

АНО «НИИ СИСТЕМНОГО СИНТЕЗА» · ОГРН 1267700083617 · ПРИБОРНЫЙ ПРОТОКОЛ ИБ 3.0 · РЕДАКЦИЯ V1.0 ОТ 25.10.2026

Паспорт приборной верификации и заводских испытаний ИБ 3.0: MITR Sentinel · Полигон INDRA · ISS Ledger 117

Назначение документа: настоящий технический паспорт фиксирует инструментально измеренные показатели промышленного стека детерминированной безопасности автономных ИИ-агентов (ИБ 3.0), разработанного АНО «НИИ Системного синтеза» (научный руководитель — **Фищук Александр Алексеевич**), по состоянию на **25 октября 2026 года**. Все приведённые метрики получены автоматическими верификационными стендами без ручной правки и доступны для перепроверки аудиторами заказчика и инвестора.

Обязательная регуляторная оговорка (Принцип № 2 «Claim ≤ Reality»): настоящий документ является официальным **протоколом заводских приёмо-сдаточных испытаний разработчика**, структурированным в соответствии с требованиями **ГОСТ Р 56939-2024** и **Приказа ФСТЭК России от 11.04.2025 № 117** (вступил в силу 01.03.2026), и не подменяет государственный сертификат соответствия ФСТЭК России.

1. Архитектура стека ИБ 3.0 и теорема детерминированного перехвата

Традиционные средства защиты языковых моделей (Guardrails второго поколения) используют вероятностный классификатор для фильтрации текста, что оставляет ненулевую вероятность пропуска состоятельной атаки ($\epsilon > 0$) и не создаёт математически воспроизводимого доказательства соблюдения политики. Стек **ИБ 3.0** переносит точку контроля на границу физического вызова инструмента (Function Calling / Model Context Protocol / REST / SQL / ACU TP) и опирается на четыре замкнутых контура:

- 1. Формализация контракта в разрешимой логике первого порядка ($QF_UF\Delta$):** каждое действие агента a с вектором параметров p и контекстом среды s отображается в систему ограничений кванторно-свободной линейной целочисленной арифметики с неинтерпретируемыми функциями. Вызов разрешается тогда и только тогда, когда решатель **Z3** доказывает совместность конъюнкции условий безопасности.
- 2. Иммунная петля самоисправления по ядру невыполнимости (*Unsat Core Injection*, заявка Роспатента № 2026128816 от 22.09.2026):** при обнаружении нарушения (UNSAT) шлюз не просто обрывает сессию, а извлекает минимальное невыполнимое подмножество аксиом C_{unsat} и возвращает его агенту в связке `retry_of`, позволяя агенту скорректировать параметры вызова в пределах `MITR_GATE_ROUNDS_MAX = 3`.
- 3. Криптографическая квитанция вместо доверия на слово (*iss-ledger-117*):** каждый вердикт (ALLOW или DENY) канонизируется по **RFC 8785 (JCS)**, хэшируется по национальному стандарту **ГОСТ Р 34.11-2012 («Стрибог-256»)**, связывается с предыдущей записью в линейную хэш-цепь и периодически якорится через деревья Меркла в блокчейне Bitcoin по протоколу **OpenTimestamps**.
- 4. Автономный полигон атаки и автосинтеза правил (*INDRA*):** в изолированной песочнице без доступа к внешним сетям генерирует мутации эксплойтов и при обнаружении пробития автоматически синтезирует новое блокирующее правило $QF_UF\Delta$ быстрее чем за **0,05 секунды**.

2. Сводная ведомость выполнения 6 рубежей промышленного шлюза MITR Sentinel (K1–K6)

К 24–25 октября 2026 года промышленный шлюз **MITR Sentinel** закрыл все **6 из 6** контрольных рубежей производственной программы (**100% готовности**).

Рубеж / Индекс	Наименование контрольного контура	Нормативный целевой порог (Обещано)	Фактический приборный замер на 25.10.2026 (Факт)	Верифицирующий инструмент / Статус
----------------	-----------------------------------	-------------------------------------	--	------------------------------------

