SWORDFISH SECURITY

КОНСОРЦИУМ ИССЛЕДОВАНИЙ БЕЗОПАСНОСТИ ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Методика тестирования безопасности систем ИИ

Теория и практика

Башарин Антон

Старший управляющий директор, Свордфиш Секьюрити
Руководитель РГ2, Консорциум исследований безопасности технологий ИИ

Зачем нужна методика

SWORDFISH SECURITY

Точка отсчёта — 1 марта 2026, ФСТЭК №117

П. 60 — защита данных, моделей, параметров, процессов

П. 61 — статистические критерии и реагирование на недостоверные ответы

Поверхность атак — весь жизненный цикл:

данные → разметка → обучение → реестр артефактов → деплой → инференс → RAG / агенты

Методика (п.3.18) — 12 апреля 2026, утверждена ФСТЭК

Методика тестирования ИИ

SWORDFISH SECURITY

Что внутри

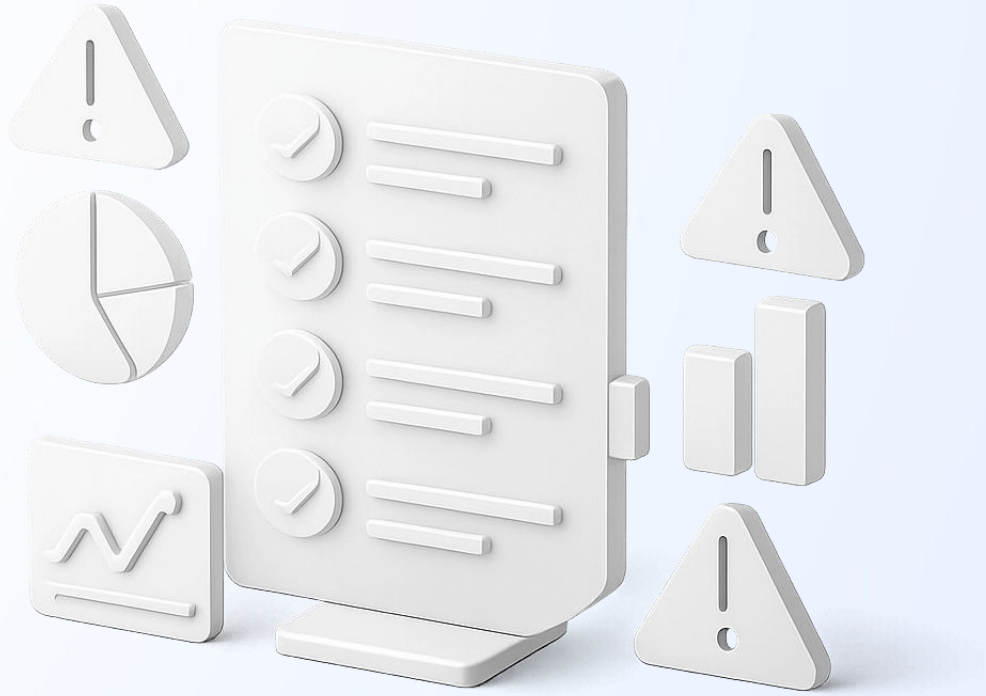
117 требований · **9** этапов ЖЦ · **99%** приходится на 6 этапов

Объекты защиты: ПО · данные · модель

Охват: предсказательные, генеративные, диалоговые модели

Опираемся на:

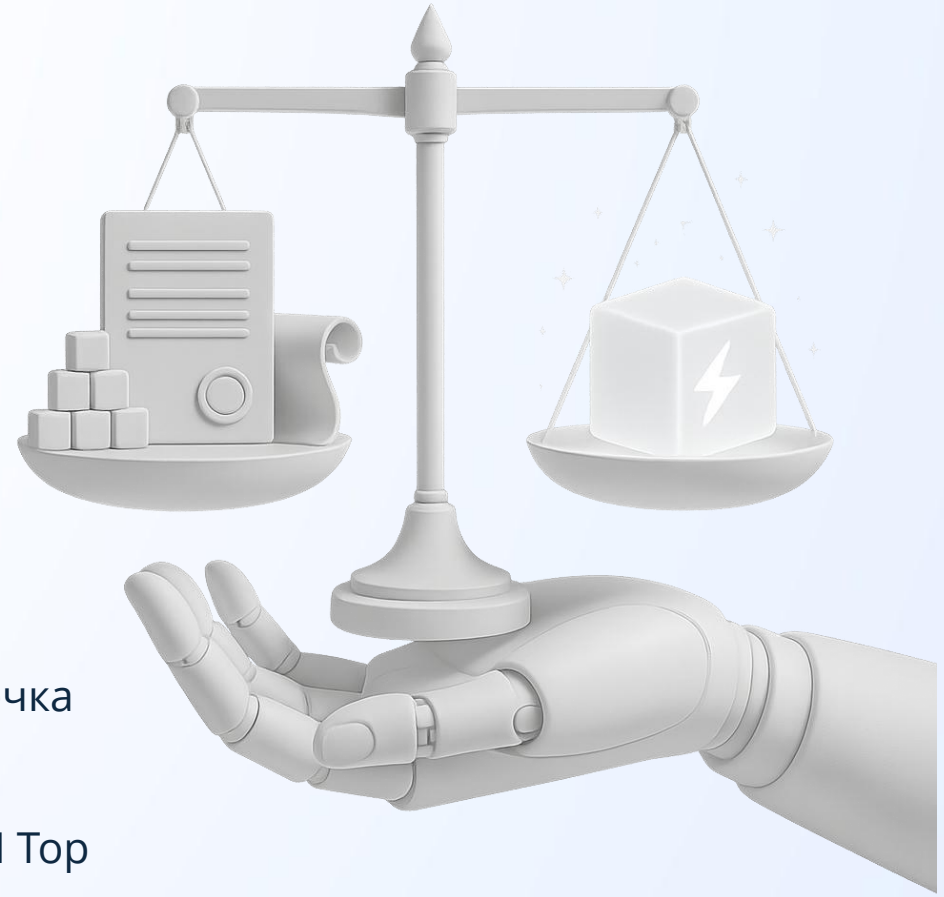
- **Российский контур:** ФСТЭК №117 (п. 60–61), БДУ ФСТЭК, Методический документ от 12.04.2026
- **Мировой контур:** OWASP AI Testing Guide, NIST AI RMF, OWASP LLM Top 10 (2025)



