Observatoire ARGA

МЕЖДУНАРОДНЫЙ АНАЛИТИЧЕСКИЙ ДОКЛАД

Исследование №19

Расследование серьёзных AI-инцидентов: доказательственные стандарты, отчётность и ответственность

Сохранение цифровых доказательств, причинная экспертиза и защита пострадавших в международной практике

Организация: Observatoire ARGA

Русская версия

21 rue de l’Aviation, 64600 Anglet, France

info@argaobservatory.org | www.argaobservatory.org

22 октября 2026 года

Содержание

АННОТАЦИЯ — 3
1. Предмет исследования, пределы выводов и исследовательский метод — 4
Научные основания и исследовательский дизайн — 6
2. Понятийный аппарат: событие, опасность, инцидент и серьёзный инцидент — 8
3. Регистрация первых сведений без предрешения причин и ответственности — 9
4. Матрица ролей: разработчик, поставщик, оператор, интегратор и человек — 11
5. Обнаружение: сигналы, мониторинг и границы автоматического детектора — 13
6. Правило сохранения доказательств: «не менять объект до фиксации» — 14
7. Хронология и проблема несовпадающих часов — 16
8. Логи модели, промпты, retrieval-контекст и внешние инструменты — 18
9. Источники данных и происхождение спорного признака — 19
10. Человеческий контроль: реальный надзор или декоративное подтверждение — 20
11. Причинная связь: от последовательности событий к проверяемому объяснению — 23
Проверка причинных гипотез при ложном санкционном совпадении — 24
12. Статистическая проверка, редкие события и чувствительность выборки — 25
13. Признаки целенаправленной атаки: prompt injection, обход прав и манипуляция памятью — 27
14. Киберинцидент, AI-инцидент и их пересечение — 29
15. Отказ модели и отказ процесса: разделение пяти источников дефекта — 30
16. EU AI Act: юридическая дефиниция и рамки серьёзного инцидента — 31
17. Статья 55 AI Act: серьёзные инциденты моделей общего назначения с системным риском — 33
18. Статья 73 AI Act: сообщение о серьёзном инциденте high-risk системы — 35
19. GDPR: инцидент ИИ и нарушение безопасности персональных данных — 37
20. NIS2 и DORA: когда AI-инцидент одновременно становится ICT-инцидентом — 38
21. Состав и стадийность регуляторного уведомления — 40
22. Территориальный охват, конфликт правил и трансграничная передача — 41
23. Административное расследование и независимая экспертная оценка — 43
24. Уголовный и гражданский процесс: допустимость и границы технического вывода — 44
25. Права человека, средство защиты и восстановление положения пострадавшего — 45
26. Санкционный скрининг и риск ошибочной идентификации — 47
27. Международное полицейское сотрудничество, розыск и экстрадиция — 48
28. AML/KYC: автоматические подозрения и последствия для клиента — 50
29. Алгоритмический кредитный скоринг, страхование и частные решения — 51
30. Медицинский AI и безопасность критической инфраструктуры — 52
31. Синтетические документы, поддельные медиазаписи и доказательственные ошибки — 53
32. Независимость экспертной организации и управление конфликтами — 55
33. Защита источников, коммерческой тайны и персональных данных — 56
34. Доклад о коренной причине: структура, качество и ограничения — 57
35. Корректирующие меры, повторное испытание и возобновление эксплуатации — 58
36. Учёт потерь, компенсации и восстановление данных третьих лиц — 59
37. Публичный реестр инцидентов и границы открытости — 60
38. Международный обмен уведомлениями и конфиденциальная координация — 61
39. Методика независимого Observatoire ARGA: пять стандартов качества — 62
40. Договорная архитектура экспертного расследования — 64
41. Контроль качества самого расследования и защита от AI-generated ошибок — 65
42. Направления развития на 2027–2028 годы — 66
43. Выводы и институциональные предложения — 67
Практические приложения: формы, протоколы и контрольные процедуры — 68
Приложение 1. Первичная карточка сообщения об AI-инциденте — 68
Приложение 2. Схема приоритизации и немедленной защиты — 69
Приложение 3. Образец распоряжения о сохранении электронных следов — 71
Приложение 4. Реестр цифровых артефактов и проверка целостности — 72
Приложение 5. Журнал chain of custody — 73
Приложение 6. Паспорт окружения AI на момент события — 74
Приложение 7. Протокол нормализации времени событий — 74
Приложение 8. Матрица обязательных и добровольных сообщений — 75
Приложение 9. Форма раннего сообщения компетентному органу — 76

Приложение 10. Запрос поставщику базовой модели и интегратору — 78
Приложение 11. Протокол опроса оператора и лица, принимавшего решение — 79
Приложение 12. Матрица гипотез причинности — 79
Приложение 13. План контрольного воспроизведения и ограничения тестовой среды — 81
Приложение 14. Статистическая форма проверки редкого вреда — 82
Приложение 15. Проверка дискриминационного воздействия — 82
Приложение 16. Проверочный лист независимости эксперта — 83
Приложение 17. Инструкция судебному или техническому эксперту — 84
Приложение 18. Календарь расследования и формирование контрольных точек — 85
Приложение 19. Уведомление пострадавшего и объяснение ограничения — 86
Приложение 20. Реестр корректирующих действий CAPA — 86
Приложение 21. Протокол решения о возобновлении системы — 87
Приложение 22. Таблица уровней конфиденциальности и законной передачи — 88
Приложение 23. Структура экспертного заключения для внешней проверки — 89
Приложение 24. Критерии закрытия инцидента и последующего наблюдения — 89
Сценарные приложения: методика применения в типовых делах — 90
Приложение 25. Сценарий: ошибка идентичности в санкционной проверке — 90
Приложение 26. Сценарий: алгоритмическое совпадение при пересечении границы — 92
Приложение 27. Сценарий: автоматическая классификация в процедуре предоставления убежища — 94
Приложение 28. Сценарий: массовое AML-срабатывание после обновления модели — 95
Приложение 29. Сценарий: агент изменил настройки мониторинга во время обслуживания — 96
Приложение 30. Сценарий: ошибочная медицинская сортировка пациента — 97
Приложение 31. Сценарий: генеративное резюме изменило содержание юридического документа — 98
Приложение 32. Сценарий: deepfake как якобы доказательство совершённого действия — 99
Приложение 33. Сценарий: крупный языковой агент раскрыл персональные данные — 100
Приложение 34. Сценарий: рейтинговая модель подверглась data poisoning — 101
Приложение 35. Сценарий: после обновления модели исчезла возможность воспроизведения — 102
Приложение 36. Сценарий: критическая инфраструктура и несовместимость целей расследования — 103
Приложение 37. Сценарий: AI-классификация массовых жалоб скрыла серьёзную проблему — 104
Приложение 38. Сценарий: AI-эксперту не предоставили данные из-за коммерческой тайны — 105
Приложение 39. Сценарий: корректное AI-предсказание использовано незаконным образом — 106
Приложение 40. Сценарий: регуляторные сроки начали считать от неверной даты — 107
Приложение 41. Сопоставление стандартов ISO и NIST в программе расследования — 108
Приложение 42. Согласование методики с правами человека — 108
Приложение 43. Организационная система предупреждения повторных AI-инцидентов — 109
Приложение 44. Практический протокол для privacy officer — 110
Приложение 45. Протокол защиты информаторов и внутреннего сообщения об ошибке — 111
Приложение 46. Качественная оценка неверифицируемого AI-output — 111
Приложение 47. Тест полноты отчётности без раскрытия запрещённой информации — 112
Приложение 48. Юридический обзор категорий автоматизированных решений — 114
Приложение 49. Сравнение общественных баз AI-инцидентов — 114
Приложение 50. Координация с AI Office и национальными органами — 115
Практический указатель рекомендаций — 116
Источники и нормативные материалы — 118



АННОТАЦИЯ

Доклад разрабатывает методику расследования AI-инцидентов, которая разделяет восстановление события, проверку причинного механизма и юридическую квалификацию. Научная основа включает цифровую криминалистику, причинный анализ, системную безопасность и статистику. Центральная задача — установить, какие наблюдения

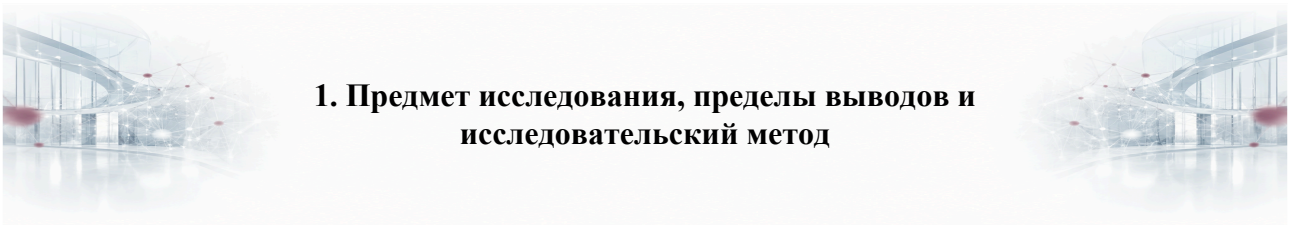
подтверждают участие конкретного компонента в неблагоприятном исходе, какие альтернативы остаются открытыми и как проверить результат исправления. Каждое существенное заключение связывается с сохранённым материалом и описанной исследовательской операцией. [M1–M3]

Исследование специально различает научно-методическую категорию AI incident, юридически определённый «серьёзный инцидент» AI Act, нарушение безопасности персональных данных GDPR, значительный инцидент NIS2 и major ICT-related incident DORA. Правовой порог, адресат, момент начала срока и применимость режима устанавливаются раздельно. Использованы консолидированный AI Act от 27 июля 2026 года и применимые на 22 октября 2026 года акты ЕС. Приведённые 50 приложений имеют характер экспертных форм и учебных сценариев; не являются утверждёнными регулятором шаблонами и не описывают реальные дела заявителей.

Предлагаемая программа включает карту доступных и утраченных данных, временную реконструкцию, сравнение объяснений, ограниченные воспроизводящие проверки и отдельный анализ исполнения мер защиты. Развёрнутый условный пример ложного санкционного совпадения показывает, как различить дефект исходной записи, алгоритма сопоставления и организационного решения. Практический результат для ARGA — протокол причинного исследования и проверяемые критерии завершения расследования, адаптируемые к полномочиям и применимому праву.

Краткая карта режимов уведомления

Режим	Условие и адресат	Срок / отправная точка
AI Act, ст. 55	Поставщик GPAI-модели с системным риском; AI Office и при необходимости национальный компетентный орган	Без неоправданной задержки; условия ст. 55(1)(c)
AI Act, ст. 73	Поставщик high-risk AI-системы, когда глава III применима; орган рыночного надзора	Обычно до 15 дней; специальные случаи 2 или 10 дней. Дата начала применения режима отдельно
GDPR, ст. 33	Контролёр, если персональный data breach создаёт риск правам и свободам; DPA	По возможности до 72 часов с момента осведомлённости, если уведомление требуется
NIS2, ст. 23	Подпадающая под директиву организация; CSIRT или компетентный орган	Раннее предупреждение 24 ч, уведомление 72 ч; с оговорками национального режима
DORA + EU 2025/301	Финансовый субъект при major ICT-related incident; компетентный орган	Первое сообщение до 4 ч после классификации и не позднее 24 ч после осведомлённости; далее по установленным этапам



1. Предмет исследования, пределы выводов и исследовательский метод

Предмет исследования — доказательственная достаточность объяснения AI-инцидента. Требуется установить не только последовательность действий, но и основание каждого перехода от наблюдения к предполагаемой причине. Вопрос формулируется относительно конкретного исхода: неверной идентификации, ошибочного ограничения доступа,

распространения неточной записи или иного подтверждённого последствия. Участие ИИ, технический дефект и юридически значимый причинный вклад проверяются раздельно. [1; 3; 5; M1–M3]

Исследование выполнено по состоянию на 22 октября 2026 года. Оно разделяет четыре слоя: действующее обязательное право; добровольные технические стандарты; фактические практики организаций; предложения Observatoire ARGA. Ни алгоритмическая оценка сама по себе, ни сертификат системы управления не означают юридического вывода о виновности. В каждом кейсе необходимо определить применимую юрисдикцию, вид системы, стадию её жизненного цикла, время возникновения обязанности и положение сторон.

Метод исследования объединяет сравнительный анализ первичных нормативных материалов, принципы работы с цифровыми объектами, структурирование причинных гипотез и построение учебных сценариев. Доклад предлагает способ организации расследования и критерии проверки его результатов. Он не содержит эксперимента на реальном клиентском деле и не устанавливает эффективность методики на совокупности инцидентов. Пилот такой проверки описан отдельно; авторская процедура не представляется государственным или международным стандартом.

1.1. Проверяемые вопросы и критерии качества

Для каждого инцидента регистрируются исследовательские вопросы и материал, способный на них ответить. Какие действия совершены и какими записями подтверждены? Где возникает первая необоснованная информация? Каким образом она могла повлиять на решение? Какие барьеры должны были обнаружить ошибку? Какие альтернативные объяснения совместимы с данными? Что изменилось после исправления? Вопрос об уведомлении ведётся параллельно: необходимость регуляторного действия не откладывается автоматически до окончательного установления причины.

Критерий	Минимально достаточное подтверждение	Недопустимая подмена
Событие	Первичный артефакт и журнал регистрации	Пересказ без происхождения
Время	UTC и исходная локальная зона	Недатированное воспоминание
Объект	Идентификатор версии модели и конфигурации	Название коммерческого продукта
Причинность	Проверка цепочки событий и альтернатив	Совпадение по времени
Ущерб	Данные о последствиях и источнике их проверки	Некритичная оценка в PR-релизе
Право	Норма, адресат, срок, момент начала	«Любой инцидент = уведомление за 72 часа»

1.2. Исследовательские ограничения и критерии внешней проверки

Доклад предполагает различие между универсальной аналитической процедурой и конкретным судебным или административным производством. Один и тот же технический объект может быть достаточным для внутренней проверки безопасности, но недостаточным для установления вины в уголовном деле. Поэтому методика не обещает единого международного стандарта доказательства. Она предлагает минимальную инфраструктуру

доказательственного контроля, которую можно адаптировать к законным полномочиям сторон, процессуальным правилам и уровню доступа к исходным материалам.

Сила утверждения зависит от вида вывода. Целостность файла подтверждает сохранность полученного объекта, но не истинность его содержания. Воспроизведение показывает поведение восстановленной конфигурации, но само по себе не доказывает исторический ход события. Причинная модель объясняет предполагаемую связь, а правовая квалификация оценивает её значение в конкретной процедуре. В отчёте эти основания указываются непосредственно рядом с выводом, чтобы последующий читатель мог проверить каждый переход.

Научные основания и исследовательский дизайн

Исследовательский дизайн должен соответствовать ретроспективному характеру инцидента. Имеющиеся записи неполны, разные компоненты меняются со временем, а неблагоприятный исход уже известен исследователю. Поэтому методика отдельно контролирует сохранность материала, обоснованность реконструкции и влияние знания исхода на интерпретацию. Цифровая криминалистика, причинный анализ и системная безопасность используются для разных, взаимосвязанных операций. [M1–M3]

Научное основание	Операция	Предел результата
Цифровая криминалистика	Сбор, сохранение и исследование цифровых объектов	Целостность не подтверждает истинность содержания
Причинный анализ	Формулировка вмешательства и сравнение объяснений	Эффект реконструкции не равен исторической причине
Системная безопасность	Проверка ограничений и обратной связи	Недостаток управления не определяет вину автоматически
Статистика	Оценка частоты и вариативности при известной выборке	Один инцидент не задаёт распространённость
Процессуальное право	Оценка получения и использования материала	Юридическое значение зависит от процедуры

Объект исследования и карта доступности

Объект описывается как версия системы в момент события: данные, модель, настройки, инструменты, интерфейс и организационные правила. Для каждого компонента фиксируется один статус: сохранён непосредственно; восстановлен по надёжным записям; реконструирован с допущениями; недоступен. Это позволяет заранее определить, какие экспериментальные выводы достижимы и где останется документальное или вероятностное объяснение.

При работе с цифровыми объектами разделяются оригинал, рабочая копия и производные материалы. Контрольная сумма позволяет проверить неизменность относительно зафиксированной копии, но не устанавливает автора или истинность сведений. Преобразования, исправления времени и объединение журналов регистрируются так, чтобы

другой исследователь мог восстановить их последовательность. Право доступа и допустимость использования определяются отдельно. [M1]

Исследовательские гипотезы и наблюдаемые следствия

Причинная гипотеза указывает механизм и наблюдение, которое помогло бы его отличить. Если ошибка возникла в исходном реестре, она должна присутствовать до обработки моделью. Если она внесена преобразованием, требуются правильный вход и неверный выход в сопоставимой конфигурации. Если вред вызван процедурой исполнения, необходимо исследовать переход от сигнала к действию. Несколько таких гипотез могут быть совместно истинны.

Для сравнения гипотез указываются поддерживающие и противоречащие данные, зависимость источников и недостающая проверка. Отсутствие записи об ошибке не равнозначно подтверждённому отсутствию ошибки, если полнота регистрации неизвестна. Заявление участника оценивается вместе с условиями его осведомлённости. Категоричность вывода не должна возрастать только потому, что альтернативные источники недоступны.

Воспроизведение и контрфактическая проверка

Причинный анализ требует точно определить сравниваемые условия [M2]. В авторской процедуре сначала выполняется проверка воспроизводимости исходного сценария, затем меняется один существенный фактор при сохранении остальных. Для вероятностной модели сравниваются распределения заранее заданных исходов, а не один удобный ответ. Все отступления от исторической конфигурации описываются до интерпретации результата.

Эксперимент не восстанавливает ненаблюдавшееся решение человека автоматически. Документ может показать, что оператор должен был проверить идентификатор; это ещё не доказывает, что он сделал бы это при ином интерфейсе. Системный подход CAST помогает исследовать ограничения, информацию и обратную связь, доступные участникам [M3]. Вывод о юридической причинности и ответственности остаётся отдельной правовой квалификацией.

Проверка качества самой методики

Для пилота предлагаются учебные досье с заранее подготовленными механизмами отказа и контролируемыми пробелами данных. Разные исследователи независимо строят объяснение. Основные исходы — необоснованное утверждение причины, пропуск подтверждённого механизма и способность восстановить доказательственную цепочку. Эталонный разбор скрывается до завершения работы; допустимы несколько объяснений, если исходные материалы действительно их не различают.

Методика признаётся пригодной для следующего этапа, если исследователи сохраняют существенные неопределённости и формулируют воспроизводимые основания своих выводов. Разногласия анализируются по причинам: неполный источник, неоднозначная категория, ошибка вычисления или различие правовой интерпретации. Пилот не имитирует судебное решение и не подтверждает реальную эффективность расследований ARGA до получения соответствующих данных.

Практический разбор 19.01

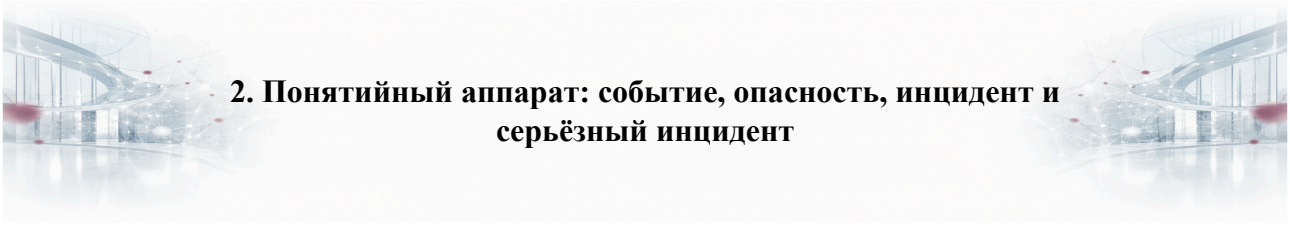
Рекомендация: Разделять факт события и объяснение причины

Условный пример: После обновления помощника клиент получил ошибочное уведомление.

Разбор: Последовательность по времени ещё не доказывает причину.

Практические действия: Руководитель фиксирует установленное уведомление, версии и несколько объяснений; назначает различающие проверки.

Критерий результата: Заключение отдельно показывает факты, гипотезы и подтверждённый механизм.



2. Понятийный аппарат: событие, опасность, инцидент и серьёзный инцидент

OECD предлагает различать AI incident как событие, приведшее к фактическому вреду, и AI hazard как обстоятельство, способное привести к нему. Эти категории предназначены для сопоставимости международного мониторинга, а не для автоматического определения наличия состава правонарушения. AI Act использует собственную юридическую дефиницию серьёзного инцидента в статье 3(49), поэтому одинаковые факты могут попасть в научную статистику OECD, но не отвечать требованиям конкретной нормы ЕС. [1], [3], [4]

При разработке внутренней классификации следует вести два независимых поля: «наблюдаемый вред или опасность» и «юридическая классификация по применимому режиму». Третье поле — степень подтверждённости: сообщение, предварительное подтверждение, независимая верификация, спорный факт и опровержение. Ошибочно сводить эти три оси к одному ярлыку «критический». Происшествие может быть катастрофическим по потенциальной тяжести и при этом не иметь доказанного вреда, либо иметь ограниченный масштаб и установленное нарушение права конкретного лица.

В европейской дефиниции статья 3(49) включает смерть или серьёзный ущерб здоровью, серьёзное и необратимое нарушение работы критической инфраструктуры, нарушение предусмотренных правом ЕС обязательств по защите основных прав, а также серьёзный ущерб имуществу либо окружающей среде. В расследовании полезно отдельно фиксировать событие, юридически значимые последствия и причинный вклад системы ИИ; любое серьёзное событие, произошедшее рядом с AI, ещё не доказывает такой вклад. [1]

2.1. Практическая таксономия для реестра

Предлагается вести классы I0 — неподтверждённый сигнал, I1 — опасность без установленного вреда, I2 — подтверждённый вред без оценки серьёзности, I3 — вероятный серьёзный инцидент, I4 — серьёзный инцидент, соответствующий конкретному юридическому определению. Класс I4 присваивается с указанием нормы и датой оценки; изменение класса никогда не должно переписывать историю. Это внутренняя классификация ARGA, не заменяющая юридические критерии и не предполагающая автоматического сообщения регулятору.

