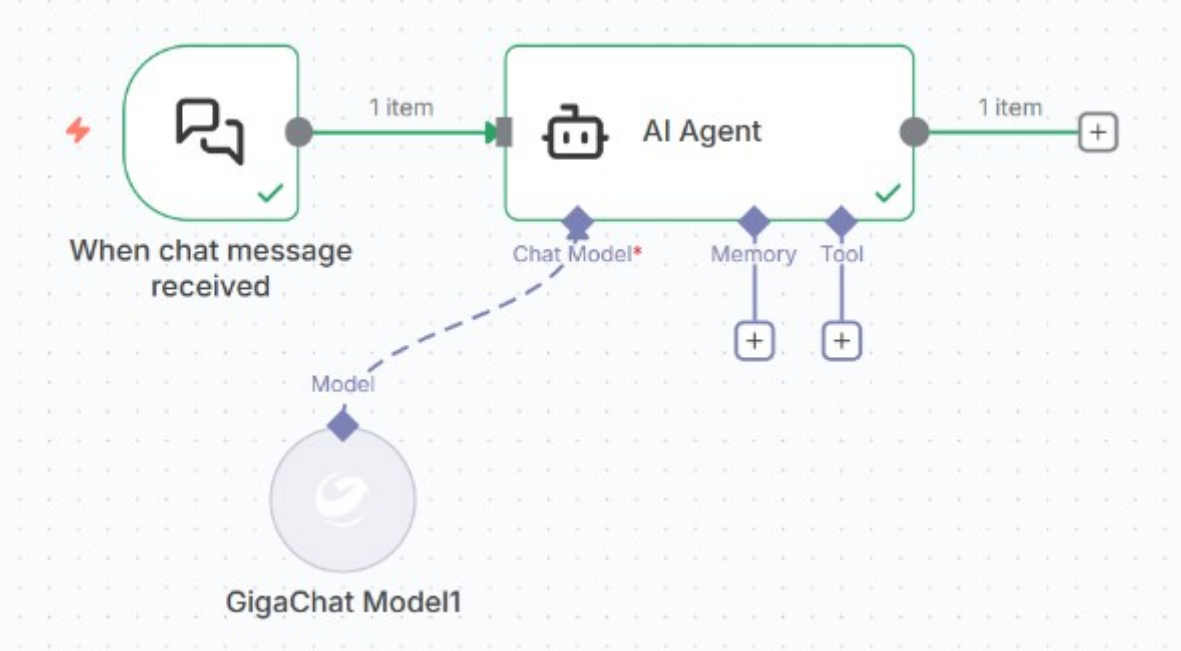
Верификация доверия в отношении систем искусственного интеллекта

Лившиц Илья Иосифович
Профессор практики, доктор технических наук
Факультет безопасности информационных технологий
Университет ИТМО

+7 (921) 934-48-46
Livshitz.il@yandex.ru

1. Система искусственного интеллекта как объект (1)

Особенностью объекта исследования является **высокая неопределённость** «генерирования» выходного результата, не позволяющая надежно **верифицировать функции безопасности** (встроенные или «наложенные») в компонентах систем ИИ, что может **негативно воздействовать** на объекты критической инфраструктуры, в состав которых встраиваются эти компоненты.



1. Система искусственного интеллекта как объект (2)

В стандарте ISO/IEC серии 42001 приведены системные требования к созданию, внедрению и эксплуатации систем ИИ, но **не приведено никаких требований** для проверки критически важного «разумно предсказуемого неправильного использования» (*reasonably foreseeable misuse*) и соответствующего выбора мер ИБ.

В стандарте ISO/IEC серии 27001 не содержит конкретных мер по обеспечению ИБ для систем ИИ, хотя, при ограничениях, некоторые организационные (A.5) и технические (A.8) меры ИБ могут рассматриваться.

В стандарте ISO/IEC серии 42005 описано влияние систем ИИ (*Artificial intelligence system impact assessment*), но **не установлены специфические процедуры в отношении мер ИБ**.

Решение задачи для объектов ИИ возможно по комбинации методов оценки возможных потерь:

a) ISO/IEC серии 22301 (обеспечение непрерывности бизнеса);
b) ISO/IEC серии 42005 (оценивание влияния систем ИИ);
c) ISO/IEC серии 27001 (системы менеджмента ИБ).

