. органов
- ИС ФГУП/МУП
- ИС гос. учреждений
(образование, медицина,
ПФ, культура...)

Мероприятия и меры по ЗИ, содержащейся в ИС

- все те же
- ЗОКИИ (приказ 239)
- ПДн (приказ 21)
- АСУ ТП (КВО) (приказ 31)
- ИС УП (ОПК) (приказ 31)

Замена приказа № 21 (проект 169583)

Методика оценки уровня зрелости деятельности в области технической защиты информации в информационных системах и обеспечения безопасности значимых объектов критической информационной инфраструктуры Российской Федерации, утв. ФСТЭК России 07.08.2026

Мероприятия и меры по ЗИ, содержащейся в ИС

3.17. Обеспечение ЗИ при использовании ИИ

- Требования к реализации (25 процессов + МД)
- Требования к документации (не предъявляются)
- Требования к усилению (не предъявляются)

Мероприятия и меры по ЗИ, содержащейся в ИС

3.18. Обеспечение ЗИ при использовании ИИ (разработка)

- Перечень объектов защиты (ОИИ, ПО разработки ИИ, API, агентов, обучающие данные, выходная модель и ее параметры)
- Выделение в изолированный сегмент разработки, использование безопасных форматов данных, целостность ПО
- Требования к набору обучающих данных (доверенные источники, АВЗ, обособленное хранение, целостность)

Мероприятия и меры по ЗИ, содержащейся в ИС

3.18. Обеспечение ЗИ при использовании ИИ (эксплуатация)

- Перечень объектов защиты (ОИИ, ПО реализации ИИ, ПО агентов, API, систем фильтрации запросов, обученная модель и ее расширения)
- При использовании ИИ в составе ИС, должны реализовываться:
 - идентификации и аутентификации пользователей;
 - управлению доступом пользователей;
 - фильтрации входных и выходных данных;
 - обеспечению изоляции системы ИИ;
 - защите данных системы ИИ;
 - защите от ВПО;
 - выявлению уязвимостей;
 - контролю функциональности системы ИИ
- Фильтрация входных/выходных запросов, квотирование, регистрация событий, КЦ параметров
- Требования к усилению: изоляция (L1) ИИ разработки; использование СКЗИ при хранении обучающих данных и обеспечении целостности модели ИИ; требования к выходной модели и параметрам; тестирование промт-атак)

Кодекс этики в сфере разработки и применения ИИ на финансовом рынке

Методические рекомендации Банка России по обеспечению
информационной безопасности при разработке и применении
искусственного интеллекта на финансовом рынке
(З-МР от 16.06.2026)



МР по обеспечению ИБ при разработке и применении ИИ на финансовом рынке

- Риски ИИ, связанные с ИБ и ОН. В критических процессах валидация результатов ИИ человеком
- Модель угроз (8), способы реализации (8) и меры защиты (21) (соблюдать пропорциональность принимаемых мер защиты выявленным рискам и масштабам последствий от их реализации)
- Тактики (6) и техники (20) реализации угроз
- Основные положения политики обеспечения ИБ при разработке и применении технологий ИИ
- Обеспечение ИБ при взаимодействии с поставщиками услуг (сервис, open-source)

ГОСТ Р (проект)



«Защита информации. Разработка безопасного программного обеспечения, реализующего технологии искусственного интеллекта. Общие требования»

- ❏ Проект направлен на процесс разработки систем ИИ и не устанавливает требования к системам ИИ, к ЖЦ, менеджменту качества, классификации и применения систем ИИ
- ❏ Соотнесение процессов ГОСТ Р 71539-2024 и ГОСТ Р
- ❏ 25 процессов РБПО (типовые и модифицированные) + 2 специфических:
 - Управление наборами данных;
 - Безопасное обучение моделей ИИ
- ❏ Состав данных описания модели ИИ (в контексте ППК)

ГОСТ Р (проект)

*ГОСТ Р 71476-2024 – Искусственный интеллект. Концепции и терминология
искусственного интеллекта

ГОСТ Р 71539-2024 (ИСО-МЭК 5338-2023) – Искусственный интеллект.
Процессы жизненного цикла систем искусственного интеллекта

ГОСТ Р 59277-2020 – Системы искусственного интеллекта. Классификация
систем искусственного интеллекта

ГОСТ Р (проект)

- ГОСТ Р 59276-2020 – Системы искусственного интеллекта. Способы обеспечения доверия. Общие положения
- ГОСТ Р 59898-2021 – Оценка качества систем искусственного интеллекта. Общие положения
- ГОСТ Р 71484.x-2024 (ИСО-МЭК 5259-x-2024) – Искусственный интеллект. Качество данных для аналитики и машинного обучения
- ГОСТ Р 70889-2023 (ИСО/МЭК 8183:2023) – Информационные технологии. Искусственный интеллект. Структура жизненного цикла данных
- ГОСТ Р ИСО/МЭК 42001-2024 - Искусственный интеллект. Система менеджмента
- ПНСТ 836-2023 (ISO-IEC DTR 5469) – Искусственный интеллект. Функциональная безопасность и системы искусственного интеллекта
- ПНСТ 837-2023 – Искусственный интеллект. Управляемость автоматизированных систем искусственного интеллекта
- ПНСТ 1046-2026 – Искусственный интеллект в критической информационной инфраструктуре. Общие положения

ГОСТ Р 56939-2024

Сейчас

**ГОСТ Р 58412-2019
(Угрозы БИ)**

**ГОСТ Р 71206-2024
(Компилятор C/C++)**

**ГОСТ Р 71207-2024
(Стат. анализ)**

ГОСТ Р 56939-2024

Сейчас

ГОСТ Р 58412-2019
(Угрозы БИ)

ГОСТ Р 71206-2024
(Компилятор C/C++)

ГОСТ Р 71207-2024
(Стат. анализ)

В разработке

Термины и определения

Методика оценки уровня РБПО

Динамический анализ

Композиционный анализ

Угрозы БИ при ИИ

Национальные стандарты (планы)

Угрозы БИ при ИИ

В планах (~2030)

Инструменты автоматизации сборки и
конфигурирования программных проектов

Правила оформления безопасного кода
для языков программирования

Национальные стандарты (планы)

Угрозы БИ при ИИ

В планах (~2030)

Инструменты автоматизации сборки и
конфигурирования программных проектов

Правила оформления безопасного кода
для языков программирования

Среда разработки РБПО

Руководство по реализации
мер по РБПО

- 🕒 NIST AI Risk Management Framework и NIST AI RMF Playbook (NIST AI 100-1), NIST SP 800-218A
- 🕒 Google Secure AI Framework 2.0
- 🕒 MITRE ATLAS и Framework SAFE-AI
- 🕒 OWASP GenAI Security Project: AI Maturity Assessment, AI Security and Privacy Guide, TOP-10 LLM...
- 🕒 China's AI Safety Governance Framework, EU AI Act, US Executive Order on AI
- 🕒 Guidelines for Secure Development and Deployment of AI Systems (Kaspersky) и Kaspersky AIST
- 🕒 AI Security Agentic Framework Essentials (AI-SAFE) 2.0 (Yandex Cloud)

Спасибо за внимание

Сергей Канивец, руководитель группы экспертного аудита

m: +7 (904) 989-43-71

w: <https://dialognauka.ru/>

e: s.kanivets@dialognauka.ru