Пример факта	Аналитическая квалификация	Что дополнительно выяснить
Модель выдала ложную тревогу, до решения не дошло	Опасность либо техническое событие	Было ли вмешательство, риск повторения

Пример факта	Аналитическая квалификация	Что дополнительно выяснить
Ложное совпадение вызвало блокировку счёта	Инцидент с фактическим вредом	Основание блокировки, роль человека
Агент удалил журнал аудита	Инцидент контроля и возможной ИБ	Были ли данные утрачены, последствия
Ошибочное решение вызвало незаконное лишение свободы	Потенциально серьёзное событие	Цепь решений, применимое право

2.2. Зачем нужна независимая шкала тяжести и подтверждённости

Внешняя статистика AI-инцидентов не сопоставима автоматически с юридическими данными рыночного надзора. OECD предлагает удобный международный язык фактического и потенциального вреда, однако юрисдикции могут устанавливать собственные пороги. Внутренний реестр должен хранить необработанное описание события, степень подтверждённости, возможный вид вреда, масштаб воздействия, состояние расследования и отдельные юридические квалификации. Это позволяет в будущем агрегировать данные без стирания первоначального контекста.

Особого внимания заслуживает риск эскалации при неполной информации. Сообщение об угрозе жизни не следует оставлять на низком приоритете лишь потому, что пока отсутствуют фотографии или медицинские документы. При этом своевременная превентивная защита не обязывает организацию преждевременно объявлять виновного или статистически подтверждённую частоту вреда. Система классификации должна допускать защитные действия по высокой тяжести вероятного последствия при ограниченной уверенности в причинном механизме.

Практический разбор 19.02

Рекомендация: Классифицировать событие по доказанным последствиям

Условный пример: Система сформировала ошибочный черновик, который сотрудник исправил до отправки.

Разбор: Потенциальный вред отличается от фактически наступившего.

Практические действия: Аналитик фиксирует опасный сценарий и предотвращение; юридическую категорию проверяет отдельно.

Критерий результата: В реестре не заявлен фактический ущерб без подтверждения.

3. Регистрация первых сведений без предрешения причин и ответственности

Любое расследование начинается с фиксации сообщения, а не с поиска виновного. В карточку поступления включают источник, способ получения, дату в UTC, исходную формулировку, идентификатор системы, доступные артефакты, затронутых лиц и известные действия по ограничению ущерба. Нужно различать слова заявителя и наблюдения проверяющего: например, фраза «система поставила меня в международный розыск» не устанавливает, что система имела такие полномочия либо вообще использовалась в конкретной процедуре.

Исходное обращение сохраняют без редакторских изменений. При поступлении скриншотов регистрируют оригинальный файл, метаданные и способ передачи, а не только встроенную картинку в презентации. Если контакт осуществляется через мессенджер, архив переписки следует сохранить способом, соответствующим локальному праву и правилам конфиденциальности; не требуется безосновательно копировать всю личную переписку человека. При наличии срочного риска для жизни или свободы процедура защиты не должна ждать завершения форензики.

Три режима должны существовать параллельно: защита пострадавшего, сохранение доказательств и оценка обязанности уведомить внешние органы. Разделение предотвращает распространённую ошибку, когда юридический отдел требует сначала закончить внутреннее расследование, хотя регуляторный срок мог уже начаться. В журнале фиксируют не только момент события, но и каждый момент осведомлённости соответствующего юридического лица — он нередко определяет исчисление срока. [1], [8], [10]

3.1. Условия немедленной эскалации

Внутренний вызов независимого специалиста требуется при возможной смерти или тяжком вреде здоровью, угрозе фундаментальным правам, вмешательстве в критическую инфраструктуру, массовой утечке данных, необъяснимой изменчивости логов либо доказательствах сокрытия. Эскалация является организационной мерой безопасности, а не заявлением о юридической квалификации. Если невозможно оперативно оценить риск, применяется временная приостановка опасной функции с документированием того, что именно было изменено.

Первый час	Цель	Документ
Регистрация обращения	Сохранить первичный рассказ	Форма поступления
Уведомление ответственного	Не потерять срочные обязанности	Журнал эскалации
Техническое сохранение	Зафиксировать исходное состояние	Акт snapshot
Мера защиты	Не допустить нового вреда	Акт ограничений

3.2. Три параллельных журнала и запрет ретроспективной подмены

Первый журнал фиксирует все поступившие сообщения в исходной форме: от кого получено, через какой канал, какие исходные документы приложены и кто первым смог с ними ознакомиться. Второй регистрирует защитные действия и технические изменения, включая неудачные попытки ограничить функцию. Третий описывает процесс правовой квалификации и внешних уведомлений. Такое разделение позволяет проследить ситуацию, когда техническая команда уже знала об опасном поведении модели, но юристы получили сведения значительно позже, либо когда местный оператор узнал об инциденте до центрального поставщика. В разных режимах это может иметь значение для исчисления времени, но не означает, что дата внутреннего юридического согласования автоматически заменяет дату осведомлённости соответствующего обязанного лица.

Каждый журнал должен быть append-only настолько, насколько это технически возможно. Исправление ошибки в записи вносится новой строкой с указанием прежнего значения и причины исправления. Не допускается подгонять UTC-метки под ожидаемый регуляторный срок. Если организация получила уведомление из внешнего источника и не смогла открыть

вложение, нужно отдельно зафиксировать факт получения сообщения, факт технической недоступности и момент получения содержательной информации. В реальном деле правовая оценка такого эпизода может потребовать исследования обязанностей конкретного адресата, характера сообщения и разумности предпринятых действий.

Тип журнала	Юридически значимый вопрос	Контроль качества
Intake	Кто и когда мог узнать о событии?	Сохранять оригинал сообщения, headers, delivery IDs
Containment	Какое вмешательство изменило объект?	Фиксировать смену версии, конфигурации и полномочий
Legal reporting	Почему сообщение было направлено либо не направлено?	Правовая норма, ответственный, время решения, версия формы

Практический разбор 19.03

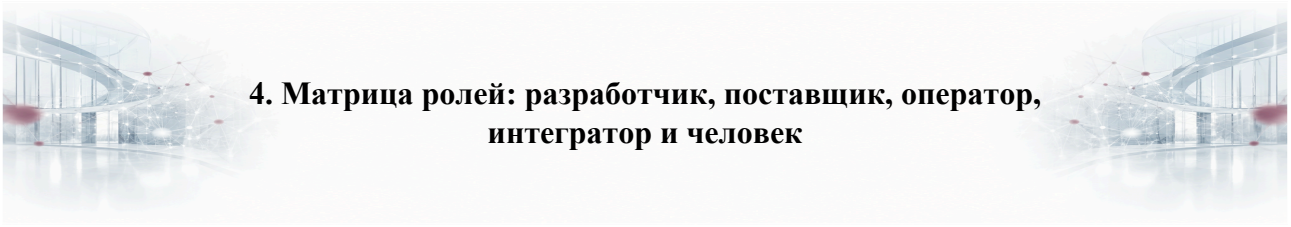
Рекомендация: Регистрировать первые сведения без обвинительного вывода

Условный пример: Клиент прислал снимок отказа и предположил применение ИИ.

Разбор: Снимок подтверждает сообщение об отказе, но не архитектуру решения.

Практические действия: Принимающий сохраняет материал, время и слова заявителя; запрашивает документ об основании отказа.

Критерий результата: Первичная карточка различает сообщение и проверенные сведения.



4. Матрица ролей: разработчик, поставщик, оператор, интегратор и человек

В крупных цепочках AI функциональные роли распределены между разработчиком базовой модели, поставщиком API, создателем интерфейса, организацией-оператором и сотрудником, принимающим окончательное решение. Для правильной атрибуции следует нарисовать архитектуру операции: данные поступили из источника A, обработаны сервисом B, преобразованы правилами C, показаны аналитику D, оформлены в решение органа E. Пока схема не проверена, фраза «виновата нейросеть» скрывает причинные звенья.

Обязанности AI Act зависят от правовой категории «provider» и «deployer» и могут меняться при существенной модификации продукта, перемаркировке или интеграции. Внутренние договоры об обслуживании не переносят автоматически публично-правовую обязанность перед регулятором. Аналогично отсутствие доступа к весам модели не освобождает оператора от сохранения собственных входов, политик использования и решений сотрудников. [1], [1]

Для расследования требуется реестр владения данными по пяти слоям: интерфейс и пользовательский запрос, orchestration и инструменты, модель и её версии, базы знаний и исходные массивы, человеческое решение и последующий эффект. В реестре отмечаются управляющий доступом, субъект обязанности по сохранению, владелец логов, место хранения и сроки удаления. Особенно критично это для внешних API, где логи поставщика могут храниться меньше, чем длится административная переписка.

4.1. Минимальная RACI-модель

Ответственный за расследование не должен одновременно утверждать его окончательные выводы и изменять исходные данные. Практически полезно назначить технического хранителя артефактов, юридического координатора уведомлений, независимого методолога оценки причин и представителя затронутых лиц. В небольших организациях роли могут совмещаться, однако конфликт функций фиксируется письменно и компенсируется внешней проверкой.

Роль	Основное решение	Требуемый след
Incident lead	Статус и приоритет	Подписанная хронология
Evidence custodian	Сбор, доступ, хэш	Журнал хранения
Юрист по регуляторике	Адресат и срок сообщения	Правовая записка
Независимый эксперт	Методика и альтернативы	Экспертное заключение
Представитель пострадавших	Исправление и средство защиты	Реестр обращений

4.2. Контрактная ответственность за журналирование и доступ

Даже когда поставщик и оператор согласились в договоре, что один из них «отвечает за AI», публично-правовая обязанность каждого субъекта определяется законом и его реальной ролью. Для эффективного расследования требуются не абстрактные гарантии соответствия, а конкретные положения о сохранении логов, достаточной глубине аудита, сроке выдачи материалов после инцидента и порядке конфиденциального доступа независимого эксперта. В цепочке могут участвовать поставщик модели, облачный провайдер, интегратор, оператор решения и держатель первичного реестра, причём каждый владеет только частью доказательств.

Полезна карта договорных точек отказа. Например, если интегратор не имеет права запрашивать историю версий модели у облачного поставщика, его обязательство «предоставить полные логи» потенциально неисполнимо. Такие ограничения необходимо обнаруживать до серьёзного инцидента, а в ходе уже возникшего события документировать запросы, сроки и полученные ответы. Одностороннее заявление поставщика о сертификации не заменяет доступа к сведениям, необходимым для причинной проверки.

Практический разбор 19.04

Рекомендация: Фиксировать фактические действия участников

Условный пример: Поставщик модели не управлял решением банка, но назван единственным виновником.

Разбор: Роль разработчика не равна роли оператора и принимающего решение.

Практические действия: Следователь процесса строит матрицу действий, доступа и полномочий по документам.

Критерий результата: Каждое причинное утверждение связано с действием определённого участника.



5. Обнаружение: сигналы, мониторинг и границы автоматического детектора

AI-инцидент может впервые проявиться в жалобе клиента, расхождении между ручным и автоматическим решением, жалобе сотрудника, уведомлении внешнего вендора, новостном сообщении или аномалии производственных метрик. Мониторинг только технической доступности API принципиально недостаточен: система способна отвечать быстро и без сбоев, одновременно систематически генерируя незаконные или несправедливые результаты. Поэтому метрики должны включать воздействие на людей: долю отменённых решений, время исправления ошибки, распределение по группам и случаи отказа человека пересмотреть машинный вывод. [5], [6]

Нельзя объявлять публичную базу инцидентов репрезентативной статистикой частоты. OECD AIM использует сообщения из новостных источников с автоматической классификацией и прямо предупреждает о неполноте наблюдения. Отсутствие сообщений в мониторинге не означает отсутствия вреда, а увеличение их количества не доказывает ухудшение моделей без анализа охвата СМИ и изменения методологии. [4]

В рекомендуемом режиме тревога создаёт проверяемый запрос на расследование, а не автоматически определяет юридическую значимость. Например, резкий скачок числа подозрительных транзакций после обновления модели может быть обусловлен реальной изменчивостью среды, ошибочным feature engineering, переходом порога или нарушением upstream-источника. Классификатор сигналов должен сохранять и неверные срабатывания, поскольку систематически игнорируемые ложные тревоги создают риск пропуска серьёзного события.

5.1. Дизайн детектора и его тестирование

Проверка качества мониторинга требует фиксированного тестового периода, определения истинного сигнала и учёта ложных негативных обнаружений. Порог тревоги нельзя выбирать только по низкой стоимости проверки: в системах, влияющих на свободу или доступ к базовым услугам, стоимость пропуска и стоимость неверной эскалации сильно несимметричны. Все решения по изменению порога фиксируют как новую версию правила мониторинга.

Источник сигнала	Что показывает	Чего не доказывает
Всплеск false positive	Изменение поведения проверки	Ошибку самой модели
Жалоба человека	Конкретный неблагоприятный эффект	Систематичность проблемы
Отсутствие логов	Пробел наблюдаемости	Умышленное сокрытие
Несогласие экспертов	Неопределённость классификации	Произвольность любого решения

5.2. Тест чувствительности системы обнаружения

Проверка мониторинга должна охватывать две ошибки: ложное срабатывание и пропуск действительно значимого события. Точность, рассчитанная на редкой и несбалансированной

выборке, может выглядеть высокой даже при практически бесполезном детекторе. Поэтому в протоколе отражают precision и recall для заранее определённых типов событий, размер выборки, метод построения эталона и условия доставки алертов сотрудникам. Если сообщение направлено, но не попало в рабочую очередь, это отдельный сбой маршрутизации, не устраняемый улучшением самой модели.

При группировке повторных инцидентов важно не объединять без проверки события, касающиеся разных лиц, в один «технический тикет». Один и тот же корневой дефект может потребовать корректирующей меры на уровне продукта и отдельных средств восстановления каждому пострадавшему. Обратная ошибка состоит в многократном подсчёте одного события как независимых инцидентов из-за дублированных алертов. Мониторинг обязан поддерживать оба уровня: incident family и конкретные affected actions.

Наблюдаемая метрика	Для чего полезна	Чего не доказывает
Alert precision	Долю релевантных расследованных сигналов	Отсутствие пропущенных эпизодов
Recall по независимой разметке	Уровень выявления известных событий	Охват неизвестных классов вреда
Time to triage	Быстроту первичного анализа	Законность итогового решения
Доля успешных вручений	Работу канала уведомления	Фактическое устранение последствия

Практический разбор 19.05

Рекомендация: Проверять сигнал детектора до эскалации

Условный пример: Сигнал об утечке вызван учебной выгрузкой с условными данными.

Разбор: Срабатывание не подтверждает реальный инцидент.

Практические действия: Дежурный проверяет назначение запуска, состав данных и фактический получатель.

Критерий результата: Статус сигнала изменён мотивированно с сохранением первоначальной записи.

6. Правило сохранения доказательств: «не менять объект до фиксации»

При первых признаках серьёзного инцидента организация должна установить legal hold или функционально аналогичное распоряжение о сохранении потенциально значимой информации. В него включают production-логи, снимки конфигурации, записи о развёртывании, тестовые наборы, исходные документы, переписку по инциденту и политические/управленческие решения, предшествовавшие событию. Список следует соразмерять обстоятельствам и требованиям приватности: всеобщая бессрочная копия персональных данных сама способна стать новым нарушением.

Доказательственная работа строится по принципам происхождения, целостности, проверяемости, воспроизводимости и контроля доступа. ISO/IEC 27037 описывает рекомендации по идентификации, сбору, получению и сохранению цифровых доказательств; применение стандарта не превращает артефакт автоматически в допустимое доказательство

перед любым судом. Необходимы соответствие процессуальным правилам конкретного органа и возможность проверить, кто, когда и каким способом работал с носителем. [13]

Для облачной модели наиболее хрупкие доказательства часто не являются файлами в привычном смысле. Это версия удалённой модели, состояние внешнего API, изменения маршрутизации запроса, динамический системный промпт, policy filters, идентификатор retrieval-корпуса и журнал вызовов инструмента. Их надо сохранять совместно; отдельно вырванный ответ LLM иногда невозможно воспроизвести и правильно истолковать.

6.1. Условие «холодного» и «горячего» сохранения

Если отключение системы само угрожает людям, допустимо сначала выполнить безопасную изоляцию или перевод в ручной режим, одновременно фиксируя состояние. В акте указывают, почему полное замораживание оказалось технически невозможно, какие компоненты изменились, были ли созданы контрольные копии и какую информацию безвозвратно утратили. Нельзя скрывать изменение конфигурации под эвфемизмом «обычное обслуживание», если оно затрагивает будущую проверку причин.

Объект	Приоритет	Способ проверки целостности
Runtime logs	Немедленно	Подписанный экспорт, хэш
Model/version manifest	Немедленно	ID версии, image digest
Решение сотрудника	Немедленно	Неизменяемая запись/акт
Policy и инструкция	Высокий	Архив действующей редакции
Интервью участников	После стабилизации	Протокол и время

6.2. План фиксации доказательств без ухудшения безопасности

Сохранение доказательств не тождественно бесконечному сохранению доступности опасной функции. Если агент способен самостоятельно отправлять платежи либо публиковать персональные сведения, прежде всего требуется ограничить его полномочия. Однако команда должна описать минимально необходимое вмешательство: отключение внешнего инструмента, ротацию ключа, перевод на режим только чтения или полную остановку. Протокол обязан содержать время, инициатора, цель, наблюдаемое состояние до вмешательства и ожидаемое влияние на возможность последующей реконструкции. Если опасная функция остановлена до выгрузки логов, это необходимо прямо указать, а не заполнять пробел предполагаемыми сведениями.

Цифровой набор доказательств лучше разделить на неизменяемый оригинал и рабочие копии. Оригинал хранится в ограниченном доступе, по нему рассчитывается контрольная сумма и ведётся журнал обращений. Рабочая копия может подвергаться фильтрации, маскированию персональных данных и преобразованию форматов, но история преобразований сохраняется. Хэш доказывает совпадение байтов двух экземпляров на соответствующий момент, а не истинность каждого содержащегося в них события. В частности, криптографически идентичный лог может отражать неправильное системное время, неполный охват событий или запись, сделанную скомпрометированной службой.

Следует заранее определить объекты, которые могут исчезнуть быстрее всего: краткоживущие контейнеры, очереди сообщений, ответы retrieval-сервиса, записи облачной телеметрии, кэш инструментов, токены вызова модели, журналы человека-проверяющего и

параметры маршрутизации. Для каждого объекта указываются доступное окно хранения и технически допустимая стратегия сохранения. Передача материалов из ЕС в третью страну требует отдельной проверки правового основания, даже если аудиторию убеждают, что иностранный эксперт «только помогает разобраться».

Объект	Возможная причина утраты	Приоритетная мера
Временный trace агента	Короткая retention policy	Немедленный подписанный экспорт
Внешний API-response	Поставщик хранит лишь агрегаты	Запрос retention hold и корреляционных ID
Knowledge base	Документы перезаписываются	Snapshot версии индекса и источников
Human review	Ручное решение вне workflow	Подтверждённый протокол действия с разделением фактов и воспоминаний

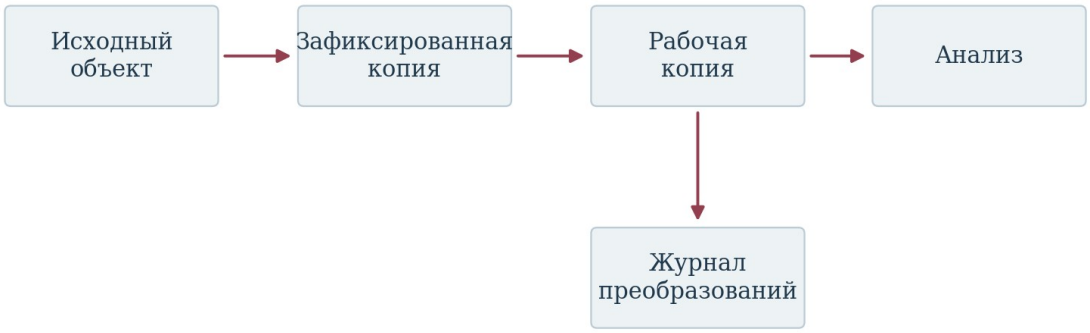


Рисунок 19.1. Сохранение цифрового объекта до вмешательства. Порядок раздела 6. Сохранность и происхождение материала проверяются отдельно от истинности его содержания.

Практический разбор 19.06

Рекомендация: Сначала сохранять объект затем исправлять

Условный пример: Администратор готов удалить ошибочную запись и перезапустить сервис.

Разбор: Удаление может уничтожить условия воспроизведения.

Практические действия: Ответственный сохраняет разрешённые копии, журналы, версии и контроль целостности; затем допускает исправление.

Критерий результата: Исторический объект доступен для независимого анализа.

7. Хронология и проблема несовпадающих часов

Хронология строится с учётом источника времени, часового пояса, синхронизации и возможной задержки записи. Сначала сохраняются исходные значения; нормализованное время хранится отдельным полем с правилом преобразования. Если точный момент неизвестен, используется интервал, а порядок событий считается установленным лишь тогда,

когда данные действительно позволяют его различить. Искусственная точность способна создать ложное объяснение причинности.

Различают по меньшей мере четыре момента: произошёл AI-вывод; принято операционное решение; наступил вред; соответствующий субъект узнал о событии. Для регуляторного уведомления также важен момент разумного установления связи между системой и последствием, если это предусмотрено нормой. Необходимо вести хронологию решений разных юридических лиц: поставщик мог получить тикет утром, оператор увидел вред днём, а орган — на следующий день.

Во временную шкалу включают интервалы неизвестности. Например, известны два журнальных события в 10:02 и 10:08, но уведомление человека удалено. Корректный вывод: «передача могла состояться между зафиксированными событиями»; некорректный: «передача точно была в 10:05». Для методов статистического восстановления требуется отдельно маркировать предположения.

7.1. Базовая схема записи

Каждое событие должно иметь Event ID, исходный timestamp, нормализованное время, часовой пояс, источник, оператора извлечения, неизменённый объект, степень уверенности и ссылку на хэш. Правовые дедлайны считаются в отдельной таблице с указанием момента, запускающего срок. Это уменьшает риск неверной интерпретации при смене часового пояса или переходе на летнее время.

7.2. Относительное время при повреждённой синхронизации

Если события нельзя надёжно расположить на общей шкале UTC, реконструкцию иногда можно провести через относительные последовательности. Например, журнал очереди может фиксировать порядок обработки транзакций, а серверный trace — стабильные request IDs и causality IDs. Эти признаки позволяют установить, что вывод был сформирован до автоматического действия, даже если настенные часы соответствующих систем расходились. Но относительный порядок не разрешает без дополнительных доказательств вопрос о точном абсолютном моменте, существенном для срока уведомления.