2. Система искусственного интеллекта. Факторы риска

Для оценки «стоимости» инцидентов ИБ рассмотрим релевантные кейсы применительно к системам ИИ:

- В экспертной оценке ExploitDog «ИИ ускорил атаки быстрее, чем компании успели перестроить защиту» показано, что **ущерб от одного киберинцидента в среднем составляет 100–200 млн. руб.**
- В отчете «Code Red 2026: Актуальные киберугрозы для российских организаций» Positive Securities показано, что в ближайшие годы активная цифровизация и внедрение систем ИИ будут расширять поверхность атак, что может привести к большему ущербу.
- В обзоре «Искусственный интеллект в промышленности: страхование рисков» научного журнала «Промышленные страницы» приведены примеры критических сбоев в работе систем ИИ:
  a) В 2023 г. в Магнитогорске произошёл сбой системы ИИ, отвечающей за контроль качества продукции на металлургическом заводе. В результате была поставлена некачественная сталь, **ущерб составил более 50 млн. руб.**
  b) В 2024 г. на химическом заводе в Казани система ИИ, управляющая АСУТП, допустила ошибку, что привело к выбросу вредных веществ и травмам 12 рабочих, только **страховые выплаты превысили 15 млн. руб.**

2. Система искусственного интеллекта. Верификация

Необходимая терминология:

- 3.8.12 верификация (*verification*): Подтверждение, посредством представления объективных свидетельств, того, что установленные **требования были выполнены**.

  - ✓ **Примечание 1.** Объективное свидетельство, необходимое для верификации, может быть результатом контроля или других форм определения, таких как **осуществление альтернативных расчетов** или **анализ документов**.

- 3.8.13 валидация (*validation*): Подтверждение, посредством представления объективных свидетельств, того, что **требования, предназначенные для конкретного использования или применения, выполнены**.

  - ✓ **Примечание 1**. Объективное свидетельство, необходимое для валидации, является результатом **испытания** или других форм определения, таких как **осуществление альтернативных расчетов** или **анализ документов.**
  - ✓ **Примечание 3 Условия, применяемые при валидации, могут быть реальными или смоделированными**.

2. Система искусственного интеллекта – факторы оценки (1)

Для целей данной публикации рассмотрим описание УБИ применительно к системам ИИ в соответствии с БДУ ФСТЭК России.

Следует принять во внимание, что национальный стандарт ГОСТ Р 59898-2021, устанавливающий требования к качеству систем ИИ, **исключает специальные характеристики (в том числе ИБ)** из перечня **верифицированных**.

УБИ	Способ реализации
УБИ в инфраструктуре разработчика системы ИИ	СП.5 – Несанкционированная модификация (отравление) наборов обучающих данных СП.6 – Несанкционированная модификация параметров (весов) модели машинного обучения и LoRA, а также данных RAG
УБИ в инфраструктуре оператора системы ИИ	СП.4 – Модификация (изменение) параметров (весов) модели машинного обучения в целях нарушения функционирования системы ИИ СП.5 – Модификация конфигураций модели машинного обучения и агентов (например, системного промпта) в целях нарушения функционирования системы ИИ

2. Система искусственного интеллекта – факторы оценки (2)

Стандарт ISO/IEC 42005:2025 определяет аспекты влияния систем ИИ

Ключевое понятие: «**Разумно предсказуемое неправильное использование» (*reasonably foreseeable misuse*, п. 3.6), как использования систем ИИ способом, изначально не предусмотренным разработчиком (поставщиком), но которое может быть результатом предсказуемого поведения конечных пользователей.**

Примеры метрик, которые могут быть применены при оценивании мер ИБ при отсутствие прямых аналогов в существующих международных стандартах ISO/IEC:

1. ожидания заинтересованных сторон применительно к оценкам воздействия систем ИИ (п. 5.4.1. e);
2. распределение обязанностей по организации процесса утверждения и эскалации (п. 5.6. f);
3. определение функциональных возможностей системы ИИ, например, типы данных модели (п. 6.3.2. a);
4. определение показателей, используемых для оценки моделей, например, точность (п. 6.5.4. g);
5. примеры ущерба (таблица C1).