K1 / P1	Полнота блокировки разрушительных векторов и отсутствие ложных срабатываний	<code>fn_rate = 0,0</code> <code>fp_rate ≤ 0,02</code>	Выборка: 157 эталонных кейсов (включая 52 вектора в живом регрессионном прогоне): <code>fn_rate = 0,0 (0 пропусков)</code> <code>fp_rate = 0,0 (0 ложных блокировок)</code>	<code>verify_p1_corpus.py</code> PASS (100%)
K2 / P2	Иммунная петля самоисправления агента по ядру невыполнимости (<code>retry_of</code>)	Сходимость $\geq 80,0\%$ за ≤ 3 раунда при 0 утечек наружу	Сходимость: 85,0% (17 из 20 сложных сценариев исправлены агентом со 2-й попытки), число утечек опасного действия наружу — 0	<code>verify_p2_immune_loop.py</code> PASS (100%)
K3 / P3	Независимая криптографическая воспроизводимость журнала квитанций	100% сходимость хэш-цепи и OTS-якоря при внешнем реплее	Полный повтор: 5 053 из 5 053 квитанций в эпохе № 32 OpenTimestamps Bitcoin (всего в непрерывной цепи 5 088 квитанций, <code>mismatches = 0</code>)	<code>verify_chain_standalone.py --replay</code> PASS (100%)
K4 / P4	Микросекундная и миллисекундная задержка решателя Z3 в рантайме	Медиана решателя $< 3,0$ мс; полный HTTP-цикл $< 10,0$ мс	Горячий такт решателя: медиана 1,08 мс (<code>p95 = 1,83 мс</code>); полный сквозной HTTP-цикл шлюза: медиана 5,22 мс (<code>p95 = 5,58 мс</code>)	<code>benchmark_p4_latency.py</code> PASS (100%)
K5 / K6	Матрица промышленных контрактов в режиме принудительного исполнения (<code>enforce</code>)	Перевод доменных контрактов из <code>shadow</code> в боевой <code>enforce</code>	20 контрактов в боевом режиме <code>enforce</code> (3 системных базовых + 17 отраслевых доменных, включая 4 контракта в логике <code>QF_UFLIA</code>)	<code>registry/contracts.json</code> PASS (20 в enforce)
K6-Финтех	Прототип правил детерминированного комплаенса маркетплейсов и эскроу (Режим Б)	Стенд 5 правил I1-I5 к рубежу 25.11.2026 (№ 289-ФЗ и ст. 860.1 ГК РФ)	Стенд 5 правил (I1-I5): 1 позитивная + 6 из 6 негативных проб (маркировка, сертификат ЕАЭС, согласие на скидку, эскроу-сделка, токены, ст. 860.1 ГК РФ)	<code>verify_marketplaces_25_11.py --five</code> PASS (прототип к 25.11)

3. Протокол заводских испытаний автономного полигона INDRA (Gates A–E: 12 / 12 PASS)

Автономный полигон **INDRA** предназначен для непрерывного стресс-тестирования агентских шлюзов (Continuous Automated Red-Teaming) и автоматического синтеза контрмер в логике `QF_UFLIA`. Испытания проведены по пяти контрольным вратам (**Gates A–E**, суммарно **12 из 12** проверок со статусом **PASS**).

Врата испытаний	Проверяемое свойство полигона INDRA	Приборный результат на выборке N = 28 и стресс-пробах	Статус приёмки
Gate A (Точность детекции)	Матрица ошибок на сбалансированной выборке N = 28 (14 боевых эксплойтов + 14 легитимных рабочих вызовов)	TP = 14, FP = 0, TN = 14, FN = 0. Чувствительность (Sensitivity / Recall) = 100,0% . Специфичность (Specificity) = 100,0% .	PASS (3/3)
Gate B (Охват таксономии)	Покрытие 8 классов угроз агентских систем (OWASP Agentic Top-10 / MITRE ATLAS)	7 классов в активном контуре AUTO_SCAN (инъекции промптов, обход лимитов, эксфильтрация, подмена ролей, SQL/OS-инъекции, нарушение инвариантов состояния, обход цепочки) + 1 класс (dos_resource) в изолированном режиме LAB_ONLY.	PASS (2/2)
Gate C (Безопасность среды)	Изоляция полигона (Zero-Damage Sandbox) при исполнении разрушительных пэйлоадов	Проверено 26 из 26 инвариантов изоляции: 0 сетевых обращений к внешним хостам, 0 модификаций файлов вне временного контейнера, 100% детерминированная очистка состояния.	PASS (3/3)
Gate D (Автосинтез правил)	Скорость и корректность автоматического синтеза патча QF_UFLIA после обнаружения нового вектора	Время синтеза и валидации нового SMT-правила: < 0,05 с (менее 50 мс) против 4–8 часов ручной работы аналитика SOC (ускорение в 10 000 раз).	PASS (2/2)
Gate E (Криптопротокол)	Подписание отчёта испытаний и совместимость с iss-ledger-117 (ГОСТ Р 34.11-2012)	Каждый прогон формирует неизменяемый JSON-манифест с хэшем Стрибог-256 и полной трассой входов/выходов решателя Z3.	PASS (2/2)

4. Реестр открытых пакетов PyPI и честная двойная метрика задержки

Для обеспечения прозрачности перед инженерным сообществом и заказчиками три базовых компонента клиентского стека ИБ 3.0 опубликованы в официальном каталоге **Python Package Index (PyPI)** под свободной лицензией **Apache-2.0**:

Пакет PyPI и версия	Функциональное назначение	Приборно измеренные характеристики	Команда установки
iss-ledger-117 версия 1.1.1	Криптографический журнал действий ИИ-агентов с канонизацией RFC 8785 (JCS)	100% детерминированная проверка целостности цепи; поддержка офлайн-	pip install iss-ledger-117==1.1.1
	34.11-2012 («Стрибог-256»)	аудита по Приказу ФСТЭК № 117	
mcp-sentinel версия 1.1.2	Детерминированный перехватчик вызовов протокола Model Context Protocol (MCP) с проверкой политик до исполнения	Двойная метрика: ~0,016 мс чистая проверка правил в оперативной памяти процесса и ~22,5 мс полный цикл с синхронной записью ГОСТ-квитанции на диск	pip install mcp-sentinel==1.1.2

iss-agent-kit версия 0.1.1	Единый SDK (GuardedMCPServer), соединяющий mcp-sentinel и iss-ledger-117 в один защищённый контур	Гороскопный пакет автотестов: 19 из 19 PASS; полная совместимость с Python 3.10–3.12 на Linux и Windows	pip install iss-agent- kit==0.1.1
-------------------------------	---	---	--------------------------------------

5. Нормативная привязка к стандартам РФ (ФСТЭК России, ГОСТ Р, 152-ФЗ, 187-ФЗ, 289-ФЗ)

- **Приказ ФСТЭК России от 11.04.2025 № 117** (вступил в силу с 01.03.2026) — пп. 15, 22, 65: регистрация событий безопасности, контроль целостности и предотвращение несанкционированных воздействий в государственных и корпоративных информационных системах.
- **ГОСТ Р 56939-2024** («Защита информации. Разработка безопасного программного обеспечения. Общие требования») — процессы статического и формального анализа, верификации архитектурных инвариантов и тестирования на проникновение (полигон INDRA).
- **ГОСТ Р 34.11-2012** («Информационная технология. Криптографическая защита информации. Функция хэширования „Стрибог“») — национальный стандарт хэширования (256 бит), используемый в iss-ledger-117 для формирования неизменяемой хэш-цепи квитанций.
- **ГОСТ Р 57580.1-2017** («Безопасность финансовых (банковских) операций. Защита информации финансовых организаций») — контур операционного контроля транзакций при применении ИИ-агентов в финансовом секторе.
- **Федеральный закон от 31.07.2025 № 289-ФЗ** «Об отдельных вопросах регулирования платформенной экономики в Российской Федерации» (вступил в силу с 01.10.2026), **ПП РФ от 28.01.2026 № 54**, **ПП РФ от 30.04.2026 № 504** и **ст. 860.1–860.10 ГК РФ** (договор номинального счёта) — нормативная база прототипа детерминированных правил I1–I5 (Режим Б к рубежу 25.11.2026).

6. Патентная защита и инструкция по воспроизводимой проверке для аудитора

Фундаментальные технические решения стека ИБ 3.0 защищены двумя заявками на выдачу патента РФ на изобретение в Федеральном институте промышленной собственности (ФИПС Роспатента):

1. **Заявка РФ № 2026116990 от 01.06.2026** — способ и система гибридного консенсуса и динамической маршрутизации в мультиагентных вычислительных средах.
2. **Вторая патентная заявка РФ № 2026128816 от 22.09.2026** — способ детерминированного перехвата и верификации действий автономных ИИ-агентов с замкнутым циклом обратной связи по минимальному ядру невыполнимости (*Unsat Core Injection*).

Пошаговый сценарий независимой приборной верификации (занимает 60 секунд):

```
# Шаг 1. Установка открытых пакетов из официального реестра PyPI
pip install iss-ledger-117==1.1.1 mcp-sentinel==1.1.2 iss-agent-kit==0.1.1

# Шаг 2. Проверка детерминированной блокировки опасного действия и выпуска квитанции ГОСТ Р
34.11-2012
python -c "from iss_ledger_117 import Ledger117; l = Ledger117(); r =
l.append({'tool':'shell','cmd':'rm -rf /','verdict':'DENY'}); print('GOST receipt hash:',
r.hash_hex); assert l.verify()"
```

Контакты для запроса расширенных протоколов испытаний и демо-доступа к стенду: АНО «НИИ Системного синтеза» (ОГРН 1267700083617, ИНН 5027128588 / 5027161602), главный руководитель: Филипп Александр Алексеевич, <https://isslab.ru/investor/>, <https://isslab.ru/mitr/>, электронная почта: info@isslab.ru.

Проверено по первоисточникам: 25.10.2026.