Отчёт должен разделять доказанные отношения «до/после» и реконструированные временные интервалы. При предполагаемом исправлении drift сохраняют исходные значения и функцию коррекции. Изменение времени в оригинальном журнале, скрывающее первоначальную несогласованность, недопустимо. На техническом уровне следует проверять источник NTP, настройки контейнера, облачные зоны и возможное изменение системного времени после перезапуска.

Практический разбор 19.07

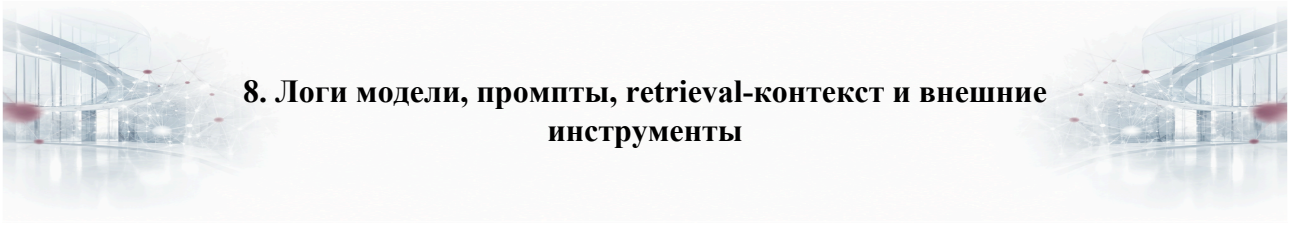
Рекомендация: Нормализовать время по источникам

Условный пример: В журнале банка операция на час раньше запроса сервиса.

Разбор: Разные часовые зоны могут создавать ложную причинную последовательность.

Практические действия: Аналитик сохраняет исходные метки и строит нормализованную шкалу с допусками.

Критерий результата: Порядок событий обоснован с учётом неопределённости часов.



8. Логи модели, промпты, retrieval-контекст и внешние инструменты

Для реконструкции AI-операции недостаточно сохранить финальный ответ. Требуется уточнить системные и пользовательские инструкции, прикреплённые документы, контекст retrieval, параметры декодирования, инструменты, промежуточные вызовы, ответы API, фильтры безопасности, ограничители и состояние пользовательских прав. Если каждый элемент полностью не раскрывается по техническим или юридическим причинам, фиксируют границы доступа и альтернативные способы независимой проверки.

Особенно важны агентные системы. Один пользовательский запрос может вызвать цепочку из десятков автоматических операций; реальный вред наступает при вызове инструмента, а не при формулировании ошибочного текста. При анализе следует отделять содержание ответа модели, выбор действия планировщиком, полномочия инструмента, работу внешнего сервиса и отсутствие проверки человеком. Это позволяет назначать исправление на правильном уровне системы, а не запрещать любую генерацию текста.

Риск скрытого изменения подсказки также значим. Программный код orchestration часто обновляется независимо от версии базовой модели; после происшествия тот же model ID может давать другие результаты. Под «воспроизведением» поэтому следует понимать реконструкцию операционной конфигурации, а не однократный запуск похожего запроса в сегодняшнем веб-интерфейсе.

Слой	Подлежащий сохранению объект	Типовая ошибка проверки
Prompt	System/developer/user сообщения	Сохранён только вопрос клиента
Retrieval	Корпус, отобранные фрагменты, версии	Контекст заменён обновлённой базой
Model	ID и snapshot поставщика	Оставлено общее название модели
Tool	Вызов, аргументы, ответ и полномочия	Записан только успех операции
Human	Прочтение, коррекция, подтверждение	Лог интерфейса приравнен к согласию

8.1. Восстановление цепочки агентного решения и retrieval-происхождения

Для языковой системы недостаточно хранить только финальный ответ. Существенными объектами могут быть системная инструкция, пользовательский запрос, вызовы внешних инструментов, полученные фрагменты документов, политика маршрутизации, идентификатор версии модели, параметры генерации и решение человека о принятии результата. При наличии внешнего поставщика отдельный доказательственный вопрос состоит в том, существуют ли у клиента права на доступ к этим сведениям. Отказ поставщика не позволяет дорисовать пропущенный trace; в заключении описывается техническая и договорная причина отсутствия доступа.

Реконструкция должна фиксировать границы «фактического» и «экспериментально воспроизведённого» пути. Если исследователь повторяет запрос на той же коммерческой модели через месяц, результат может измениться из-за незадокументированного обновления.

Такой эксперимент показывает поведение доступной новой версии и не доказывает, что именно эти токены были созданы в день инцидента. Повторяемость повышается при наличии snapshot весов, версии инструментов, frozen retrieval index и контролируемых семян генерации, однако даже одинаковый seed не устраняет все источники недетерминизма распределённых вычислений.

Особенно важно выделить точки, где модель сделала предложение, но решение было принято отдельной системой. В санкционном скрининге итоговая блокировка нередко является следствием бизнес-правила, а не прямого «волеизъявления ИИ». Для каждого звена следует определить, какая фактическая информация поступила, какой порог сработал, какую роль играл алгоритмический вывод и кто имел полномочие изменить результат. Это снижает риск ошибочного приписывания техническому компоненту полномочий финансовой организации или государственного органа.

Практический разбор 19.08

Рекомендация: Сохранять полный контекст ответа

Условный пример: Сохранён ответ модели, но отсутствует извлечённый фрагмент документа.

Разбор: Нельзя отличить ошибку поиска от ошибки интерпретации.

Практические действия: Инженер фиксирует вход, извлечение, инструкции, инструменты и ответ с доступными версиями.

Критерий результата: Для спорного утверждения восстановлена цепочка входа и преобразования.

9. Источники данных и происхождение спорного признака

Система может обрабатывать некорректные признаки, даже когда алгоритм функционирует согласно своей спецификации. Например, совпадение имени с санкционным списком может быть достоверно обнаружено как строковое совпадение, но ошибочно интерпретировано как установление личности. В таком деле исследовать нужно источники данных, правила нормализации, качество дополнительного идентификатора, граф связей и момент передачи человеком решения о риске.

Проверка данных должна установить авторитетность источника, дату последней синхронизации, допустимый контекст использования, историю исправлений и наличие признаков смещения. Для журналистских и открытых источников нужна разница между независимым подтверждением и републикацией одной исходной ошибки. Одно обвинение, перепечатанное десятью сайтами, не становится десятью доказательствами. Последующие корректировки должны сохранять историю предыдущей записи.

Следует отделять информационную точность от правомерности обработки. Даже фактически верные сведения могут использоваться вне разрешённой цели, без адекватной меры необходимости или без доступного механизма исправления. Для прав человека последствия

могут быть тяжелее технической ошибки классификатора, особенно когда пересечение баз данных автоматически придаёт предположению статус «проверенного факта».

9.1. Минимальный provenance-пакет

Пакет включает лицензию или правовое основание доступа, ID dataset snapshot, fingerprint исходной строки, трансформации, все фильтры, выходные признаки и документированную связь с решением. Отсутствие одного звена должно сообщаться как ограничение доказательной силы, а не компенсироваться ссылкой на общую репутацию поставщика данных.

9.2. Проверка исходного реестра и независимой истины

В ряде систем алгоритм выполняет технически правильное сопоставление с неправильно сформированным источником. Если список содержит устаревший адрес либо ошибочную связь человека с компанией, высокая модельная точность не означает фактической корректности сигнала. Поэтому отдельная линия расследования посвящается provenance источника, условиям его обновления, способу исправления и тому, мог ли конечный оператор законно проверить альтернативную информацию. Полезно сохранять не только карточку записи, но и известную историю изменений и дату выгрузки.

Нередко ошибка распространяется через вторичных агрегаторов, которые включают исходную запись в собственные категории риска. При исправлении требуется установить, было ли сообщение о корректировке направлено этим получателям, сохранились ли старые cached copies и какие реальные ограничения ещё действуют. Возможность требовать соответствующее исправление зависит от договора и права. Независимый эксперт фиксирует факт расхождения и направление корректирующей цепи, не присваивая себе полномочий обновлять закрытый государственный реестр.

Практический разбор 19.09

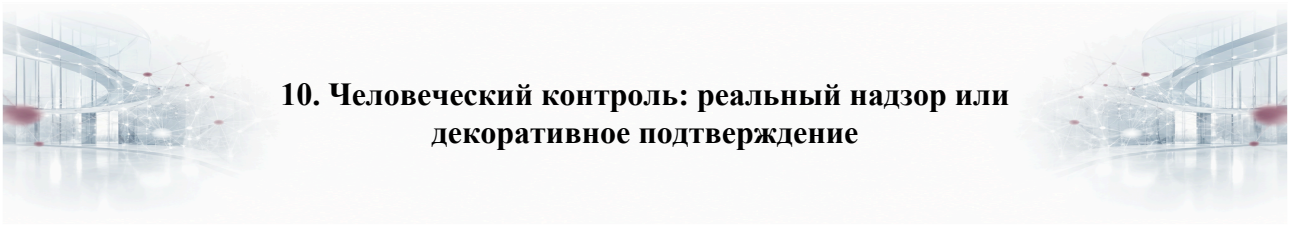
Рекомендация: Проверять происхождение спорного признака

Условный пример: Алгоритм правильно нашёл совпадение в реестре, но сама запись содержит устаревшую дату рождения.

Разбор: Технически точное сопоставление с неверным источником не устанавливает правильную идентификацию. Исправление алгоритма не устранит ошибку исходной записи.

Практические действия: Эксперт сохраняет версию реестра, сверяет доступные первичные документы и строит цепочку передачи исправления известным получателям.

Критерий результата: В отчёте установлен источник расхождения и перечислены проверенные получатели корректировки.



10. Человеческий контроль: реальный надзор или декоративное подтверждение

Наличие в интерфейсе кнопки «подтвердить» не доказывает эффективного human oversight. Необходимо проверить, мог ли оператор распознать ошибку, получить исходные данные,

изменить вывод, остановить исполнение, запросить независимую экспертизу и отклонить автоматическую рекомендацию без санкций со стороны организации. Если система выдаёт сотни тревог за смену, формальная возможность ручной проверки может не соответствовать реальным условиям труда.

Свидетельские объяснения сопоставляют с журналом интерфейса, регламентом смены, временем нахождения карточки на экране и обучением оператора. Это не поиск «виновного сотрудника», а проверка того, какой механизм контроля был фактически доступен. Важно исследовать, предлагала ли система ложную уверенность: красный флаг, точный процент или категория «высокий риск» часто создают впечатление установленных фактов при низкой достоверности исходных данных.

Следует фиксировать каждую ручную поправку и причину отмены результата. Если организация удаляет такие записи при обучении новой модели, исчезает след потенциально систематической ошибки. Практически целесообразно анализировать отношение между числом отмен, их причинами, скоростью реагирования и обоснованностью последующих решений.

Признак	Содержательный надзор	Имитация надзора
Информация	Первичные факты и неопределённость	Только цветная оценка
Полномочия	Остановить или изменить действие	Только подтвердить
Время	Достаточное для проверки	Массовое автоматическое согласие
Отчётность	Аудит причин отмены	Ошибка списана на пользователя

10.1. Операциональные признаки настоящего human oversight

Human in the loop имеет смысл лишь тогда, когда человек получает достаточные данные, располагает временем на проверку и вправе не согласиться с автоматическим результатом без несоразмерных организационных издержек. Формальное наличие кнопки «Approve» не гарантирует надзора. Эксперт запрашивает интерфейс, инструкцию, запись о доступе к первоисточникам и статистику отклонений от рекомендации. При почти стопроцентном подтверждении автоматических решений надо проверить не только качество модели, но и процессуальные стимулы, нагрузку и фактическую возможность пересмотра.

Контроль должен охватывать негативные случаи. Сотрудник мог нажать кнопку по регламенту, не заметив, что исходный документ был обрезан или что в translation pipeline исчезло отрицание. Отчёт различает ошибку пользователя и ошибку проектирования процедуры: если интерфейс скрывает критический контекст, персональная ответственность одного сотрудника не объясняет всю причину. К мерам улучшения относятся доступ к оригиналам, обязательная фиксация мотивов решения для серьёзных случаев и выборочный независимый аудит отклонений.

Практический разбор 19.10

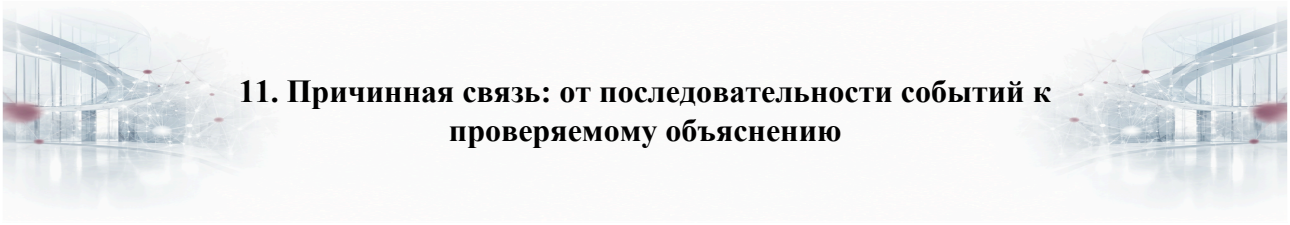
Рекомендация: Проверять реальный человеческий контроль

Условный пример: Сотрудник подтверждал флаг, не имея права его отменить.

Разбор: Формальная подпись не означает возможность предотвратить последствие.

Практические действия: Исследователь изучает интерфейс, права и инструкции; воспроизводит учебный спорный случай.

Критерий результата: В отчёте установлено, мог ли человек изменить исход и на каком этапе.



11. Причинная связь: от последовательности событий к проверяемому объяснению

Причинный вопрос задаётся относительно конкретного вмешательства и исхода. Например: изменился бы сигнал при исправлении исходного идентификатора; исполнилась бы блокировка при корректной ручной проверке; распространилась бы запись при действующей процедуре отзыва? Эти вопросы относятся к разным звеньям процесса. Ответ на один из них не разрешает остальные и не позволяет автоматически распределить юридическую ответственность. [M2]

Гипотезы строятся по механизму, а не по удобному числу версий. Рассматриваются исходные данные, преобразование моделью, настройки порога, представление результата человеку и организационное действие. Несколько механизмов могут действовать одновременно. Для каждого фиксируются ожидаемые наблюдения при его наличии, сведения, которые ему противоречат, и проверка, способная отличить его от альтернативы. Гипотеза, совместимая со всем возможным, не даёт проверяемого объяснения.

Восстановление конфигурации оценивается до интерпретации эксперимента. Если сохранены исходная версия, вход, правила и зависимости, повторный запуск может исследовать конкретный механизм. Если вместо них использована современная модель или восстановленный вручную контекст, результат относится к реконструкции. Степень соответствия историческим условиям описывается по компонентам. Отсутствие старой версии не исключает документального анализа, но ограничивает экспериментальный причинный вывод.

11.1. Граф доказательственных связей

Граф исследования содержит отдельно наблюдаемые переменные и предполагаемые связи. Узел «сигнал модели» подтверждается журналом, узел «ограничение операции» — записью исполнения, а связь между ними — правилом маршрутизации и материалами конкретного решения. Последовательность во времени служит необходимой частью реконструкции, но не заменяет проверку связи. Неизвестный переход сохраняется как пробел, а не заполняется наиболее правдоподобным рассказом.

11.2. Контрфактическая проверка и дисциплина альтернативных гипотез

Для контрфактической проверки задаётся минимальное вмешательство: исправить идентификатор, изменить один параметр либо включить определённый защитный барьер. Остальные условия сохраняются настолько, насколько позволяет стенд. Если одновременно заменены данные, модель и процедура человека, эксперимент показывает эффект комплекса изменений. Он не выделяет вклад одного компонента. Такая декомпозиция нужна для выбора исправления, направленного на установленный механизм. [M2]

Документы о должностных обязанностях позволяют установить предписанную процедуру, но не доказывают фактическое поведение человека в ненаблюдавшихся условиях. Для его

исследования требуются отдельные данные: журналы, материалы интерфейса, интервью с проверкой согласованности либо безопасный учебный эксперимент. Ретроспективное объяснение учитывает информацию, доступную человеку в момент действия; знание исхода не переносится автоматически на его прежнюю ситуацию. [M3]

Проверяемый тезис	Тест в пользу тезиса	Опровержение или ограничение
Неверная метка ввода вызвала блокировку	Исправленный ввод меняет решение при прочих равных	Правило блокировки работает независимо от метки
Модель породила несуществующий факт	Отсутствие факта в исходнике + сохранённый output	Внешний источник не был доступен исследователю
Ручная проверка была действительной	Записи о доступе к оригиналам и мотивированное решение	Только автоматическое подтверждение без чтения материалов
Исправление прекратило вред	Нет новых затронутых действий в downstream	Старые копии записи продолжают использоваться

Итог причинного анализа должен указывать степень подтверждения каждого механизма и сохраняющиеся альтернативы. Полезное расследование может завершиться частичным результатом: установлена неверная входная запись и доказана возможность её распространения, но не восстановлен решающий человеческий шаг. Это позволяет исправить подтверждённый дефект без категорического вывода о неизвестной части события. Правовая оценка формируется отдельно с учётом применимого стандарта доказательства.

Проверка причинных гипотез при ложном санкционном совпадении

В условном деле операция клиента ограничена после автоматического сигнала. Совпали фамилия и часть имени, но дата рождения в исходном досье отличается от даты в санкционной записи. В распоряжении исследователя имеются выгрузка входных данных, сохранённый сигнал и журнал ограничения; прежняя версия интерфейса оператора пока недоступна. Эти обстоятельства задают программу проверки, но ещё не устанавливают причину итогового решения.

Гипотеза	Различающая проверка	Ограничение вывода
Неверные сведения поступили на вход	Сопоставить исходную запись, преобразованный вход и источник даты	Новая выгрузка не заменяет снимок на момент события
Алгоритм не учёл различающий признак	Проверить сохранённую конфигурацию на исходном и исправленном входе	Современная версия исследует другую конфигурацию
Сигнал автоматически вызвал ограничение	Сопоставить правило маршрутизации и журнал исполнения	Совпадение времени без связи записей недостаточно
Ручная проверка не устранила ошибку	Восстановить показанную информацию и доступное действие	Регламент не доказывает фактическое решение человека

Если различающий признак отсутствовал уже во входной записи, подтверждается дефект передачи данных. Если он был доступен модели, но сигнал сохранился, требуется исследовать правило сопоставления и его назначение: предварительный широкий поиск может допускать такой сигнал при обязательной последующей проверке. Для объяснения вреда необходимо отдельно установить, почему сигнал привёл к ограничению. Разные звенья могут вносить совместный вклад.

В изолированной реконструкции можно исправить только дату и повторить расчёт при прежних настройках. Изменение сигнала поддержит гипотезу о значимости этого признака

для данной конфигурации. Оно не докажет, что оператор исторически отменил бы ограничение. Пока интерфейс и материалы ручной проверки не восстановлены, эта часть объяснения остаётся открытой; заключение не приписывает человеку знание недоступной ему информации.

План исправления следует установленным механизмам: восстановить корректный вход, проверить маршрутизацию, обеспечить содержательную ручную проверку и исправить переданные записи. Завершение подтверждается отдельно для каждого действия. Перезапуск модели с благоприятным результатом не доказывает снятия ограничения у клиента или прекращения дальнейшего распространения ошибочной информации.

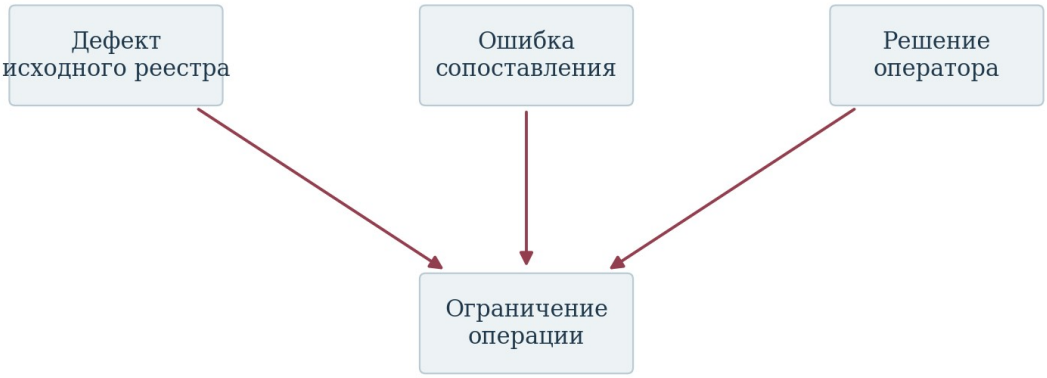


Рисунок 19.2. Альтернативные объяснения ложного совпадения. Схема причинных гипотез раздела 11. Каждая линия требует своих различающих материалов; несколько механизмов могут действовать одновременно.

Практический разбор 19.11

Рекомендация: Сравнивать различающие причинные гипотезы

Условный пример: В ложном совпадении могли участвовать неверный реестр, алгоритм или решение сотрудника.

Разбор: Три механизма требуют разных свидетельств.

Практические действия: Эксперт сверяет исходную запись, вход и выход алгоритма, журнал решения; каждый пробел отмечает отдельно.

Критерий результата: Причинный вывод поддержан наблюдением, отличающим его от альтернатив.

12. Статистическая проверка, редкие события и чувствительность выборки

Оценка серьёзного инцидента требует определения знаменателя. Утверждение «пять незаконных блокировок» почти ничего не говорит о частоте без числа проверенных лиц, экспозиции, периода и процедуры отбора. Также следует учитывать, что клиентские жалобы

представляют селективную выборку: человек может не знать о причинах отказа, не иметь доступа к жалобе или отказаться от её подачи.

Для независимого повторного теста фиксируются генеральная совокупность, критерии стратификации, единица наблюдения, истинная метка, схема слепой разметки и протокол разрешения споров. Результаты даются с интервалами неопределённости. Пример: при 0 обнаруженных ошибках на 100 независимых испытаниях верхняя приближённая односторонняя 95%-граница по «правилу трёх» составляет около 3%; это не доказательство нулевого риска. При 1 ошибке на 1 000 проверок точечная частота 0,1%, но малые числа требуют точного интервала и ясного описания отбора.