3. Система искусственного интеллекта. Методы оценки

Меры ИБ, применимые для оценки систем ИИ

Мера	Норматив	
	Приказ ФСТЭК № 117	ГОСТ Р ИСО/МЭК 27001
Обеспечение защиты информации при использовании ИИ	п. 34. т)	A.5.1
Выполнение мониторинга ИБ с применением доверенных технологий ИИ	п. 49	A.5.22
Определение шаблонов запросов пользователей, направляемых в системы ИИ и обеспечение контроля соответствия запросов установленным шаблонам	п. 61 а)	A.8.26
Разработка статистических критериев для выявления недостоверных ответов систем ИИ для последующего сбора и анализа недостоверных ответов	п. 61 в)	A.8.28 A.8.33
Обеспечение реагирования на недостоверные ответы систем ИИ посредством ограничения области принимаемых решений и (или) реализации функций ИС на основе недостоверных ответов систем ИИ	п. 61 г)	A.8.33 A.8.34
Проведение **регулярного независимого аудита ИБ**	-	9.2.2
Выполнение оценки и обработки рисков (**остаточных рисков**) **ИБ**	-	6.1.3

3. Система искусственного интеллекта. Пример расчета (1)

Оценка систем ИИ. Формула и расчёты

$$R_{IT-Security} = \frac{Loss_R - (Sec_{Emb} + Sec_{Add})}{Loss_R}$$

где:

R $_{IT\text{-}Security}$ — результирующая оценка эффективности мер ИБ (%);

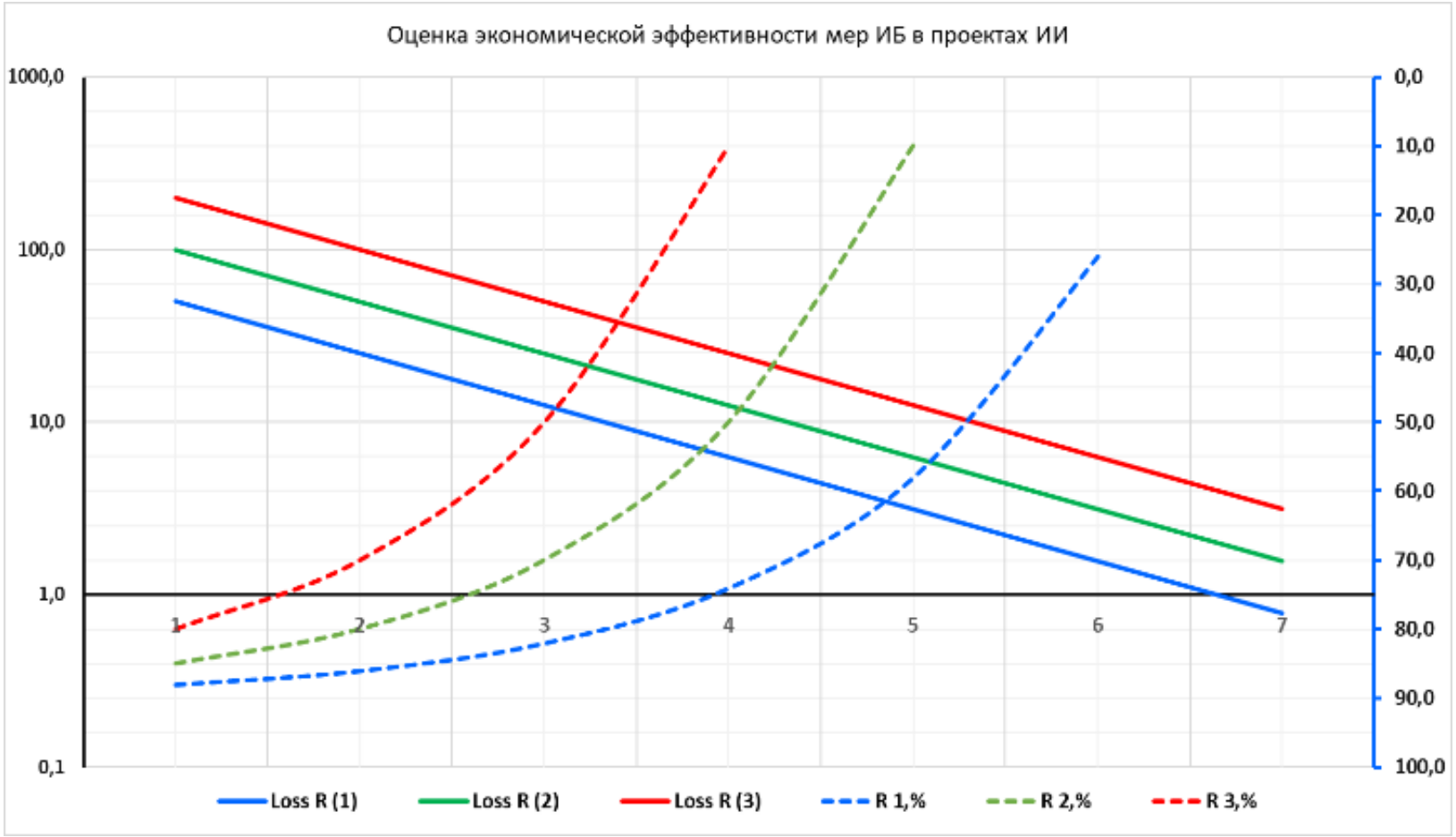
Loss $_{R}$ — результирующая оценка потерь;