Ключевые требования

SWORDFISH SECURITY

25 / 117



Приоритизация:

1. **Регуляторы** — ФСТЭК №117 + БДУ ФСТЭК + Методика ФСТЭК
2. **Критичность** — необратимые последствия: отравление, утечка PII, prompt injection, jailbreak, supply chain RCE
3. **Частота и популярность** — массовые угрозы из OWASP LLM Top 10 2025 и публичных AI-инцидентов 2024–2025

Ключевые требования по фазам ЖЦ

SWORDFISH SECURITY

Фаза	Ключевые ТР	Что проверяем
Данные / разметка	ТР-2.1.1, 2.1.3, 2.4, 4.5, 4.16	Отравление, PII, утверждённые источники, ML-BOM, криптоподпись версий
Обучение / деплой	ТР-2.9.4, 6.5, 6.11, 7.2, 7.10	Безопасность предобученных моделей, окружение, подпись при деплое, bias и уязвимости, верификация экспорта
Runtime / мониторинг	ТР-9.1.x, 9.4.x, 9.6.x, 9.7, 9.10.x	Guardrails, robustness, галлюцинации, защита параметров и API, DoS / exfiltration

В каждой фазе — свой класс контроля

Автоматизация проверок

SWORDFISH SECURITY

Класс контроля	OSS	Коммерческое
Контроль данных и lineage	Great Expectations, DVC, lakeFS	SOMA MLOps GIS, SafeERP
Supply chain ML	Trivy, ModelScan, Sigstore Cosign, ML-BOM (CycloneDX)	**AppSec.Track**, CodeScoring
Состязательное тестирование	Garak, PyRIT, IBM ART, TextAttack	**AppSec.GenAI**, HiveTrace
Runtime / LLM Firewall	NeMo Guardrails, Presidio, Detoxify	**AppSec.AIGate**, HiveTrace
Drift и качество	Evidently, Cleanlab, NannyML	SOMA MLOps GIS

Гейты: SCA — на CI · red-team — pre-release · guardrails и drift — runtime

Ручные проверки

SWORDFISH SECURITY

- **Архитектура и threat modeling** — TP-1.1, 1.2 — сценарная сессия по STRIDE / LINDDUN
- **Утверждение источников данных** — TP-2.4 — правовой статус, лицензия, due diligence
- **Двойная аннотация и согласие аннотаторов** — TP-4.3, 4.8 — экспертное решение о достаточности разметки
- **Human-in-the-loop при высокорисковых действиях** — TP-9.1.5
- **Сценарный red-team** — chained attacks, prompt injection → tool use → exfiltration

Автоматизация снимает 70–80% объёма; оставшиеся 20–30% — экспертиза и governance

Развитие методики

SWORDFISH SECURITY

В ближайшее время

- **RAG** — манипуляция индексом, подмена источников, тестирование релевантности и устойчивости к подмене контекста
- **Эмбеддинги** — семантический дрейф, утечка векторных представлений, защита механизма поиска
- **Агентные системы и RL** — reward hacking, ограничение автономии, сценарное тестирование в симулированной среде

В перспективе

- **Уровни доверия** — единый язык для регулятора, заказчика и интегратора
- **Профили зрелости** — минимальный, корпоративный, КИИ

В заключении

SWORDFISH SECURITY

Теория

- Методика как единый язык: 117 требований, 9 этапов, российские и мировые контуры в одном
- ~25 ключевых требований из 117 — фильтр «регулятор × критичность × частота»
- Граница автоматизации проходит через управление и процессы
- Методика — живой стандарт; следующая итерация — RAG, эмбеддинги, агенты, уровни доверия

Практика

- Инвентаризация ИИ-артефактов: модели, датасеты, RAG-источники, агенты
- Карта применимых требований по своему ландшафту
- Три практики: OSA/SCA, red-team, LLM Firewall
- Ручные проверки: регламент, ответственный — не «по запросу»

SWORDFISH SECURITY

КОНСОРЦИУМ ИССЛЕДОВАНИЙ БЕЗОПАСНОСТИ ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Башарин Антон

Старший управляющий директор

@nirahsab
anton@swordfishsecurity.ru

Telegram-канал