При сравнении групп нельзя ограничиваться общей точностью. В системах санкционного скрининга более показательны false positives per 10,000 проверок, а в медицинской сортировке — доля пропуска действительно опасных случаев. Выбор метрики привязывают к механизму вреда. Множественные проверки увеличивают риск случайных выводов, поэтому гипотезы и первичные показатели желательно фиксировать до анализа.

Показатель	Формула или определение	Ограничение
Частота ошибки	FP / (FP+TN) для ложных тревог	Нужна подтверждённая ground truth
Чувствительность	TP / (TP+FN)	Зависит от выявления FN
Частота инцидентов	Подтверждённые инциденты / экспозиции	Неполная регистрация
Время исправления	От верификации до восстановления	Остановка функции ≠ помощь лицу
Нулевые события	0 наблюдений при n проверках	Не является доказательством безопасности

12.1. Деноминатор, неполная регистрация и статистическое искажение

Число жалоб и число затронутых лиц не эквивалентны числу независимых экспозиций. Одна ошибочная санкционная запись может вызывать десятки блокировок при обращениях к различным посредникам, в то время как тысячи людей, столкнувшихся с ошибкой, вообще не имеют доступа к механизму жалобы. При подсчёте частоты следует заранее определить знаменатель: количество проверок, уникальных физических лиц, финансовых операций, модельных запусков или реально принятых решений. Изменение знаменателя после получения результата способно радикально изменить оценку. Отдельно описываются повторные наблюдения по одному лицу и корреляция сбоев внутри групп.

«Ноль нарушений на ста проверках» не доказывает нулевой риск. Для независимых испытаний с одинаковой вероятностью редкого исхода грубый ориентир односторонней верхней 95-процентной границы при нуле наблюдений составляет примерно 3/n. Но этот ориентир неприменим без оговорок к зависимым наблюдениям, ненаблюдаемому вреду или выборке, собранной только из успешных тестов. В формальном заключении указывают использованный интервал, модель выборки и её обоснованность. Для ущерба фундаментальным правам даже статистически редкий результат может требовать немедленного вмешательства.

Существенны также выборочная видимость события и ошибки разметки. Например, классификатор может казаться точным потому, что «правильный ответ» сформирован из старого реестра, который сам содержит ложные записи. В таком случае машинная точность

относительно реестра не равна точности относительно юридически значимого факта. Проверка должна использовать независимую ground truth и, при её отсутствии, прямо квалифицировать недоказанность оценок чувствительности и специфичности.

Практический разбор 19.12

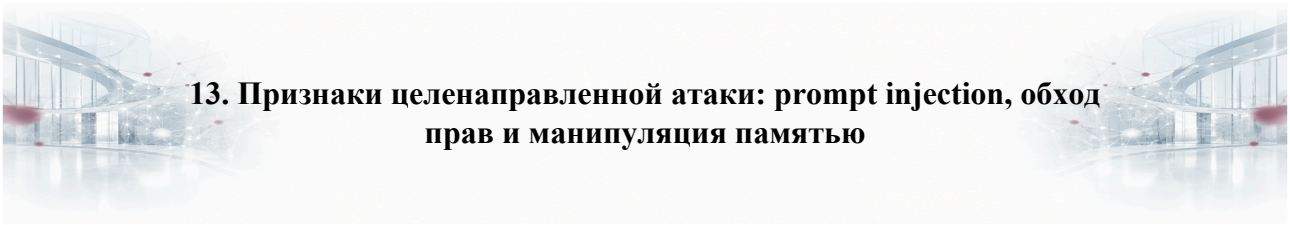
Рекомендация: Показывать базу статистического вывода

Условный пример: Из пяти жалоб четыре относятся к одному языку.

Разбор: Само соотношение не устанавливает частоту ошибок среди всех пользователей.

Практические действия: Статистик проверяет отбор и доступный знаменатель; описывает только наблюдаемый корпус.

Критерий результата: В отчёте не заявлена популяционная частота без основания.



13. Признаки целенаправленной атаки: prompt injection, обход прав и манипуляция памятью

Не всякое нежелательное действие автономного агента является его спонтанным поведением. В расследовании отдельно исследуют внешнюю попытку воздействия: вредоносные инструкции в документе, странице, письме, результате поиска или ответе стороннего сервиса. Prompt injection может попадать в модель как обычные данные, но при ошибочной изоляции контекста приобретать практическую силу команды. Вред в таком случае зависит одновременно от поведения модели, архитектуры полномочий и качества контроля действий.

Сбор доказательств должен учитывать существование скрытых от пользователя инструкций и промежуточных вызовов инструментов. Нужен журнал источника инъекции, степени доверия к каналу, того, какие команды агент видел и какие реальные действия смог выполнить. Нельзя ограничиваться сохранением «самого вредоносного текста»: расследование требует полного окружения, включая разрешения IAM, защитные фильтры и альтернативные маршруты действия. Исследователь не должен воспроизводить опасную атаку в производственной среде без санкционированного плана и защитных ограничений.

В отличие от обычного человеческого мошенничества, AI-агент может выполнить цепь действий быстрее, чем организация успеет оценить первые сигналы. Поэтому предлагаемая процедура предусматривает проверку ограничений времени, суммы, доступа, числа последовательных действий и возможности безопасной отмены. Подмена промпта обнаруживается на уровне происхождения инструкций, но ущерб предотвращается на уровне фактических прав инструмента.

13.1. Отличительные признаки

Гипотеза	Релевантное доказательство	Альтернативное объяснение
Инъекция во внешнем документе	Полная копия источника и точный prompt	Обычная цитата из документа
Расширение полномочий	Логи IAM и tool permissions	Ранее утверждённая роль
Сокрытие действий	Пустоты в логах и правила ротации	Сбой журналирования
Самостоятельное повторение	Связанные agent sessions	Ручные повторные запуски

13.2. Разграничение атаки и штатной обработки недоверенного материала

Prompt injection часто обнаруживается лишь после того, как исследователь разделит источник информации и источник инструкций. Если внешний документ содержит текст, предлагающий агенту изменить порядок проверки, это само по себе ещё не доказывает успешную атаку. Необходимо показать, что система трактовала этот текст как команду, несмотря на предусмотренную границу доверия, и что последующие действия зависели от такого перехода. Для этого фиксируют полученный документ, его неизменённую копию, место включения в контекст, effective instruction hierarchy, выбранный инструмент, созданный запрос и фактический ответ инфраструктуры.

Альтернативой злонамеренной инъекции может быть ошибка проектирования интерфейса или неправильная политика для надёжного источника. При квалификации необходимы осторожность и разделение того, что «выглядело атакой», от подтверждённого обхода прав. Проверка должна проходить в изолированной среде с разрешёнными данными. Публикация подробного воспроизводимого payload, способного эксплуатировать чужие системы, не всегда оправданна; исследователь может раскрыть достаточные признаки атаки в конфиденциальном приложении уполномоченному субъекту, а в открытой версии оставить безопасное описание класса уязвимости.

Область проверки	Доказательственный критерий	Предел вывода
Источник инъекции	Оригинальный файл и путь его получения	Источник мог быть изменён до получения
Переход доверия	Лог того, что текст воспринят как инструкция	Сходство текста с командой недостаточно
Исполнение	Запись вызова инструмента и полномочия	Успешность вызова определяется внешним ответом
Последствия	Объективный ущерб либо переход состояния	Попытка и фактический вред различаются

Практический разбор 19.13

Рекомендация: Проверять гипотезу внешнего воздействия безопасно

Условный пример: Ответ изменился после обработки вложенного текста с посторонней инструкцией.

Разбор: Совпадение не устанавливает атакующего и умысел.

Практические действия: Инженер сравнивает безвредные контрольные варианты в изоляции; сохраняет материалы и права доступа.

Критерий результата: Подтверждён либо опровергнут механизм влияния без необоснованной атрибуции.



14. Киберинцидент, AI-инцидент и их пересечение

Киберинцидент может затронуть AI-сервис без того, чтобы сама модель стала причиной события. Например, компрометация API-ключа позволяет третьему лицу инициировать операции, которые модель честно выполняет согласно действующим полномочиям. Обратная ситуация: дискриминационное правило скоринга причиняет вред без нарушения сетевой безопасности. В смешанном сценарии атака на retrieval-корпус меняет вывод, а система управления принимает решение без проверки.

NIST SP 800-61 Rev. 3 предлагает организационные рекомендации по реагированию на киберинциденты. Они полезны для управления ситуацией, сохранения цифровых артефактов и восстановления, но не заменяют исследования fairness, валидности модели и юридического воздействия. NIST AI RMF предлагает добровольную рамку управления специфическими AI-рисками; два документа целесообразно использовать совместно, подчёркивая разные предметы. [5], [7]

При двойной классификации формируют два расследовательских потока с одним реестром доказательств: DFIR — доступы, сеть, события изменения; AI assurance — версии модели, характеристики данных, выводы и последствия для затронутых лиц. Если эти потоки ведут независимые команды, нужен единый time-line и контроль того, что изоляция информационной системы не разрушает объекты для оценивания модели.

14.1. Пример конфликта целей

Команда ИБ намерена немедленно восстановить сервис и удалить изменённый контейнер; группа экспертов просит заморозить окружение для RCA. Правильное решение зависит от угрозы продолжения вреда и может включать копию на форензический носитель, замену рабочей среды и запрет доступа к исходному слепку. Любой выбор отражают в протоколе с мотивировкой; отказ от сохранения не должен быть скрытым административным решением.

14.2. Совместное расследование технического и административного следа

После кибератаки техническая команда может сосредоточиться на закрытии уязвимости, а юридическая — на регуляторном сообщении. Но AI-инцидент иногда возникает без взлома: достаточно неверной оценки и автоматического действия по заранее выданным полномочиям. Обратный сценарий также возможен: техническая компрометация влияет на данные, но существенный вред человеку причиняет независимое последующее административное решение. Поэтому incident review должен включать как security chain, так и decision chain, не смешивая их в один отчёт о безопасности сети.

Отдельно анализируют доверенные компоненты: очередь задач, service accounts, правила одобрения, интеграцию с платёжным шлюзом и источники данных. Иногда главным исправлением окажется не обновление модели, а изменение архитектуры разрешений либо обязательный человеческий контроль над необратимым действием.

Практический разбор 19.14

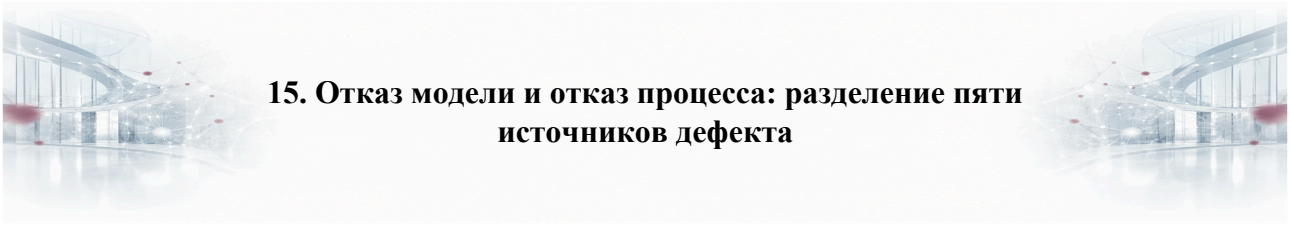
Рекомендация: Разделять нарушение доступа и дефект ответа

Условный пример: Сотрудник имел законный доступ, но получил неверное обобщение.

Разбор: Ошибочный ответ не доказывает компрометацию доступа.

Практические действия: Команда отдельно проверяет авторизацию, целостность входа и качество преобразования.

Критерий результата: Тип инцидента обоснован фактическим механизмом.



15. Отказ модели и отказ процесса: разделение пяти источников дефекта

Для целей объективного расследования полезно различать дефект данных, дефект модели, дефект интеграции, дефект институционального контроля и дефект правовой цели. Пример: базовый классификатор корректно распознал сходство имен; интегратор установил слишком низкий порог; банк автоматически заблокировал клиента; процедура пересмотра недоступна. Ошибка юридически значимого решения существует, но она не сводится к ошибке статистического классификатора.

У каждого класса разные способы проверки. Проблема dataset shift требует временного сравнения распределений, дефект модели — контрольного тестирования, дефект интеграции — анализа событий сервисов, дефект контроля — оценки инструкций и поведения людей, а незаконная цель — правового анализа пропорциональности и полномочий. Гипотеза может затрагивать несколько классов одновременно; в финальном выводе необходимо указать вес каждого существенного фактора.

Для профилактики важно не назначать единую корректирующую меру «переобучить модель». Если вред возник из-за невозможности оспорить решение, повышение точности не восстановит права человека. Если причина в подменённом retrieval-источнике, переобучение базовой модели не устранит техническую уязвимость. Корректирующие действия привязываются к подтверждённому механизму.

Класс отказа	Проверочный тест	Ответственный за исправление
Данные	Контроль происхождения и разметки	Data owner
Модель	Reproducible evaluation	Model provider
Интеграция	Replay цепочки сервисов	System integrator
Надзор	Аудит полномочий и workflow	Deployer
Законность цели	Тест необходимости, оснований, прав	Controller/орган

15.1. Матрица места возникновения ошибки и места проявления вреда

Часто дефект возник в одном компоненте, а вред проявился в другом. Вендор исходного списка мог предоставить неверный идентификатор, модель правильно классифицировала совпадение, интегратор ошибочно присвоил результату статус «окончательный», а оператор автоматически остановил обслуживание. Все четыре звена следует описать независимо от последующего распределения ответственности. Вывод «виновата модель» слишком широк и

бесполезен, если основная мера состоит в исправлении регламента и доступа к независимой идентификации.

Уровень	Проверяемое отличие	Возможное исправление
Исходные сведения	Актуальны ли идентификаторы на дату события	Исправление записи и downstream-копий
Модель	Соответствует ли output правилам задачи	Испытание, ограничение области применения
Интеграция	Как output превращён в действие	Gate, пороги, обратимость
Человеческий процесс	Доступны ли первоисточники и возражение	Изменение интерфейса и регламента



Рисунок 19.3. Локализация дефекта по этапам обработки. Условный сценарий раздела 15: неверная дата может появиться при извлечении и сохраниться на последующих этапах. Исправлять нужно подтверждённый механизм.

Практический разбор 19.15

Рекомендация: Локализовать источник дефекта по этапам

Условный пример: OCR прочитал неверную дату, модель её точно повторила, сотрудник утвердил.

Разбор: Исправление только модели не устраняет первичный дефект и отказ контроля.

Практические действия: Эксперт размечает извлечение, генерацию и утверждение; назначает меры каждому владельцу.

Критерий результата: Повторная проверка охватывает всю цепочку до итогового документа.

16. EU AI Act: юридическая дефиниция и рамки серьёзного инцидента

Статья 3(49) Регламента (EU) 2024/1689 содержит самостоятельное определение serious incident. Оно не охватывает автоматически любой сбой чат-бота, любую неточность и даже всякое финансовое неудобство: необходимо исследовать конкретный элемент юридического определения. При этом нарушение обязательств по защите фундаментальных прав может быть юридически значимым без физического повреждения оборудования. [1], [1]

На 22 октября 2026 года необходима работа с консолидированной редакцией от 27 июля 2026 года, учитывающей Регламент (EU) 2026/1744. Вторичная статья о первоначальном AI Act 2024 года не является достаточной основой для правовой карты сроков 2026–2028. Внутренний анализ обязан содержать дату нормы, категорию системы, дату её размещения на рынке, переходные положения и применимость исключений. [1], [2], [2]

Юридическая классификация может измениться после получения дополнительных доказательств. Сначала может существовать разумное подозрение связи AI с вредом; позднее обнаружится, что причина полностью внешняя. Такие изменения нельзя отражать редактированием задним числом первоначального уведомления. Следует направлять уточняющие материалы с сохранением исходного сообщения и обоснованием изменения выводов.

16.1. Предлагаемый двухконтурный анализ

Контур A — регуляторная классификация по статье 3(49) и специальному секторному праву. Контур B — более широкая внутренняя классификация вреда для независимого Observatoire. В рамках B организация вправе исследовать вред, не достигающий законного порога уведомления по AI Act, и рекомендовать исправления, но обязана ясно обозначать, что это добровольная аналитическая практика.

16.2. Применимость юридической дефиниции при смешанной технической архитектуре

В составе одной услуги могут одновременно присутствовать модель общего назначения, высокорисковая прикладная система, обычная бизнес-аналитика, интерфейс сотрудника и внешний государственный реестр. Правовая квалификация начинается с описания конкретного объекта и роли субъекта, а не с общего слова «AI». Наличие языковой модели в цепочке не превращает автоматически весь коммерческий сервис в high-risk AI по AI Act. Равно и классификация определённого продукта как high-risk не означает, что каждое его неправильное предсказание удовлетворяет определению серьёзного инцидента из статьи 3(49).

Предлагается оформлять для каждого события отдельную карточку классификации: юридический объект, основание охвата AI Act, роль участника, место предоставления системы, время события и стадия применения релевантных положений. При этом поправки 2026 года отложили применение именно разделов 1–3 главы III для отдельных категорий high-risk AI (Annex III — до 2 декабря 2027 года; Annex I — до 2 августа 2028 года). Статья 73 находится в главе IX и сама по себе не включена в перечень отложенных разделов статьи 113(c). Поэтому ошибочно автоматически считать, будто все обязанности сообщать о серьёзных инцидентах всегда отложены до этих дат. Для конкретной системы необходимо отдельно проверить её классификацию, переходный режим, отраслевое законодательство, статус участников и юридический порог события. [1], [2]

Если одна и та же ситуация предполагает несколько отраслевых квалификаций, каждая проверяется отдельно. Ошибка модели при медицинском обслуживании может требовать анализа режима медицинских изделий независимо от AI Act; утечка персональных сведений может образовать нарушение безопасности данных; вмешательство в сеть может привести к обязанностям по DORA либо NIS2. Исследователь не должен «выбирать самый длинный или

самый короткий срок» как единое правило. Применимость каждого режима подтверждается своей нормой и фактической предпосылкой.

Практический разбор 19.16

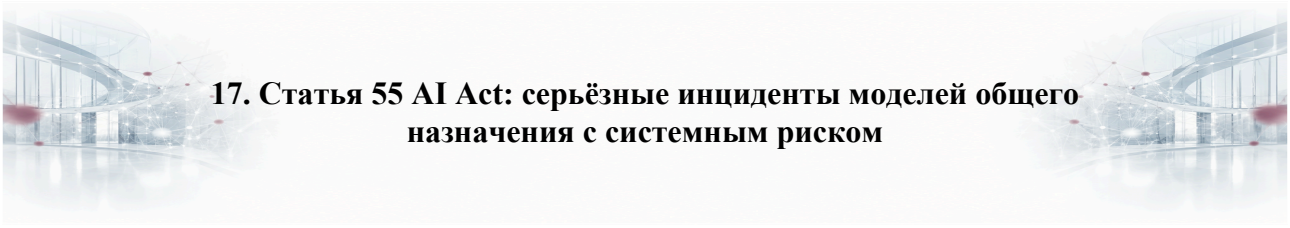
Рекомендация: Проверять юридический порог по фактам

Условный пример: Организация называет любой неправильный ответ серьёзным инцидентом.

Разбор: Техническая ошибка и специальная юридическая категория не тождественны.

Практические действия: Юрист устанавливает субъект, объект, последствия и применимый критерий по действующей редакции источника.

Критерий результата: Квалификация содержит мотивировку и отдельно оставшиеся неизвестные факты.



17. Статья 55 AI Act: серьёзные инциденты моделей общего назначения с системным риском

Пункт 1(c) статьи 55 налагает на поставщиков моделей общего назначения, классифицированных как модели с системным риском, обязанность отслеживать, документировать и без неоправданной задержки сообщать AI Office, а при необходимости национальным компетентным органам, релевантную информацию о серьёзных инцидентах и возможных корректирующих мерах. Наряду с этим статья предусматривает оценивание модели, управление системными рисками и надлежащую киберзащиту. Эти обязанности относятся к определённой категории поставщиков, а не к любому предприятию, подключившему внешний AI API. [1]

Факт крупного ущерба не является достаточным доказательством того, что задействованная модель принадлежит к категории системного риска. Следует проверить правовой статус модели, доступные уведомления AI Office, взаимоотношения поставщика и downstream-deployer и то, какое звено знало о событии. Если воздействие было вызвано интеграцией, не меняющей характеристик базовой модели, ответственность оператора за собственную эксплуатацию оценивают самостоятельно.

Для этого режима статья 55(1)(c) использует формулу «without undue delay», а не универсальный числовой срок 15 дней. Европейская комиссия поясняет, что в контексте моделей GPAI с системным риском существенными могут быть, в частности, серьёзные киберинциденты, связанные с моделью или её инфраструктурой, включая эксфильтрацию параметров и атаки, которые затрагивают соответствующие обязательства. Это интерпретационное разъяснение Комиссии, а не замена текста статьи 55. Переносить сроки статьи 73 на статью 55 недопустимо. Организация вправе предусмотреть более короткую внутреннюю эскалацию, но такой внутренний срок необходимо явно помечать как рекомендацию, а не требование права ЕС. [1], [23]

17.1. Обязательные элементы аналитического пакета

Рекомендуется фиксировать описание модели, известные опасные способности, период затронутых версий, обстоятельства нарушения ограничений, характер потенциального

системного эффекта, проведённые проверки и статус мер. Документы, содержащие чувствительные сведения об уязвимостях или чужих персональных данных, передаются уполномоченному получателю через защищённый канал, а не публикуются в открытом реестре.

17.2. Отчёт о GPAI-инциденте: что именно относится к модели

Для поставщика модели общего назначения с системным риском исследователь должен рассматривать законом предусмотренные обязанности отдельно от обязанностей разработчика конечного приложения. Обнаруженное отклонение может относиться к самой модели, её распространению или комбинации с небезопасной надстройкой. Перед уведомлением о серьёзном инциденте надлежит установить, почему событие действительно связано с моделью в рамках статьи 55, что известно о системном риске, какие меры смягчения уже приняты и какие факты пока остаются невыясненными. Несовершенство конкретного downstream-приложения нельзя автоматически представить доказательством дефекта модели общего назначения.