Sec $_{Emb}$ — совокупная стоимость встроенных мер ИБ;

Sec $_{Add}$ — совокупная стоимость дополнительных мер ИБ.

Сценарий 1 (Консервативный)				**Сценарий 2** (Базовый)				**Сценарий 3** (Стресс-тестирование)			
Loss $_{R}$	**Sec** $_{Emb}$	**Sec** $_{Add}$	**R** $_{IT\text{-}Sec}$, %	**Loss** $_{R}$	**Sec** $_{Emb}$	**Sec** $_{Add}$	**R** $_{IT\text{-}Sec}$, %	**Loss** $_{R}$	**Sec** $_{Emb}$	**Sec** $_{Add}$	**R** $_{IT\text{-}Sec}$, %
50,00	1	5,00	**88**	100,00	5	10,00	**85**	200,00	20	20,00	**80**
25,00	1	2,50	**86**	50,00	5	5,00	**80**	100,00	20	10,00	**70**
12,50	1	1,25	**82**	25,00	5	2,50	**70**	50,00	20	5,00	**50**
6,25	1	0,63	**74**	12,50	5	1,25	**50**	25,00	20	2,50	**10**
3,13	1	0,31	**58**	6,25	5	0,63	**10**	12,50	20	1,25	-
1,56	1	0,16	**26**	3,13	5	0,31	-	6,25	20	0,63	-
0,78	1	0,08	-	1,56	5	0,16	-	3,13	20	0,31	-

3. Система искусственного интеллекта. Пример расчета (2)



Определена закономерность:

При фиксировании затрат на меры ИБ (Sec $_{Add}$) по норме 10% от стоимости предполагаемого ущерба и фиксировании затрат на встроенные меры (Sec $_{Emb}$) по нормам для различных сценариев оценка эффективности R $_{IT\text{-}Security}$ смещается «наверх», т.е. снижение до **88 %** при увеличении допустимых ущербов.

Библиотека:

1. Лившиц И.И. Интеграция требований к системам менеджмента искусственного интеллекта в автомобильной промышленности // Стандарты и качество. 2025. № 12. С. 46-50.
2. Лившиц И.И. Влияние современных технологий искусственного интеллекта на безопасность промышленных систем автоматизации // Автоматизация в промышленности. 2025. № 6. С. 34-37.
3. Лившиц И.И. Применение современных методов оценивания рисков информационной безопасности объекта критической информационной инфраструктуры // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25. № 4. С. 727-736.
4. Лившиц И.И. Исследование влияния "наложенных" мер защиты на оценку уровня полноты функциональной безопасности для объектов критической информационной инфраструктуры // Автоматизация в промышленности. 2024. № 7. С. 9-13.

Верификация доверия в отношении систем искусственного интеллекта

Благодарю за внимание, пожалуйста, ваши вопросы

Лившиц Илья Иосифович
Профессор практики, доктор технических наук
Факультет безопасности информационных технологий
Университет ИТМО

+7 (921) 934-48-46
Livshitz.il@yandex.ru