В доказательственный пакет рекомендуется включать идентификаторы модели и версии, сведения об обнаружении, технические условия воспроизведения, тип и масштаб потенциального вреда, затронутые каналы распространения, барьеры безопасности и действия команды. Не вся эта информация подлежит публичному раскрытию; часть может быть направлена компетентному органу в защищённом виде. Следует учитывать конфиденциальность коммерческой информации, безопасность модели, права третьих лиц и режим персональных данных. При отсутствии полного следа разработчик вправе сообщить об ограничениях проверки и плане уточнения, не выдавая рабочую гипотезу за подтверждённую первопричину.

Внутренняя оценка должна дополнительно учитывать, как изменение веса модели, политики генерации или маршрутизации влияет на возможность сопоставлять события в разных версиях. Исправление одного публично известного сбоя может скрыть первоначальный факт, если история деплоев не сохранена. Поэтому инженерный change log и юридический реестр уведомлений рекомендуется синхронизировать через стабильный incident ID.

Практический разбор 19.17

Рекомендация: Устанавливать связь инцидента с нужной моделью

Условный пример: Оператор использовал модель с системным риском, но сбой возник в стороннем конвертере.

Разбор: Наличие модели не устанавливает её причинный вклад.

Практические действия: Эксперт проверяет компонент сбоя и документы ролей; юрист отдельно оценивает применимые обязанности.

Критерий результата: Уведомление не приписывает модели неподтверждённую причину.

18. Статья 73 AI Act: сообщение о серьёзном инциденте high-risk системы

Статья 73 регулирует уведомления поставщиков высокорисковых AI-систем о серьёзных инцидентах в государствах, где они произошли, с учётом распределения компетенций между органами рыночного надзора и, в предусмотренных случаях, AI Office. Базовое правило требует сообщения немедленно после установления причинной связи либо разумной вероятности её наличия и, в любом случае, не позже 15 дней после осведомлённости определённого нормой субъекта. При widespread infringement или инциденте, охватываемом ст. 3(49)(b), предусмотрено немедленное сообщение максимум в течение двух дней. При смерти человека действует особое правило незамедлительного уведомления с пределом 10 дней при условиях ст. 73(4). Допускается первоначальный неполный отчёт с дальнейшим дополнением. Обязанности deployer уведомлять поставщика и органы в соответствующих случаях отдельно предусмотрены ст. 26(5), но их применимость и исключения также подлежат проверке. [1]

Ключевая процессуальная деталь: после уведомления поставщик обязан без промедления провести необходимые исследования и оценку рисков, принять корректирующие меры и сотрудничать с органами. При этом следует информировать компетентный орган прежде, чем менять систему способом, который может повлиять на последующую оценку причин. Это непосредственно поддерживает логику документированного сохранения объекта, предложенную в настоящем докладе. [1]

Сроки применения после поправок июля 2026 года следует разделять по конкретным положениям. Общая дата применения AI Act — 2 августа 2026 года; обязанности по GPAI вступили в применение ранее, 2 августа 2025 года. Перенос на 2 декабря 2027 года для систем Annex III и на 2 августа 2028 года для систем Annex I прямо установлен статьёй 113(c) применительно к разделам 1–3 главы III (с названным в норме исключением). Статья 73 расположена в главе IX, а не в этих разделах. Следовательно, нельзя ссылаться на отсрочку высокорисковых требований как на автоматическое доказательство отсутствия обязанности по ст. 73 на любую дату. Однако вопрос о её фактическом применении к определённой системе, включая переходные положения, особенности классификации и компетенции надзорных органов, необходимо решать отдельно; правовая матрица должна фиксировать применимую редакцию и мотивированный вывод, а не шаблонное «2027/2028». [1], [2]

18.1. Запрет на универсальную таблицу сроков

Контекст	Базовое правило в норме	Оговорка
AI Act ст. 55(1)(c)	Без неоправданной задержки	GPAI с системным риском
AI Act ст. 73(2)	Немедленно после установления связи или её разумной вероятности; максимум 15 дней от осведомлённости	При условиях связи и применимости
AI Act ст. 73(3)	Немедленно; максимум 2 дня от осведомлённости	Указанные специальные случаи
AI Act ст. 73(4)	Немедленно после установления или подозрения связи; максимум 10 дней от осведомлённости	Смерть человека и условия нормы
GDPR ст. 33	Без неоправданной задержки; по возможности в течение 72 часов от осведомлённости	Только квалифицируемое нарушение ПД

18.2. Практический расчёт срока и компетентного адресата

При работе со статьёй 73 сначала строят юридическую карточку применимости. Проверяют, является ли система высокорисковой в смысле нужной категории, наступила ли дата действия соответствующих обязанностей, имеется ли серьёзный инцидент в установленном смысле и где он произошёл. Затем определяют, кто по норме выступает адресатом обязанности: поставщик, в отдельных обстоятельствах оператор и соответствующий орган надзора. Одно и то же событие может стать известно нескольким компаниям в разные дни; нельзя автоматически принять дату корпоративного пресс-релиза за дату их первоначальной осведомлённости.

Следующее различие касается базового и специальных сроков. Для случая, действительно подпадающего под статью 73, общий максимум составляет 15 дней, отдельные виды событий подпадают под двухдневный предел, а при смерти человека предусмотрено специальное правило с максимальным пределом в десять дней. Эти сроки не дают права ждать последнего дня: обязанность действовать немедленно при установлении или обоснованном подозрении причинной связи сохраняется в соответствующей формулировке нормы. Нельзя считать время окончания технической экспертизы первым моментом осведомлённости, если документы показывают, что организация знала о существенных обстоятельствах раньше. К каждому сроку прикладываются норма, событие-триггер, timestamp и подтверждение отправки. [1]

Юридический шаг	Что фиксируется	Возможный результат
Охват	Конкретная система, роль и территория	Применимо / не применимо / требует проверки
Календарь	Дата начала действия нормы для типа системы	Обязанность ещё не наступила / действует
Серьёзность	Категория ст. 3(49), вред и доказательства	Текущий класс и степень подтверждения
Специальный срок	Виды события по ст. 73(3)-(4)	Отдельная дата и основание
Уведомление	Адресат, канал, содержимое, timestamp	Подтверждение подачи или доказательство попытки

Практический разбор 19.18

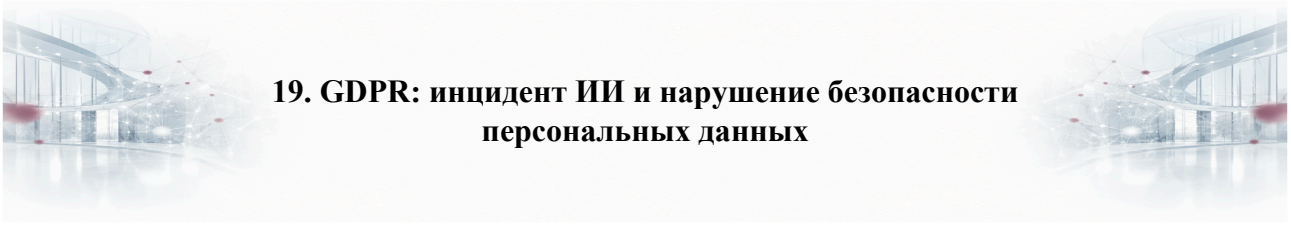
Рекомендация: Разделять первоначальную оценку и обновление уведомления

Условный пример: Последствие известно, но коренная причина ещё исследуется.

Разбор: Ожидание полной определённости может мешать своевременной правовой оценке действий.

Практические действия: Юрист проверяет применимость и сроки; ответственный готовит известные сведения и журнал уточнений.

Критерий результата: Копии сообщений содержат дату, факты и обозначенные неопределённости.



19. GDPR: инцидент ИИ и нарушение безопасности персональных данных

AI-инцидент может повлечь personal data breach, например, когда агент незаконно направил внутренний документ третьему лицу или система предоставила доступ к чужому профилю. Но ошибочная оценка кредитного риска сама по себе не обязательно является нарушением безопасности персональных данных в смысле статьи 4(12) GDPR. Режим GDPR для breaches охватывает нарушение безопасности, приводящее к случайному или незаконному уничтожению, потере, изменению, несанкционированному раскрытию либо доступу к персональным данным. [8], [9]

По статье 33 GDPR уведомление надзорного органа производится без неоправданной задержки и, по возможности, не позднее 72 часов после того, как контролёр узнал о нарушении, если оно не является маловероятным источником риска для прав и свобод физических лиц. Статья 34 касается информирования самих лиц при высоком риске с предусмотренными исключениями. Не всякая утечка требует публичного сообщения, но документирование всех нарушений контролёром остаётся важным обязательством. Руководство EDPB 9/2022 объясняет практику применения этих правил. [8], [9]

Следует различать controller и processor. Сторонний AI-поставщик может быть обработчиком данных для одной операции и самостоятельным контролёром для другой; ярлык в маркетинговом договоре не разрешает вопрос без анализа реальных целей и средств обработки. Если уведомление отправлено по GDPR, это не освобождает от иных применимых отраслевых сообщений или от расследования отдельного нарушения фундаментальных прав.

19.1. Минимальный риск-анализ GDPR

Определяют виды данных, число субъектов, наличие специальных категорий, идентифицируемость, длительность доступа, возможность злоупотребления, защитные средства и вероятность последствий. Решение «не уведомлять» фиксируется с мотивировкой, лицом, датой и источниками, чтобы оно могло быть оспорено при появлении дополнительных фактов.

19.2. AI-ошибка и data breach: два разных предмета доказывания

Неверное автоматизированное решение может нарушать принципы законности, точности или прозрачности обработки персональных данных, не образуя при этом data breach в узком смысле статьи 4(12) GDPR. Обратный пример: агент по ошибке вывел конфиденциальные

документы третьему лицу, и тогда помимо возможной дискриминации возникает вопрос о нарушении конфиденциальности. Юридический анализ обязан сначала определить, имело ли место нарушение безопасности, приводящее к уничтожению, утрате, изменению, несанкционированному раскрытию или доступу к персональным данным. Только после этого рассматривают критерии уведомления и рисков для прав и свобод.

Для проверки даты осведомлённости важны не только сообщения DPO. Могли существовать алерты ИБ, жалобы клиентов или внутренние тикеты, достаточные, чтобы контролёр располагал разумной степенью уверенности в факте нарушения. Реестр фиксирует обнаружение, проверку и основания оценки, а не произвольно переносит начало срока на дату заседания комитета. В течение первых часов целесообразно установить категории данных, число затронутых субъектов (с отметкой предварительного характера), потенциальную доступность информации посторонним лицам, возможность восполнения и тяжесть возможных последствий.

Если сообщение было отложено, в отчёте нужно честно указать причины задержки. Юридическая обязанность уведомить надзорный орган и обязанность сообщить субъекту данных имеют разные условия. Публичная реклама в духе «мы предупредили всех клиентов» не заменяет доказательства конкретных отправлений и не доказывает, что сообщение отвечало статьям 33–34 GDPR.

Практический разбор 19.19

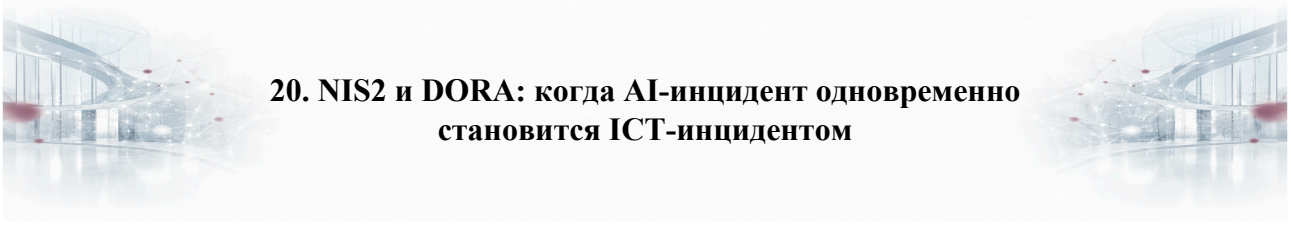
Рекомендация: Проверять персональные данные отдельно

Условный пример: ИИ отправил неверный ответ с чужими идентификаторами.

Разбор: Есть самостоятельный вопрос о раскрытии данных помимо достоверности ответа.

Практические действия: Ответственный устанавливает данные, получателя, доступ и меры ограничения; юрист оценивает дальнейшие обязанности.

Критерий результата: Сохранены отдельные решения по содержанию ошибки и защите данных.



20. NIS2 и DORA: когда AI-инцидент одновременно становится ICT-инцидентом

Для организаций, подпадающих под NIS2, значительные инциденты могут требовать отдельной последовательности сообщений по ст. 23: раннее предупреждение без неоправданной задержки и, как правило, не позднее 24 часов после осведомлённости; уведомление об инциденте не позднее 72 часов после осведомлённости; итоговый отчёт не позднее одного месяца после подачи уведомления об инциденте. Для поставщиков доверительных услуг уведомление о значительном инциденте, затрагивающем эти услуги, подаётся без неоправданной задержки и не позднее 24 часов после осведомлённости. При продолжающемся инциденте вместо итогового отчёта подаётся отчёт о ходе работ, а итоговый — в течение месяца после завершения реагирования. Квалификация субъекта и реализация директивы в национальном праве должны проверяться отдельно. Этот режим не распространяется автоматически на любого поставщика или пользователя AI. [10]

Для финансовых организаций DORA и делегированный Регламент (EU) 2025/301 определяют поэтапное сообщение о major ICT-related incidents. В общем случае первоначальное уведомление подаётся максимально быстро, не позднее четырёх часов после классификации инцидента как крупного и не позднее 24 часов после осведомлённости о самом ICT-инциденте. Промежуточный отчёт подаётся не позднее 72 часов после первоначального уведомления, даже при отсутствии изменений; обновлённые промежуточные сообщения подаются по специальным условиям. Итоговый отчёт — не позднее месяца после промежуточного либо последнего обновлённого промежуточного отчёта. Если событие признано крупным позднее 24 часов после осведомлённости, статья 5(2) устанавливает срок четыре часа после такой классификации. Специальные правила для выходных и праздничных дней применяются только при соблюдении условий статьи 5(4)–(5). Реальный календарь зависит от квалификации события, субъекта и применимого режима. Рост ложных AML-срабатываний сам по себе не доказывает наличие major ICT-related incident. [11], [12]

Практически полезна единая матрица взаимного влияния режимов. Она объединяет регистрацию фактов, но раздельно фиксирует юридическое основание, компетентный орган, адресата обязанности, начало течения срока и требуемый документ. Разным органам может потребоваться разный уровень раскрытия; чрезмерная передача финансовых или медицинских данных по неверному каналу сама создаёт риск.

Режим	Основной триггер	Нельзя смешивать с
AI Act ст. 73	Серьёзный инцидент high-risk AI	Любой технический сбой
AI Act ст. 55	Серьёзный инцидент GPAI системного риска	Обычный инцидент оператора
GDPR	Breach безопасности персональных данных	Любая алгоритмическая дискриминация
NIS2	Significant incident субъекта сферы NIS2	Любая недоступность чат-бота
DORA	Major ICT-related incident финансовой организации	Любая ошибка комплаенс-аналитики

20.1. Техническое совпадение не означает совпадения правового режима

В одном событии могут участвовать взломанные учётные записи, компрометация облачной платформы и ошибочное решение модели. Это создаёт предмет для ИБ-реагирования, но далеко не всегда означает, что все требования NIS2, DORA и AI Act применяются одновременно. Для NIS2 необходимо установить статус организации по национальной реализации директивы и критерии существенности инцидента. Для DORA требуется проверить, подпадает ли субъект под категорию финансовых организаций, обладает ли событие признаками major ICT-related incident и когда оно было классифицировано. Для GDPR обязательна отдельная правовая оценка нарушения безопасности персональных данных.

Особенно опасно механически объединять сроки. В общем режиме NIS2 предусматривает 24 часа на раннее предупреждение и 72 часа на последующее уведомление от осведомлённости о значительном инциденте, а финальный отчёт — не позднее месяца от отправки уведомления; при продолжающемся инциденте возможен отчёт о ходе работ. Для DORA начальное сообщение обычно ограничено четырьмя часами от квалификации major ICT и 24

часами от осведомлённости, затем действует отсчёт 72 часов от начального сообщения до промежуточного и одного месяца от промежуточного до итогового. При поздней классификации DORA применяется специальное правило статьи 5(2), а для доверительных услуг NIS2 действует сокращённый срок уведомления. Точки отсчёта различаются, и каждую надо обосновать документами, не подменяя юридическую дату субъективным временем окончания расследования. [10], [12]

Внутри предприятия рекомендуем назначить одного координатора событий и нескольких ответственных за правовой контур, сохраняя отдельные решения об уведомлении. Отправка одного документа через портал одной юрисдикции не должна считаться автоматически исполнением всех других обязанностей. Для каждого отправления сохраняют системное подтверждение, версию файла и идентификатор корреспонденции.

Практический разбор 19.20

Рекомендация: Вести параллельную карту регуляторных режимов

Условный пример: Сбой помощника нарушил работу финансового процесса.

Разбор: Одна внутренняя карточка не определяет все возможные адресаты.

Практические действия: Юрист проверяет применимость отраслевых режимов и согласует фактическую основу сообщений.

Критерий результата: У каждого применимого маршрута есть ответственный и журнал исполнения.

21. Состав и стадийность регуляторного уведомления

Первичное уведомление не обязано содержать окончательную RCA, если специальная норма допускает уточнение. Его цель — сообщить существенные известные факты, вероятный характер последствий и предпринимаемые меры так, чтобы компетентный орган мог реагировать. Источники неопределённости должны быть явными. Недостоверная видимость точности может быть вреднее честно обозначенной неизвестности: оценка числа затронутых лиц без методики непригодна для контролирующего органа.

Второй этап — обновление статуса после первичного контроля ущерба и сбора существенных артефактов. Третий — итоговая фиксация коренной причины, масштаба, корректирующих мер и мероприятий для недопущения повторения. Точный состав и сроки каждого сообщения определяются применимым режимом, а не внутренней универсальной формой. Для каждой коррекции создают версию документа с журналом изменений.

В информационном взаимодействии не следует смешивать раскрытие по обязанности и публичную коммуникацию. Регулятору может требоваться персонализированная информация по защищённому каналу; публично допустимо раскрыть только обезличенные выводы. В интересах затронутых лиц важно избегать преждевременных заявлений «проблема полностью устранена», когда remediation ещё не прошла независимое повторное тестирование.

21.1. Принцип правдивой неполноты

Если причинность ещё не установлена, формула «связь системы с событием исследуется; обнаружены такие-то признаки» обычно точнее, чем «ошибку вызвал AI». При отправке первого сообщения отдельно перечисляют известное, предположительное и пока недоступное. Это сохраняет доверие к отчётности и уменьшает потребность ретроспективно оправдывать ранние гипотезы.

21.2. Структура версий сообщения и регуляторная прозрачность

Первая версия сообщения должна отражать ровно известное на момент отправки. При получении новых данных сохраняются и первоначальный документ, и дополнение с перечнем новых фактов, исправлений, заменённых оценок и всё ещё неизвестных обстоятельств. Важнейшее свойство такого пакета — отсутствие «исчезающих выводов»: если первоначально предполагалась утечка, а позднее установлено, что данные не покинули среду, это изменение необходимо объяснить доказательно. Замена содержания без версионирования подрывает доверие к организации и делает аудит срока практически невозможным.

Каждый орган может требовать свою форму и канал подачи. Универсальное внутреннее досье служит источником данных для разных сообщений, но не заменяет специального регуляторного формата. До направления проверяют тип инцидента, субъект, компетентный адресат, точные события и допустимость раскрытия конфиденциальных сведений. Подтверждение отправки хранится в том же реестре, что и версия содержания.

Практический разбор 19.21

Рекомендация: Сообщать проверенные сведения с версиями

Условный пример: В первоначальном сообщении подозревали модель, затем обнаружили ошибку исходника.

Разбор: Непомеченная замена текста скрывает развитие знания.

Практические действия: Координатор направляет уточнение с причиной изменения и сохраняет обе версии.

Критерий результата: Получатель может различить первоначальное предположение и уточнённый факт.

22. Территориальный охват, конфликт правил и трансграничная передача

Одна AI-система может обслуживаться американским поставщиком, обрабатываться на серверах в ЕС, использоваться финансовой организацией во Франции и причинить вред клиенту за пределами Европы. Нельзя автоматически выбирать право по местонахождению облачного сервера. Следует последовательно определить применимость территориальных норм, договорные отношения, местоположение последствий, статус затронутого лица и компетенцию органа.

В трансграничном деле особенно важны запреты на несанкционированную передачу персональных, медицинских и правоохранительных данных. Доказательственная копия не может свободно передаваться любому привлечённому консультанту только потому, что она относится к расследованию. Сначала определяют правовую основу доступа, пределы полномочий, защитные гарантии и потребность в псевдонимизации. Для полицейских данных может действовать отдельный режим, не тождественный общему GDPR.

В случае несовпадения требований к срокам и объёму раскрытия юридический координатор готовит карту, позволяющую соблюсти каждую применимую обязанность отдельно. Если возникает неразрешимая коллизия, требуется локальная правовая оценка и при необходимости обращение в компетентный орган. Настоящий доклад не утверждает, что международного «одного окна» AI-инцидентов уже существует.

22.1. Доступ эксперта из другой страны к доказательствам

Международный характер цифровой инфраструктуры создаёт ошибочную иллюзию, что все материалы доступны эксперту независимо от географии. На практике запись может храниться на сервере в одной стране, содержать данные лиц из другой, а принадлежать контролёру из третьей. Прежде чем копировать весь набор в общий архив, требуется установить договорный режим обработки, правовое основание раскрытия, минимальность передаваемых данных и ограничения профессиональной тайны. Иногда оптимальным решением является проверка в защищённой среде владельца с выдачей обезличенного протокола вместо массовой передачи исходных данных.

Разграничиваются трансграничная передача персональных данных, использование облачной услуги, доступ иностранного специалиста из-за пределов ЕЭЗ и процессуальное раскрытие по судебному приказу. Наличие иностранной юрисдикции не запрещает автоматически проверку, но требует осознанной схемы законного доступа. В экспертном задании следует указать, какие элементы могут быть скопированы, кто вправе их читать, можно ли делать derivative data, как фиксируются экраны и команды исследователя и что подлежит уничтожению по завершении законного срока хранения.

В трансграничном споре одной стороне зачастую известен только итоговый отказ банка или пограничного органа. Эксперт должен обозначить, какие факты подтверждаются доступными документами, а какие могут быть установлены только законным ответом компетентной структуры. Нельзя изображать наличие доступа к закрытым системам INTERPOL, правоохранительных органов или банков без соответствующего подтверждения.

Практический разбор 19.22

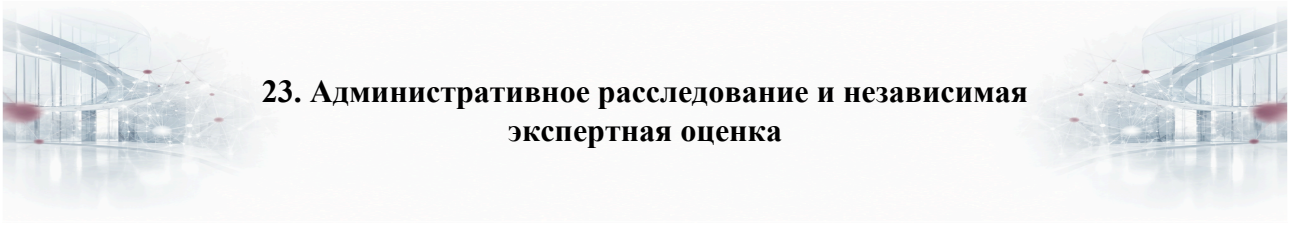
Рекомендация: Проверять основание каждой передачи

Условный пример: Зарубежный эксперт просит полный архив клиента для анализа одной записи.

Разбор: Передача всего архива может быть избыточной.

Практические действия: Ответственный определяет нужные фрагменты, допустимый канал и ограничения; юрист проверяет основание.

Критерий результата: Передан минимальный согласованный пакет с журналом получателя.



23. Административное расследование и независимая экспертная оценка

Административный надзор и независимый аудит имеют разные цели и полномочия. Орган рыночного надзора может требовать информацию, исследовать соответствие и применять предусмотренные законом меры; частный эксперт проводит оценку только в пределах доступа и поручения. Эксперт не должен выдавать свою просьбу о документах за обязательное государственное распоряжение, а отрицательный отзыв частной организации за юридически окончательное решение об ответственности.

При внешней оценке стороны должны заранее согласовать предмет, критерии, доступ, сохранение доказательств, ограничения публичности и правила работы с несогласными выводами. Право заказчика комментировать фактические ошибки не должно превращаться в право удалить неудобный вывод. Если эксперт получает вознаграждение, связанное с благоприятным заключением, конфликт интересов должен быть раскрыт; в чувствительных проектах такая модель нежелательна.

Отдельная проблема — доказательства, доступ к которым предоставлен только через демонстрацию интерфейса поставщиком. Эксперт должен обозначить, что наблюдал управляемую демонстрацию, а не самостоятельно воспроизвёл рабочую систему. При невозможности глубокой проверки вывод формулируется как оценка ограниченного набора свидетельств.

Вид проверки	Полномочие	Стандарт осторожности
Внутренняя RCA	Организационный доступ	Независимое подтверждение
Частная экспертиза	Договор и согласие	Раскрытие ограничений
Регуляторная проверка	Применимое право	Компетенция и процедура
Судебная экспертиза	Процессуальное назначение	Контроль допустимости

23.1. Пределы мандата независимого исследователя

Независимый отчёт может быть подготовлен по поручению стороны спора, организации, профессионального объединения либо компетентного органа. В каждом случае нужно раскрыть основание доступа и объём полномочий. Частный эксперт не получает автоматически права изымать сведения, обязывать иностранный банк раскрыть данные или проводить процессуально значимые действия вместо следователя. Отсутствие таких полномочий не уничтожает ценность исследования, но определяет предел допустимых выводов. Запросы к третьим лицам должны быть корректно оформлены и сопровождаться подтверждением согласия либо иного законного основания.

Профессиональная независимость не означает отсутствие заказчика. Она проявляется в раскрытии мандата, отказе от заранее предрешённого результата, документировании негативных находок и проверке альтернатив. Если специалист ранее участвовал в разработке проверяемой системы, его знания могут быть полезны для объяснения архитектуры, но вывод о независимости требует отдельного управления конфликтом интересов.

Практический разбор 19.23

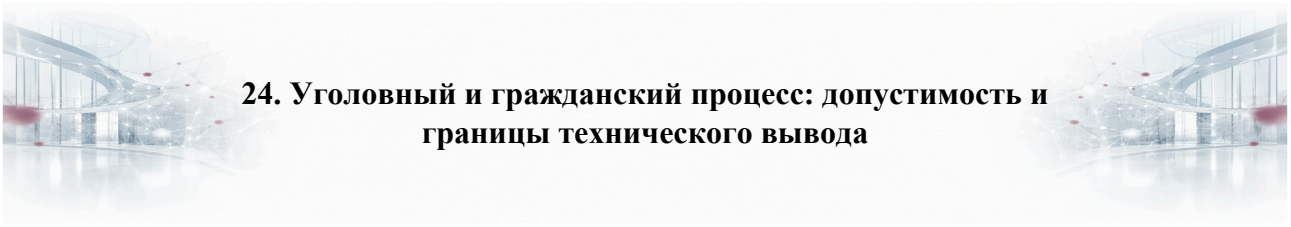
Рекомендация: Сохранять независимость экспертного исследования

Условный пример: Заказчик просит эксперта подтвердить заранее выбранную версию причины.

Разбор: Поставленный желаемый результат искажает исследование альтернатив.

Практические действия: Эксперт согласует нейтральные вопросы, доступ и право сообщить недостаточность данных.

Критерий результата: Заключение отвечает на вопросы по материалам без заданного виновника.



24. Уголовный и гражданский процесс: допустимость и границы технического вывода

Если результаты AI повлияли на задержание, уголовное обвинение, оценку доказательства или отказ в экстрадиции, расследование должно отдельно оценивать процессуальное значение источника. Машинная рекомендация, созданная без прозрачной методологии и доступа к исходным данным, не превращается в судебно допустимое доказательство только потому, что её распечатали на бланке органа. Необходимо проверить процессуальный статус, происхождение, надежность, возможность оспаривания и доступ защиты к релевантной информации.

В гражданском деле могут отличаться правила распределения бремени доказывания, раскрытия документов и использования экспертных заключений. Данный доклад не устанавливает единый международный стандарт допустимости для всех юрисдикций. Он предлагает доказательственную структуру, совместимую с задачами последующей локальной экспертизы: независимое получение данных, сохранённая цепочка хранения, точное описание методов и прозрачный перечень альтернативных причин.

Существование коммерческой тайны не следует воспринимать как абсолютную преграду проверке, но и требование предоставить полный исходный код всегда не пропорционально. Возможны защищённая экспертиза, судебно контролируемый доступ, проверка конкретного компонента, экспертиза output и независимая оценка процедур. Эти решения зависят от национального права и полномочий суда.

24.1. Вопросы к эксперту

Какие факты установлены непосредственно, какие основаны на предоставленных стороной материалах, какие тесты можно повторить, какова погрешность и какие выводы не допускаются полученными данными? Ответы на эти вопросы должны быть видимы в самом заключении, а не оставаться в переписке команды.

24.2. Отчёт эксперта и процессуальный стандарт доказательства

Технический эксперт может установить, что определённый вывод модели отсутствует в первоисточнике, что лог содержит указанный hash, что конфигурация изменилась и что определённая последовательность запросов воспроизводится. Он не должен без

специального правового мандата квалифицировать это как преступление, устанавливать заведомость действий руководства или объявлять недействительным решение государственного органа. Выводы о допустимости, относимости и достаточности доказательств подчиняются процессуальному праву конкретной юрисдикции и оценке компетентного лица.

Для каждого существенного заключения полезна таблица: проверяемое утверждение, исходные данные, метод, альтернативные объяснения, степень неопределённости и вопрос, который остаётся за судом. Например, фраза «лог отображает срабатывание совпадения» может подтверждаться напрямую, а вывод «банк незаконно заморозил активы по причине работы ИИ» требует дополнительного анализа юридического основания и всей цепочки принятия решения. Разделение повышает пригодность заключения для судебной процедуры и защищает эксперта от преувеличения компетенции.

При подготовке материалов для сторон следует сохранять возможность проверки противоположной стороной в допустимых рамках. Конфиденциальная версия может содержать сведения об архитектуре и источниках, открытая — описывать проверенный вывод и его ограничения. Если публично раскрыть метод невозможно по соображениям безопасности или тайны, следует прямо указать это как ограничение внешней верифицируемости, а не ссылаться на «секретную проверку» как на бесспорный факт.

Практический разбор 19.24

Рекомендация: Разграничивать технический вывод и правовую ответственность

Условный пример: Повторный тест воспроизвёл сбой, сторона требует назвать виновное лицо.

Разбор: Воспроизведение не устанавливает все юридические элементы ответственности.

Практические действия: Эксперт описывает механизм и ограничения; правовую квалификацию передаёт компетентному специалисту.

Критерий результата: В заключении не смешаны наблюдение, причинный вывод и решение о вине.

25. Права человека, средство защиты и восстановление положения пострадавшего

Эффективное расследование должно привести не только к исправлению модели, но и к оценке прав пострадавшего. Если человек лишился доступа к банковским операциям, испытал ограничение свободы или был исключён из процедуры убежища, успешное обновление алгоритма само по себе не возмещает последствия. Необходимо обеспечить доступную информацию о решении, возможность подать возражения, независимый пересмотр, исправление данных и, при наличии оснований, эффективную компенсацию.

Права на справедливое разбирательство, уважение частной жизни и эффективное средство правовой защиты могут быть затронуты алгоритмическими операциями; квалификация требует анализа применимого права, роли органа и реального вмешательства. Отдельный акцент — уязвимость человека, который не знает, что автоматизированный показатель

участвовал в решении. Отсутствие возможности обнаружить AI-компонент снижает практическую эффективность оспаривания. Основанием для анализа в правопорядке ЕС служит также Хартия основных прав ЕС, в частности гарантии эффективного средства правовой защиты, уважения частной жизни и защиты персональных данных. [17]

Предлагаемая модель ARGA требует два реестра: технических исправлений и индивидуальных восстановительных мер. В первом фиксируют новый порог, тесты и deployment; во втором — отмену ошибочного результата, уведомление третьих лиц, исправление корреспондирующих баз и проверку, что вредное решение не воспроизводится в последующих системах. Только вместе они позволяют обоснованно закрывать дело.

Мера	Вопрос для подтверждения	Риск фиктивного исполнения
Исправление базы	Все копии исправлены?	Ошибка осталось в downstream
Пересмотр решения	Проводился независимым лицом?	Автоматическое повторение отказа
Уведомление партнёров	Получено и принято?	Сообщение ушло не по адресу
Компенсация	Применимая правовая основа?	Формальная извинительная записка

25.1. Эффективное восстановление нарушенного положения как показатель завершения расследования

Профессиональная дисциплина расследования требует отдельно установить вред и отдельно проверить эффективность его устранения. Для ошибочной санкционной метки недостаточно снять блокировку в одном банке, если данные продолжают распространяться через общую систему контрагентских проверок. Для ошибочного результата, который повлиял на административную процедуру, недостаточно исправить внутренний score, если решение по делу не пересмотрено компетентным органом. В реестре CAPA должны существовать задания по каждому downstream-получателю, а их исполнение подтверждаться самостоятельными доказательствами.

Программу восстановления следует строить исходя из юридически доступных средств. Она может включать исправление фактических данных, рассмотрение возражений, независимый человеческий пересмотр, извещение получателей, разблокирование документации, устранение незаконно созданной записи и, если применимо, рассмотрение вопроса о возмещении ущерба. Организация не должна обещать пострадавшему результат, который может обеспечить только суд, регулятор или иной независимый орган. При наличии конфликта между раскрытием технических деталей и защитой персональных данных применяют пропорциональный доступ с сохранением возможности эффективного оспаривания.

Критерий закрытия инцидента рекомендуется формулировать как набор доказуемых состояний: прекращён текущий вред, приняты предупредительные меры, внешние обязанности исполнены или обоснованно не возникли, затронутому лицу предоставлен доступный способ защиты и организован контроль повторения. Закрытие инженерного тикета не равно юридическому или общественному устранению последствий.

Практический разбор 19.25

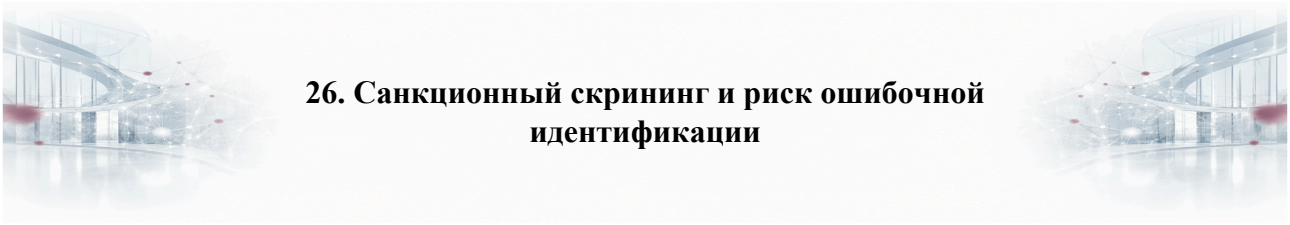
Рекомендация: Обеспечивать исправление для затронутого лица

Условный пример: Ошибка устранена в модели, но клиент продолжает получать отказ по старой записи.

Разбор: Техническое исправление не восстановило положение клиента.

Практические действия: Координатор проверяет записи у получателей и доступный пересмотр конкретного решения.

Критерий результата: Есть подтверждение исправления последствий либо мотивированное описание препятствия.



26. Санкционный скрининг и риск ошибочной идентификации

В санкционных системах AI чаще всего не принимает юридически окончательное решение, но формирует сигнал, который может оказать сильное влияние на комплаенс-сотрудника или алгоритм блокировки. Расследование ошибки начинается с выяснения, что именно было сопоставлено: полное имя, транслитерация, дата рождения, гражданство, адрес, корпоративная роль, бенефициарная связь или внешнее предположение об аффилированности. При отсутствии дополнительных идентификаторов строковое совпадение может быть технически истинным и юридически недостаточным.

Следует различать два вреда: необоснованное ограничение операции конкретного клиента и систематический перекос, когда определённые транслитерации, страны происхождения или длинные имена порождают чрезмерные ложные совпадения. В первом случае проверяют chain of decision и возможность исправления; во втором — распределение ошибок по обоснованно выбранным группам при соблюдении принципов защиты данных. Необходима осторожность с чувствительными атрибутами: статистика для аудита не даёт общего разрешения собирать их без правового основания.

Отчёт должен устанавливать, обновлялся ли исходный санкционный список, кто подтвердил совпадение, была ли передана информация в другие финансовые учреждения и какие последствия сохранились после исправления. Важнейшая рекомендация — выстраивать downstream correction: если база клиента исправлена в одном банке, но ошибочное имя осталось в контрагентском провайдере, инцидент не закрыт.

26.1. Контрольные сопоставления

Признак	Возможная причина	Обязательный для проверки материал
Только полное имя	Коллизия имён	Матч-логика и порог
Неоднозначная транслитерация	Нормализация алфавитов	Таблица преобразований
Ложная корпоративная связь	Устаревший ownership graph	Версия графа и источники
Блокировка после ручного «OK»	Автоматический workflow	Audit trail сотрудника

26.2. Иллюстративная архитектура санкционного ложноположительного совпадения

Предположим учебную ситуацию: финансовая организация получила автоматическое совпадение по транслитерации фамилии, после чего операционный модуль без независимой проверки остановил платежи. Исследование разбивается на четыре цепочки. В первой устанавливают точную запись публичного санкционного списка и дату его обновления. Во

второй восстанавливают правила сопоставления, дополнительные идентификаторы и порог confidence. В третьей проверяют, существовал ли настоящий человеческий контроль и какие данные были видны сотруднику. В четвёртой устанавливают юридическое основание остановки операции, возможность ручного рассмотрения и действия по восстановлению прав клиента. Даже если алгоритм ошибся, законность банковской меры не выводится автоматически из одного технического факта.

Задача эксперта — отличить совпадение имени от тождества человека, различить источник нормативной записи и частный агрегатор и доказать, где именно ошибка была размножена. Особое внимание уделяется компании-посреднику, которая могла распространить устаревший ownership-graph на другие учреждения. В отчёте не рекомендуется публиковать полные идентификаторы третьих лиц, чьи имена совпали в списках. Для клиента важнее документированная коррекция и механизм предотвращения рецидива, чем публичная демонстрация всех совпадений.

Практический разбор 19.26

Рекомендация: Проверять идентичность по различающим данным

Условный пример: Совпали имя и фамилия, но даты рождения отличаются.

Разбор: Автоматический флаг не подтверждает тождество лиц.

Практические действия: Специалист сохраняет оба источника и проверяет дополнительные идентификаторы до окончательной квалификации.

Критерий результата: Решение мотивировано совокупностью данных, ошибочное объединение отменено.



27. Международное полицейское сотрудничество, розыск и экстрадиция

Исследование AI-инцидента в международном розыске не должно исходить из предположения, что INTERPOL автоматически принимает решения на основе определённой модели. Применимость алгоритмической обработки и конкретная архитектура системы устанавливаются только доказательствами. Однако общий риск заключается в передаче неточного машинного совпадения между юрисдикциями, где последующее решение о задержании принимает отдельный национальный орган.

В расследовании разбирают происхождение полицейского сообщения, статус информации, историю обновлений, наличие запроса об исправлении и фактическое использование данных государством. Предмет отдельной проверки — могло ли машинное сообщение быть воспринято как официально подтверждённое основание, хотя содержало ограниченную гипотезу. Техническое совпадение не отменяет требований необходимости, точности, компетенции и процессуальных гарантий.

Правовая стратегия защиты может включать оспаривание национального решения, запрос на уточнение источника, использование действующих механизмов коррекции полицейских данных и сохранение доказательств о трансграничном распространении ошибки.

Утверждение, что один орган способен одним решением отменить все записи всех государств, было бы неверным. Независимая AI-экспертиза здесь помогает установить происхождение ошибки, но не подменяет специальных процедур национального и международного права.

27.1. Перечень доказательств для защиты

Нужны оригиналы уведомлений, документы о задержании или отказе во въезде, указанные идентификаторы, официальный ответ об исправлении, логи внутреннего доступа (если законно доступны), временная шкала и сведения об органе, принявшем решение. При отсутствии доступа к закрытым системам экспертиза прямо указывает на процессуальные ограничения.

27.2. Когда техническое совпадение взаимодействует с полицейской записью

В делах международного розыска техническое обнаружение сходства личности и юридическая обязанность органа осуществить проверку не эквивалентны. Доклад должен разделять наличие доступных подтверждённых сообщений, данные национального правоохранительного органа, решение о задержании, судебную оценку и предположительный вклад алгоритмического инструмента. Из того, что человека остановили на границе, невозможно надёжно заключить, какой именно тип международного уведомления был активен или какое программное обеспечение использовалось. Доступ к ограниченным материалам возможен лишь через надлежащие процедуры, и независимый исследователь не подменяет официальные ответы.

Доказательственный пакет в учебном кейсе включает документы о личности, записи о пограничном действии и его времени, выданные процессуальные решения, известные коммуникации компетентных органов и законно полученные сведения об источнике совпадения. Техническую гипотезу можно проверить лишь при доступе к правилам сопоставления либо проверенным процессуальным материалам. Отдельно рассматриваются омонимия, транслитерация, устаревшая запись, ошибка маршрутизации и неверное толкование данных человеком. Последствия для свободы человека требуют незамедлительного применения доступных правовых средств, даже когда цифровая форензика ещё не закончена.

Практический разбор 19.27

Рекомендация: Не подменять официальный розыск алгоритмической справкой

Условный пример: Помощник сформировал сводку из нескольких неофициальных публикаций.

Разбор: Она не подтверждает существование и содержание официальной записи.

Практические действия: Юрист разделяет публикации и официальные материалы, запрашивает надлежащее подтверждение по процедуре.

Критерий результата: В позиции указан действительный статус каждого источника.

28. AML/KYC: автоматические подозрения и последствия для клиента

AML-модель способна правомерно выявлять сомнительные закономерности и одновременно совершать систематические ошибки. В делах об отказе в обслуживании или чрезмерной заморозке средств нужно различать обязанность организации соблюдать AML-правила, запрет на раскрытие отдельных сведений о сообщениях об операциях и право клиента на справедливую, недискриминационную и соразмерную процедуру в применимом праве.

Установление ошибки требует анализа версии модели, списка признаков, source data, правил изменения risk score, действий аналитика, времени принятия решения и юридической меры. Не следует требовать публичного раскрытия деталей, которые закон специально защищает; возможна закрытая независимая оценка с ограниченным кругом допущенных экспертов. Вопросы обоснованности решения и возможности его обжалования оцениваются в рамках соответствующего государства.

Особенно важна проверка массовых систематических отказов, которые не отражаются в одном «крупном» инциденте: сотни небольших ошибочных блокировок могут затронуть больше людей, чем одно заметное происшествие. Отдельный показатель должен описывать частоту повторения по идентичным условиям и сроки восстановления обслуживания после обоснованного возражения.

28.1. Массовая AML-эскалация как проверяемый управленческий сбой

Когда после обновления модели подозрительными внезапно оказываются тысячи операций, независимый анализ должен учитывать не только метрики точности, но и реальное распределение труда. Следует установить изменившуюся базовую частоту сигналов, долю повторных клиентов, изменение порогов и нагрузку на сотрудников, а затем выяснить, стали ли автоматические отказы заменять предусмотренный ручной пересмотр. Количество сигналов само по себе не равно количеству случаев отмывания средств или доказанных нарушений клиентов.

В отдельной матрице фиксируются версия модели, период сравнения, критерии ground truth, число независимых операций и количество затронутых субъектов. Нельзя сравнивать новый поток с прошлым месяцем без поправки на сезонность, изменение клиентского портфеля и внешние изменения законодательства. Если ошибочная система вызвала долгосрочные закрытия счетов, необходимо проверить правовые основания и доступность процедуры оспаривания, а не ограничиваться возвратом уровня ложноположительных срабатываний к исходному значению. Корректирующая мера включает восстановление данных, независимую перепроверку особо тяжёлых решений и прозрачный процесс исправления системных ошибок.

Практический разбор 19.28

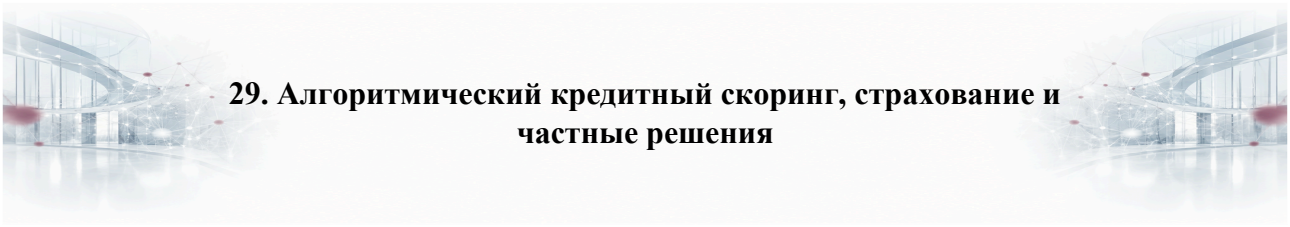
Рекомендация: Отделять подозрение от установленного нарушения

Условный пример: AML-модель выделила необычную операцию, сотрудник назвал её доказанным преступлением.

Разбор: Индикатор риска не устанавливает состав правонарушения.

Практические действия: Рецензент исправляет квалификацию, сохраняет исходные данные и проверяет основание последующего решения.

Критерий результата: Материал различает сигнал, проверенные факты и юридическую оценку.



29. Алгоритмический кредитный скоринг, страхование и частные решения

Дело о необоснованном отказе в кредите требует точного отделения оценки вероятности дефолта от окончательного решения банка. На исход влияют лимиты, доход, отраслевые правила, человеческое согласование, а иногда и внешние антифрод-проверки. Если организация использует стороннюю модель, в расследовании анализируют, какие факторы были переданы оператору и способен ли тот проверить ошибки или апелляцию.

Для оценки discriminatory impact нужно выбрать юридически допустимые группы сравнения, корректировать важные смешивающие факторы и не выдавать простое различие средних результатов за установленную дискриминацию. При подготовке доказательств желательно показать не только disparity, но и возможный причинный механизм: неточные proxy, исторический отбор данных, неправомерно использованные личные признаки или нестабильная логика по группам.

Технические метрики особенно ненадёжны без оценки отказов, которые не дошли до клиентского обращения. Реестр инцидентов должен учитывать обращения третьих лиц, повторные заявки и обратную связь от операционного подразделения. Если банк изменяет порог после жалобы, фиксируют предыдущую редакцию и независимый повторный прогон, иначе доказательство исчезает.

29.1. Автоматическое принятие решения и частный правовой эффект

В кредитовании и страховании сложность часто возникает на границе модели и бизнес-правила. Алгоритм может возвращать риск, а договорный или управленческий порог преобразует score в отказ. Для разбора необходимо получить значение и интерпретацию score, действующую на ту дату политику, полноту исходных данных и историю доступных клиенту возражений. Техническая объяснимость предсказания не всегда равна юридически достаточному объяснению решения. Если используется обработка персональных данных, отдельно анализируются применимые требования GDPR, включая при соответствующих условиях статью 22.

Ошибочный скоринг может иметь последствия за пределами одного продукта: запись о риске передаётся между подразделениями, влияет на иной счёт либо воспроизводится позднее. Поэтому эксперт должен проверить карту downstream-использования и доступность коррекции. Правовое заключение оценивает не «плохой score» в вакууме, а конкретное действие и его основание.

Практический разбор 19.29

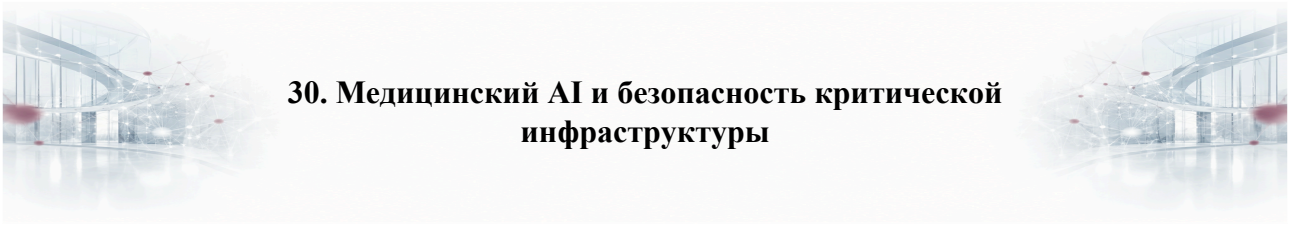
Рекомендация: Проверять вход скоринга и пересмотр

Условный пример: Отказ в кредите основан на устаревшей записи о доходе.

Разбор: Разбор весов модели не исправляет неверный вход.

Практические действия: Ответственный сверяет запись с документами клиента и организует повторное рассмотрение.

Критерий результата: Решение пересмотрено с исправленным входом; результат зафиксирован.



30. Медицинский AI и безопасность критической инфраструктуры

В медицинской практике расследование не должно ограничиваться анализом точности алгоритма. Необходимо выяснить клиническую ситуацию, роль врача, точность входных данных, ограничение показания, пользовательскую инструкцию и характеристики пациентов. Для медицинских изделий в ЕС существуют специальные секторные режимы, включая собственную процедуру vigilance. AI Act предусматривает координацию с ними и исключения из обычного режима отчётности статьи 73 в определённых обстоятельствах. [1]

В критической инфраструктуре приоритетом является безопасность физической системы и непрерывность жизненно важной услуги. Операционную модель нельзя произвольно отключить ради полного сохранения памяти, если это усугубит угрозу. Организация должна заранее разработать безопасный режим изоляции и альтернативный маршрут управления, чтобы сохранение доказательств и защита людей не оказались несовместимыми.

Особую опасность создают ошибки оценки собственной уверенности. Высокий показатель probability может представлять только внутреннюю калибровку классификатора и не гарантирует клинической достоверности вывода. Экспертное заключение обязательно раскрывает выборку, целевую популяцию, ограничения и то, какие выводы не переносимы на конкретного пациента.

Практический разбор 19.30

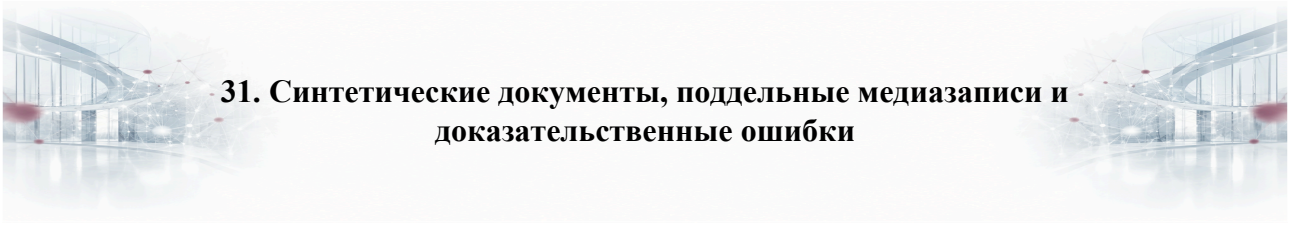
Рекомендация: Приоритизировать предотвращение продолжающегося вреда

Условный пример: В критическом процессе обнаружен воспроизводимый ошибочный выход.

Разбор: Длительное исследование не должно оставлять опасную функцию без контроля.

Практические действия: Оператор вводит безопасный ограниченный режим и сохраняет материалы; профильная команда исследует причину.

Критерий результата: Снижение текущего риска и расследование имеют отдельных ответственных.



31. Синтетические документы, поддельные медиазаписи и доказательственные ошибки

Генеративные модели могут создавать убедительные документы, звук и изображения. При расследовании нельзя объявлять материал поддельным лишь потому, что автоматический детектор deepfake дал высокую оценку. Чувствительность детектора зависит от формата, сжатия, новых методов генерации и частоты ложных срабатываний. Напротив, отсутствие следов синтеза не доказывает подлинность.

Техническая проверка предполагает получение оригинального файла, анализ доступных метаданных, историю передачи, независимые подтверждения времени и места, а при необходимости экспертные методы медиафорензики. В делах уголовного преследования принцип состязательности требует, чтобы сторона имела возможность оспорить метод, контрольную выборку и ограничение оценщика. Для документов, возможно, прошедших несколько конвертаций, отсутствие EXIF само по себе нормально и не равняется фальсификации.

Практическая схема включает три независимых ветви: происхождение носителя, согласованность внешних обстоятельств и техническая экспертиза содержания. Даже если один компонент не доступен, остальные могут иметь самостоятельную доказательственную ценность. Итоговое заключение не должно подменять вероятностный технический вывод категорическим утверждением об авторстве.

31.1. Аутентичность синтетических файлов и цепочка трансформаций

В делах о подозрительных аудио-, видео- и текстовых материалах нельзя проводить экспертизу только на основании вероятности, выданной коммерческим «детектором дипфейков». Требуются сведения о происхождении файла, извлечении с устройства, контрольных суммах, истории передачи, кодировании и возможных преобразованиях платформой. Даже подтверждённое использование синтетического инструмента не означает автоматически, что отдельный факт в записи ложен. И наоборот, документ без признаков синтеза может содержать реальную сцену, снятую в неверном контексте или смонтированную с искажением хронологии.

При сравнении записи со следственными и судебными материалами эксперт составляет отдельный список утверждений: что видно или слышно непосредственно, что предполагается из контекста и какие элементы подтверждены независимыми источниками. Неизвестная степень достоверности детектора указывается явно; результаты разных инструментов не следует механически усреднять. Для важных решений рекомендуются оригинальный носитель, независимый медиаэксперт и процессуально допустимые доказательства альтернативного происхождения. Служебная метка «AI generated» сама по себе не является судебным выводом о подделке.

Практический разбор 19.31

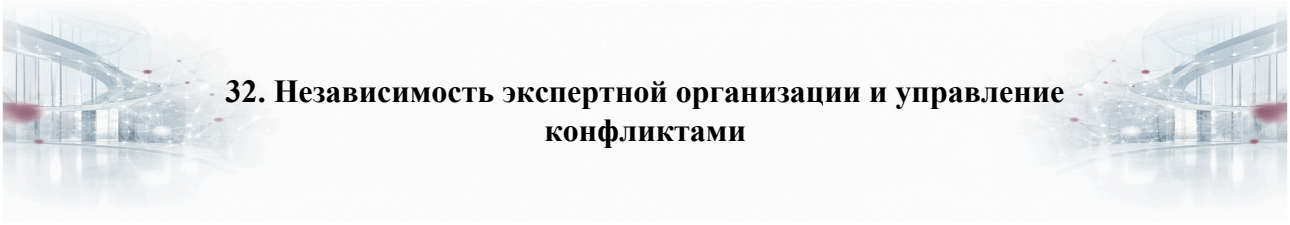
Рекомендация: Проверять происхождение спорного медиафайла

Условный пример: Запись выглядит убедительно, но предоставлена только пересланная копия.

Разбор: Внешний вид не подтверждает подлинность и автора.

Практические действия: Эксперт запрашивает оригинал и происхождение, фиксирует преобразования; при нехватке данных ограничивает вывод.

Критерий результата: Степень уверенности соответствует доступным первичным сведениям.



32. Независимость экспертной организации и управление конфликтами

Разработчик AI заинтересован в демонстрации ограниченности инцидента; пострадавшая сторона — в подтверждении заявленного вреда; государственный орган — в защите легитимности собственных решений. Независимый исследователь не может игнорировать эти стимулы. Перед назначением экспертов оформляют реестр интересов: коммерческие отношения, участие в разработке системы, вознаграждение, предоставление вычислительных ресурсов, возможность публикации и связи с заинтересованными институциями.

Степень независимости может быть полной, ограниченной или недостаточной для определённого вопроса. Например, эксперт, ранее проводивший тесты модели, может объяснять собственную методику, но его независимость при оценке качества этих тестов ограничена. Рабочая группа должна иметь право документировать dissenting opinion, сохранять оригиналы материалов и обозначать препятствия доступу, даже если вывод неудобен заказчику.

Практически предлагается двухуровневая проверка: первый эксперт воспроизводит технический сценарий, второй оценивает соответствие процесса доказательственным критериям. Особенно полезна рецензия специалиста по правам человека, когда формально корректная статистика скрывает тяжёлое индивидуальное последствие.

32.1. Запрет на ложную институциональную атрибуцию

Вывод Observatoire ARGA должен распространяться как экспертное исследование частной организации, если иное не подтверждено отдельным официальным актом. Нельзя обозначать документ как решение Еврокомиссии, ООН, INTERPOL, национального суда или сертификационного органа. Направление мнения на консультацию не означает, что регулятор признал выводы.

32.1. Пределы институционального имени Observatoire ARGA

В публичных материалах важно не выдавать независимую исследовательскую организацию за орган власти, официальную инспекцию или аккредитованную государством лабораторию без соответствующего документального статуса. Наличие экспертного опыта, участия в международных консультациях и аналитических публикаций не тождественно процессуальному назначению экспертом по конкретному делу. Для каждого документа указываются авторы, технический мандат, раскрытые ограничения, характер рецензии и принадлежность окончательных юридических выводов компетентному специалисту.

При представлении методических предложений другим учреждениям корректно говорить о потенциальном стандарте сопоставимости, а не об «обязательных международных правилах ARGA». Это повышает профессиональное доверие и позволяет использовать материалы как аналитическую основу при разработке внутренних и публичных процедур.

Практический разбор 19.32

Рекомендация: Раскрывать конфликт и назначать независимого рецензента

Условный пример: Эксперт ранее консультировал поставщика по спорному компоненту.

Разбор: Возникает риск оценки собственной работы.

Практические действия: Координатор документирует отношения и передаёт соответствующий вопрос независимому специалисту.

Критерий результата: Решение по компоненту не утверждено его прежним консультантом.

33. Защита источников, коммерческой тайны и персональных данных

Важные доказательства могут включать медицинские карты, сведения об уголовном преследовании, внутренние модели антифрода, персональные переписки и уязвимости критической инфраструктуры. Расследование должно строиться по принципам необходимости, минимизации и разграничения доступа. Нельзя формировать «полный экспертный архив» из чужих сведений без правового основания и определённого срока хранения.

Рекомендуется вести три уровня доступа: рабочий, ограниченный экспертный и публичный. В рабочем содержатся описания без лишних идентификаторов, в защищённом — оригиналы и доступно проверяемые артефакты, в публичном — только редактированные факты, которые не позволяют воссоздать опасную атаку или определить пострадавшего вопреки его правам. Необходимо сохранять доказательство самой редактировки: хэш оригинала, журнал преобразования, перечень удалённых полей и полномочие лица.

Особая проблема — участие внешних AI-сервисов в подготовке самого расследования. Передача полного закрытого досье в генеративный инструмент требует самостоятельной оценки правового основания, договорных гарантий, сохранения данных и конфиденциальности; модель может ошибочно суммировать материалы и не должна быть единственным экспертом или хранителем доказательств.

Практический разбор 19.33

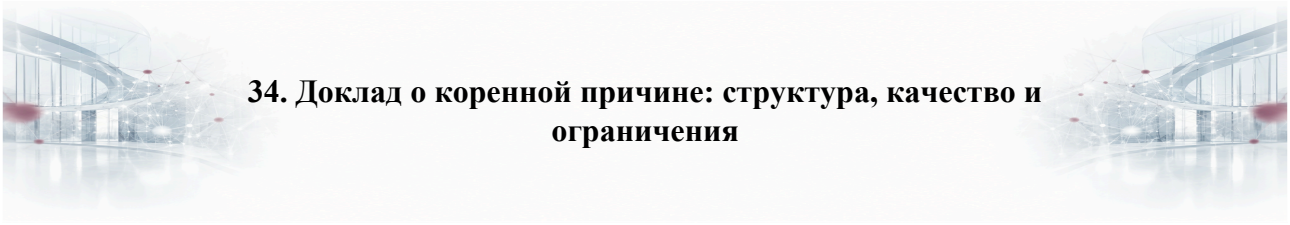
Рекомендация: Защищать источники при передаче результата

Условный пример: Публичный отчёт содержит детали, позволяющие узнать заявителя.

Разбор: Удаление имени не устраняет косвенную идентификацию.

Практические действия: Редактор проверяет сочетание дат и обстоятельств, готовит закрытую доказательственную версию.

Критерий результата: Открытый текст сохраняет вывод без ненужного раскрытия личности.



34. Доклад о коренной причине: структура, качество и ограничения

Основной отчёт строится вокруг исследовательских вопросов. Для каждого приводятся установленное наблюдение, первичный источник, выполненная проверка, существенная альтернатива и предел вывода. После этого описываются связанные корректирующие действия и критерии повторной проверки. Такая структура позволяет увидеть, какие меры следуют из подтверждённого механизма, а какие предложены для уменьшения неопределённости или предотвращения сходного сценария.

Для каждого вывода применяется маркировка: F — установленный факт; E — экспертная оценка; H — рабочая гипотеза; U — неизвестно. Цифровой идентификатор источника связывает текст с evidence register. Доклад должен позволять стороннему проверяющему реконструировать ход рассуждения без личного знакомства с автором. Если внешний контрагент не предоставил часть данных, сведения об этом фиксируются в перечне неполученных доказательств.

Коренные причины следует разделять на ближайшие и системные. Ближайшая причина может заключаться в неверном ответе конкретной модели; системная — в отсутствии политики тестирования, неизменяемых логов или возможности ручной отмены. Для предотвращения повторения одних технических патчей недостаточно. Обе категории описываются раздельно.

Раздел RCA	Содержание	Контроль качества
Установленные факты	Только F + source ID	Нет неподтверждённых цитат
Гипотезы	H1, H2, H3 и основания	Учтены альтернативы
Вывод о механизме	Связь действий и последствий	Пропущенные звенья раскрыты
Ограничения	Доступ, выборки, конфликты	Не скрыты за осторожной фразой
CAPA	Действие, владелец, тест	Результат измерим

34.1. Защита отчёта от кругового подтверждения

Независимость подтверждения оценивается по происхождению данных. Журнал модели, автоматически составленное по нему резюме и экспертный текст, использующий только это резюме, остаются одной информационной цепочкой. Для ключевого вывода ищут материал, полученный через иной механизм: запись внешней операции, исходный документ, независимую временную отметку или свидетельство с установленными ограничениями. При отсутствии такого материала зависимость источников явно ограничивает уверенность.

Предлагается раздельно фиксировать факты, результаты технической проверки, причинные интерпретации и правовые выводы. Термин «подтверждено» сопровождается основанием: подтверждено происхождением записи, воспроизводящим тестом либо совокупностью согласованных материалов. Неустранённое противоречие показывается рядом с выводом. Рекомендации не повышают степень его доказанности; они определяют действие при имеющихся данных.

Уровень	Допустимая формулировка	Нежелательная подмена
Факт	«В журнале имеется запись со временем... »	«Система объективно сделала...»
Тест	«В указанной среде результат воспроизведён...»	«Так происходило во всех версиях...»
Гипотеза	«Данные согласуются с версией...»	«Виновный установлен...»
Юридический вывод	«Применимость режима требует проверки...»	«Любой инцидент автоматически нарушает закон...»

Практический разбор 19.34

Рекомендация: Строить отчёт вокруг проверенной причинной цепочки

Условный пример: Заключение перечисляет документы, но не объясняет переход к выводу.

Разбор: Перечень источников не показывает, какая гипотеза подтверждена.

Практические действия: Автор связывает каждый вывод с наблюдением и альтернативой; отдельно перечисляет ограничения.

Критерий результата: Рецензент воспроизводит цепочку аргумента по материалам.

35. Корректирующие меры, повторное испытание и возобновление эксплуатации

Корректирующая мера должна быть связана с подтверждённым механизмом или с явно обозначенной предосторожностью. Если неверный идентификатор поступил из источника, изменение только порога модели не исправляет источник. Если корректный сигнал превратился в автоматическую блокировку, требуется проверить правило исполнения. Для каждой меры задаются изменяемый компонент, ожидаемое наблюдение, независимый критерий успешности и возможный побочный эффект.

Проверка исправления включает исходный сценарий, новый набор сходных случаев и ситуации, где мера способна ухудшить результат. Повторение только известного примера подтверждает устранение воспроизводимого дефекта на нём, но не перенос на другие случаи. Для изменения порога сопоставляются ложные сигналы и пропуски; для нового подтверждения человеком — своевременность и фактическая возможность отказа. Данные настройки отделяются от итоговой проверки.

Go/no-go решение оформляется компетентным органом управления и включает открытые остаточные риски. Формула «критических ошибок не найдено» не означает отсутствия риска; решение должно опираться на экспозицию, целевую область применения, неопределённость выборки и меры контроля в дальнейшем. В случае угрозы фундаментальным правам может потребоваться отдельная правовая оценка, независимая от коммерческой необходимости запустить сервис.

35.1. Проверка результативности исправлений и запрет косметической CAPA

Результативность мер оценивают по механизму и последствиям. Технический критерий отвечает на вопрос, перестал ли воспроизводиться дефект при заданных условиях. Процессный — исполняется ли новая проверка в рабочем потоке. Критерий восстановления

— прекращено ли неблагоприятное действие и исправлены ли связанные записи у установленных получателей. Выполнение первого критерия не означает автоматического выполнения остальных.

Решение о возобновлении указывает проверенную конфигурацию, выполненные критерии, оставшиеся ограничения и сигнал для повторного вмешательства. Испытатель по возможности отделён от внедрившей исправление команды. При недоступности независимой проверки это отражается как ограничение. Наблюдение после возобновления планируется по риску и доступной экспозиции; единичный благоприятный случай не устанавливает устойчивость эффекта.

Практический разбор 19.35

Рекомендация: Подтверждать корректирующую меру повторным тестом

Условный пример: Поставщик сообщает, что ошибка исправлена, но предъявляет только новую инструкцию.

Разбор: Заявление об изменении не подтверждает работоспособность.

Практические действия: Инженер повторяет исходный сценарий и близкие контрольные задачи; оператор проверяет процесс утверждения.

Критерий результата: Возобновление допускается после документированного результата проверки.

36. Учёт потерь, компенсации и восстановление данных третьих лиц

Последствия AI-инцидента часто распределены между прямыми и косвенными потерями: отмена операции, стоимость кредита, юридические расходы, репутационный вред, ухудшение здоровья, ограничение свободы и нарушение процессуального положения. Эксперт по AI не вправе самостоятельно присваивать денежный эквивалент всем последствиям без юридической и экономической оценки. В отчёте о вреде следует разграничить документированные убытки, обоснованную оценку и спорные требования.

Для организации полезно фиксировать стоимость исправления собственных систем отдельно от возмещения потерпевшим. Если remediation учтена как «расход на инцидент», это не показывает, получили ли люди реальное средство защиты. При трансграничном распространении ошибочных данных важно подтвердить, что сведения действительно исправлены у каждого получателя и повторное использование прекращено.

Показатели закрытия: доля подтверждённых ошибок, устранённых в срок; доля downstream-получателей, подтвердивших исправление; доля обращений людей, рассмотренных вне AI-конвейера; число повторных необоснованных отказов; остаточные открытые требования. Эти показатели относятся к предлагаемой управленческой методике, не являются автоматическими юридическими критериями.

36.1. Разделение материального и нематериального последствий

Расследование не должно автоматически сводить вред к сумме прямых финансовых потерь. Задержание при неверной идентификации, ограничение доступа к медицинскому обслуживанию, утрата репутации и невозможность оспорить автоматическое решение могут иметь существенные последствия без сразу измеримой денежной стоимости. Реестр фиксирует вид и продолжительность воздействия, документально подтверждённые расходы, ограничения свободы, административные процедуры и меры восстановления. Юридическая оценка компенсации требует указания применимого материального и процессуального права.

Для сравнимости полезно отделять наступивший вред, подтверждённый документами, от вероятного будущего вреда и от затрат на предотвращение повторения. Расходы компании на собственный аудит не являются автоматически компенсацией пострадавшему. Исполнение мер восстановления требует доказательств именно для соответствующего лица.

Практический разбор 19.36

Рекомендация: Учитывать прямые и последующие последствия

Условный пример: Неверная запись вызвала отказ, а её копия осталась у партнёра.

Разбор: Исправление локальной базы не устраняет вторичный эффект.

Практические действия: Координатор составляет карту получателей и подтверждений корректировки; потери фиксирует по документам.

Критерий результата: Проверены все известные получатели; спорные суммы не выданы за установленные.

37. Публичный реестр инцидентов и границы открытости

Публичная база нужна для накопления знаний, но её структура должна учитывать неполноту сообщений, законодательные запреты на раскрытие и риск повторной идентификации жертв. OECD AIM демонстрирует возможность международной классификации и связывания новостных сообщений с событиями; в то же время его методика не превращает СМИ в полный официальный перечень всех инцидентов. Разработка публичного реестра на базе одной категории источников требует явного указания sampling bias. [4]

Предлагается вести два связанных набора: закрытый evidence register с полными оригиналами и публичный incident abstract с анонимизированными фактами. Публичная версия содержит тип системы, класс угрозы, сведения об оценке источника, тип последствий, стадию расследования и статус исправления. Не публикуются номера дел, персональные данные, токены доступа или технические инструкции, способные облегчить злоупотребление.

Реестр не должен создавать впечатление, что включённая публикация доказывает юридическую виновность. Для каждой записи необходимо показывать уровень верификации, дату последнего изменения, тип источника, ссылку на публичный материал, а при опровержении — сохранённую запись о коррекции. Закрытие кейса означает достижение

критериев расследования, а не автоматическое отсутствие права человека продолжить обжалование.

37.1. Защита от статистических иллюзий публичной отчётности

Публичный реестр, построенный на добровольных сообщениях и новостях, неизбежно отражает не только частоту происшествий, но и интенсивность прессы, доступность журналистов и стимулы организаций сообщать о проблемах. Его нельзя принимать за полную статистику всех AI-инцидентов. Полезно показывать уровень подтверждения, первичный источник, тип вреда, наличие независимой проверки и дату последнего обновления. Спорные события могут сопровождаться заметкой о неподтверждённой причинной связи.

Общая прозрачность не требует публиковать идентификаторы пострадавших либо эксплуатационные подробности опасного сбоя. Открытая запись должна содержать достаточные агрегированные сведения для общественного контроля и иметь законную политику корректировки ошибок. Непроверенные машинно составленные резюме нельзя считать равными официальному отчёту регулятора.

Практический разбор 19.37

Рекомендация: Публиковать статус а не необоснованное обвинение

Условный пример: В реестре есть обращение клиента, но нет подтверждения архитектуры системы.

Разбор: Публичная категоричность может опередить доказательства.

Практические действия: Редактор маркирует сообщение, подтверждённое событие и нерешённые вопросы раздельно.

Критерий результата: Карточка позволяет понять доказательный статус без чтения закрытого досье.

38. Международный обмен уведомлениями и конфиденциальная координация

Одни и те же инциденты могут быть значимы для разных компетентных органов: AI Office, национального рыночного надзора, DPA, органа кибербезопасности, финансового надзора и судебной власти. Однако конкретное уведомление должно направляться по собственному правовому основанию. Предложение ARGA создать межведомственную точку координации не следует описывать как существующий универсальный международный механизм.

Для унификации коммуникации пригоден нейтральный фактологический пакет с событием, цифровыми идентификаторами, воздействием, мерами безопасности и версионностью, из которого формируются юридически специфичные приложения каждому адресату. Такая архитектура сокращает риск рассогласования фактов: например, чтобы в одном сообщении указать десять затронутых лиц, а в другом — тысячу, не объяснив разницу методики подсчёта.

При получении запроса из другой страны специалист сначала проверяет полномочие, цель, защищённый канал и законность международной передачи. Коммерческие обещания о «полном доступе регуляторов» не отменяют правил банковской тайны, персональных данных и адвокатской тайны. При конфликте доступа и конфиденциальности необходима надлежащая процедура или судебное разрешение, а не неформальная рассылка файлов.

38.1. Международная координация и контроль дублирования уведомлений

В одном трансграничном деле может возникнуть необходимость обращения к нескольким независимым органам. Для этого каждому сообщению присваивается отдельный номер и сохраняется карта того, какие факты переданы и в какой версии. Различия в законодательной терминологии фиксируются отдельно: государственный орган может использовать собственное понятие significant incident, не совпадающее с определением AI incident OECD. Даже корректная отправка сообщения одной организации не подтверждает автоматически получение его всеми нужными адресатами.

Обмен документами должен учитывать ограничения международной передачи данных и полномочия лица, инициирующего контакт. Если один орган ответил, что вопрос находится вне его компетенции, такой ответ сохраняют и используют для корректировки маршрутизации, не маскируя его под подтверждение существа экспертных выводов.

Практический разбор 19.38

Рекомендация: Согласовывать факты международного уведомления

Условный пример: Два партнёра направили разные описания одного события.

Разбор: Различие версий может вызвать ошибочные решения адресатов.

Практические действия: Координатор использует единый идентификатор, согласованную хронологию и журнал обновлений.

Критерий результата: Получатели имеют актуальную версию и сведения об исправлениях.

39. Методика независимого Observatoire ARGA: пять стандартов качества

Предлагаемая программа независимой оценки основана на пяти требованиях: прослеживаемость каждого существенного факта, воспроизводимость методики, правовая раздельность режимов, доказанная независимость проверяющих и измеримость мер исправления. Эти требования могут использоваться для добровольных экспертных заключений ARGA, аналитических консультаций и представления предложений регуляторам, но не учреждают новую юридически обязательную форму сертификации.

Процедура открывается квалификационным меморандумом: предмет, цели, стороны, конфликт интересов, сохранность данных, безопасность пострадавших и предварительная карта компетенций. Затем формируется реестр доказательств с уровнями доступа, программа тестов с запретом эксплойта в production без санкции и матрица правовых обязанностей. Результатом является не один «красивый PDF», а связанный пакет source register — analyses — remediation — review.

Предлагается публиковать методику в виде открытого стандарта качества с упором на корректное употребление терминов и контроль неопределённости. При этом ARGA не должна изображать себя органом официального надзора или обладателем полномочий по доступу к закрытой государственной информации. Такой подход повышает пригодность материалов для профессиональной дискуссии с институциональными партнёрами.

39.1. Этапы внутреннего контроля

Этап	Выход	Критерий завершения
Intake	Первичная карточка	Триггеры и безопасность оценены
Preservation	Evidence register	Источники и доступ зафиксированы
Investigation	RCA + правовая матрица	Рассмотрены альтернативы
Review	Независимая рецензия	Возражения разобраны
Closure	CAPA + восстановление	Исполнение измерено

39.2. Независимость и повторяемость проверки в составе экспертной группы

Observatoire ARGA может применять пятикомпонентную процедуру качества: регистрация источника, защита цифровой целостности, сравнение альтернативных гипотез, раздельная правовая оценка и мониторинг восстановления. В каждом блоке требуется проверка лицом, которое не принимало исходное техническое решение либо не связано с оценкой собственного результата. Конфликт интересов может возникать не только при прямом финансовом участии, но и при зависимости эксперта от поставщика модели, организации-клиента или ожидаемого судебного исхода. Обнаруженный конфликт раскрывают; не каждый конфликт автоматически исключает участие, но должен быть управляемым и понятным читателю заключения.

Перед внешней передачей доклад рекомендуется пройти через двойную проверку: технический специалист проверяет воспроизводимость и ограничения, юрист — правовой охват, даты и адресатов обязательств. Списки источников и утверждения должны быть машиночитаемо сопоставлены настолько, насколько это возможно без раскрытия конфиденциальных данных. При обнаружении ошибки в уже выданном заключении допускается только версия с явно обозначенным исправлением и сохранением предыдущей истории. Молчаливое замещение документа без отметки об изменении противоречит самой идее доказательственной прослеживаемости.

Практический разбор 19.39

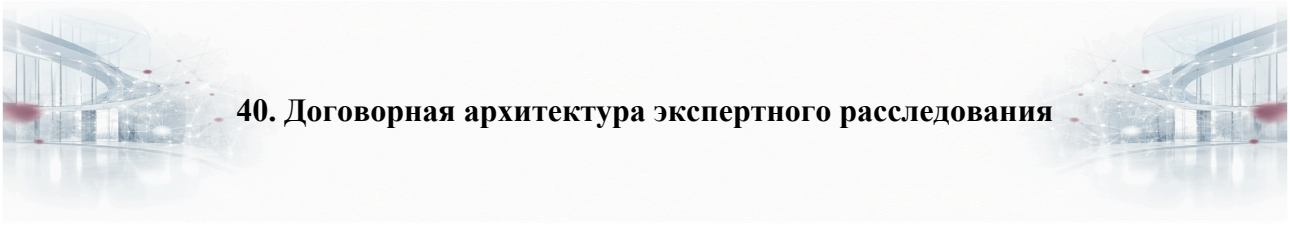
Рекомендация: Проверять пять элементов качества расследования

Условный пример: Досье содержит вывод и приложения, но не карту утрат и альтернатив.

Разбор: Внешний рецензент не видит ограничений анализа.

Практические действия: Куратор проверяет происхождение, причинность, права, независимость и исполнимость мер по связанным записям.

Критерий результата: Каждый элемент подтверждён документом либо указанным ограничением.



40. Договорная архитектура экспертного расследования

В договоре с независимым исследователем заранее определяют предмет проверки, законные основания доступа, объём конфиденциальности, хранение и возврат артефактов, участие специалистов, ограничение ответственности, публичное использование имени и право на объективное заключение. Отдельное приложение устанавливает, какая сторона является контролёром персональных данных, какие передачи допустимы и какие документы поставщик обязуется сохранить.

Не рекомендуется договорное положение, позволяющее заказчику беспрепятственно изменять фактические выводы по своему усмотрению. Допустима процедура fact checking, при которой заказчик предоставляет возражение с источником; рецензент фиксирует, принято ли оно и почему. По завершении работы сохраняют оригинал отчёта и версионный журнал, чтобы новые редакции нельзя было задним числом выдать за первые.

Для международных кейсов договор учитывает возможность применения национальных правил о конфиденциальности, санкционных ограничениях, экспортном контроле и сохранении специальных данных. Даже при законном договоре исполнитель не получает автоматически доступ к закрытым полицейским и медицинским системам.

40.1. Что требуется в договоре до начала расследования

В задании фиксируются юридический статус сторон, предмет независимого исследования, список доступных источников, разграничение технических и адвокатских услуг, условия доступа к конфиденциальной информации, режим интеллектуальных прав на заключение и порядок хранения цифровых оригиналов. Отдельно устанавливаются правила публикации, включая право человека или организации на возражения против раскрытия персональных сведений. Срок работы зависит от доступности доказательств; подрядчик не должен обещать подтвердить определённую гипотезу вне зависимости от результата проверки.

Конфликт между интересами плательщика и защитой пострадавших решается процедурами независимости и раскрытия ограничений. Если заказчик запрещает отражать существенные опровергающие данные, эксперт обязан определить, может ли продолжать работу и в каком качестве, вместо выпуска заведомо одностороннего заключения.

Практический разбор 19.40

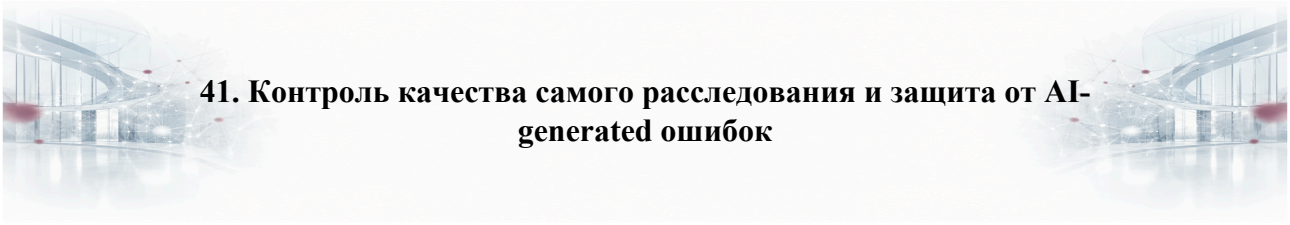
Рекомендация: Закреплять условия доступа и завершения договором

Условный пример: Исполнитель обязан назвать причину, хотя заказчик не даёт исторические логи.

Разбор: Договорное обещание превышает доступные возможности исследования.

Практические действия: Стороны определяют доступ, этапы, критерии результата и допустимый вывод о недостаточности данных.

Критерий результата: Приёмка не требует выдуманной определённости при отсутствии материала.



41. Контроль качества самого расследования и защита от AI-generated ошибок

Внутреннее использование генеративного ИИ для построения таймлайна, поиска несоответствий и подготовки текстовых таблиц допустимо только как вспомогательный инструмент. Каждая цитата и нормативная ссылка должна проверяться по первичному документу. Система может уверенно придумать номер решения суда, дату принятия закона или статистический показатель; без ручной верификации это создаёт новый доказательственный дефект.

Контроль качества расследования включает повторное восстановление ключевой цепочки другим исследователем. Он проверяет наличие первичного объекта, корректность временной связи, воспроизводимость вычисления и соответствие причинного вывода выполненной процедуре. Отдельно проверяются правовая редакция, конфиденциальность и допустимость внешнего раскрытия. Замечания классифицируются по влиянию на заключение; разрешение существенного возражения требует основания в материалах.

Каждое утверждение заключения должно иметь одну из трёх маркировок: опора на проверенный документ, вывод эксперта с описанной методикой, рабочее предложение Observatoire. Авторская рекомендация не может оформляться так, будто её прямо требует GDPR или AI Act. При обновлении закона сохраняют дату правового снимка и журнал изменений текста.

41.1. Контроль использования генеративных инструментов при подготовке заключения

Использование генеративной модели для первоначального поиска документов или редакционной помощи не превращает её ответ в независимый источник. Каждый процитированный судебный акт, нормативная норма, номер документа, дата и статистическое утверждение должны быть перепроверены по первоисточнику. Особый риск представляют вымышленные номера дел, смешение редакций закона, ложные цитаты из стандартов с платным доступом и автоматическое переписывание предположения в категорическую констатацию. Поэтому создают claim register: утверждение, источник, дата проверки, проверяющий и ограничения. Непроверенные утверждения удаляют либо чётко маркируют как исследовательские гипотезы.

Проверка содержательных пропусков начинается с вопроса, какое наблюдение могло бы изменить основной вывод. Рецензент выбирает существенную альтернативу и оценивает, была ли она исследована. Если альтернативу нельзя различить из-за утраченных данных, это должно менять формулировку результата. Наличие длинной хронологии и заполненных таблиц не заменяет такой проверки. Финальный статус расследования зависит от достаточности ответа на вопросы, а не от объёма документа.

Практический разбор 19.41

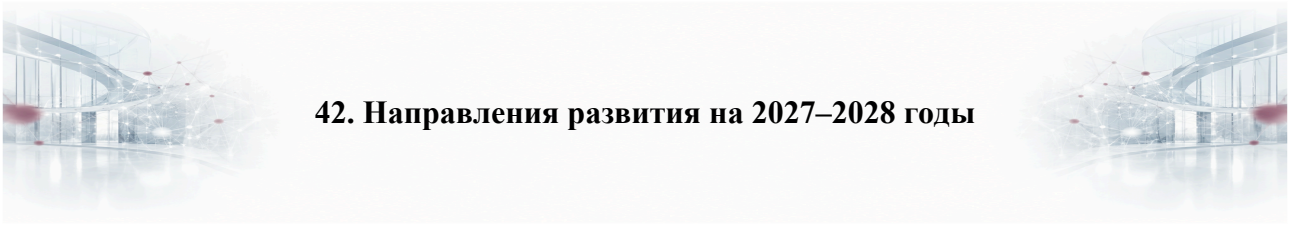
Рекомендация: Независимо проверять AI-помощь в расследовании

Условный пример: Помощник вставил в отчёт ссылку на несуществующий журнал.

Разбор: Автоматическое обобщение создало новое недостоверное доказательство.

Практические действия: Рецензент сверяет все существенные ссылки с реестром, расчёты пересчитывает отдельно.

Критерий результата: Каждое существенное утверждение подтверждено доступным материалом.



42. Направления развития на 2027–2028 годы

Прогноз правового развития должен исходить из известного календаря, но не выдавать ожидаемые проекты за принятые обязанности. Изменённая в июле 2026 года архитектура AI Act предусматривает последующие этапы высокорискового регулирования, что делает особенно ценными сравнительные исследования процедур добровольной отчётности и практической готовности организаций. Объективным направлением является мониторинг появляющихся официальных руководств, решений надзора и методик оценки. [2], [2]

Для международной программы целесообразно анализировать возможности взаимной совместимости схем инцидентной классификации OECD, отраслевого законодательства ЕС и национальных административных практик. Унификация должна касаться полей доказательственного сообщения, а не искусственного стирания различий между правовыми триггерами.

Отдельной задачей является создание механизмов безопасного обмена анонимизированными данными об ошибках AI в полицейской и финансовой аналитике. Подобное исследование должно начинаться с оценки конфиденциальности, законности и риска злоупотребления: данные о задержаниях, санкционных совпадениях и банковских операциях нельзя механически публиковать в общедоступной базе.

42.1. Сценарии международного развития стандартов

Можно ожидать дальнейшего сближения терминологии описания AI incident и практик независимого аудита, но прогноз не следует выдавать за действующее право. Важно сохранять гибкость внутренних форм под изменения национального законодательства, обновление NIST AI RMF и новые методики OECD. При последующих изменениях отраслевых правил юридическая карта должна обновляться с фиксацией даты, редакции и конкретных изменённых обязанностей, не стирая историю ранее направленных сообщений.

К перспективным вопросам относятся проверяемость адаптивных моделей, права доступа к trace для независимых специалистов, защита затронутого лица при использовании трансграничных AI-сервисов и управление ответственностью в длинной цепочке поставщиков. Методологические инициативы ARGA целесообразно формулировать как вопросы для публичной консультации, допускающие сопоставление разных юрисдикционных решений.

Практический разбор 19.42

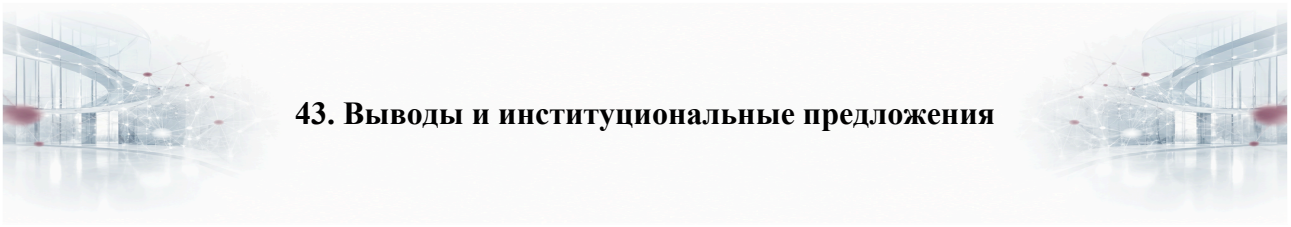
Рекомендация: Развивать программу по выявленным пробелам

Условный пример: Команда хочет новые направления, хотя в текущих делах не сохраняются входы модели.

Разбор: Расширение не исправит базовый дефект доказательств.

Практические действия: Руководитель сначала вводит сохранение контекста и пилот повторного анализа; затем оценивает новые задачи.

Критерий результата: Этап завершён по проверяемому результату, а не календарной отметке.



43. Выводы и институциональные предложения

Расследование AI-инцидента требует согласовать три самостоятельных результата: проверяемую картину событий, обоснование причинного механизма и правовую оценку последствий. Цифровая криминалистика сохраняет объекты, причинный анализ проверяет связь изменений с исходом, системная безопасность исследует отказ ограничений и обратной связи. Ни один из этих методов сам по себе не устанавливает вину конкретного лица. Их совместное применение позволяет точно определить, что подтверждено и какое действие из этого следует.

Для ARGA предлагается протокол с реестром исследовательских вопросов, картой источников и утрат, матрицей альтернатив, описанием воспроизводящих проверок и критериями завершения. Перед внешним использованием ключевое объяснение должно пройти повторный анализ по материалам дела. Если причинная часть остаётся неполной, подтверждённые факты и необходимые меры защиты могут быть изложены отдельно без вымышленной определённости.

Закрытие инцидента должно подтверждаться по технической, процессной и восстановительной части. Исправление модели не завершает работу, если ошибочная запись продолжает передаваться либо ограничение прав сохраняется. Для каждого затронутого результата нужен подтверждаемый исход: исправленная запись, прекращение ограничения, сообщение установленному получателю или мотивированное описание оставшегося препятствия. Эти критерии связывают научное исследование механизма с практической защитой лица.

Итоговые предложения ARGA носят исследовательский характер, предназначены для экспертных консультаций и не являются официальной позицией государственного или международного органа. Данный доклад не заменяет оценку местного адвоката, судебного эксперта или компетентного надзорного органа применительно к конкретному спору.

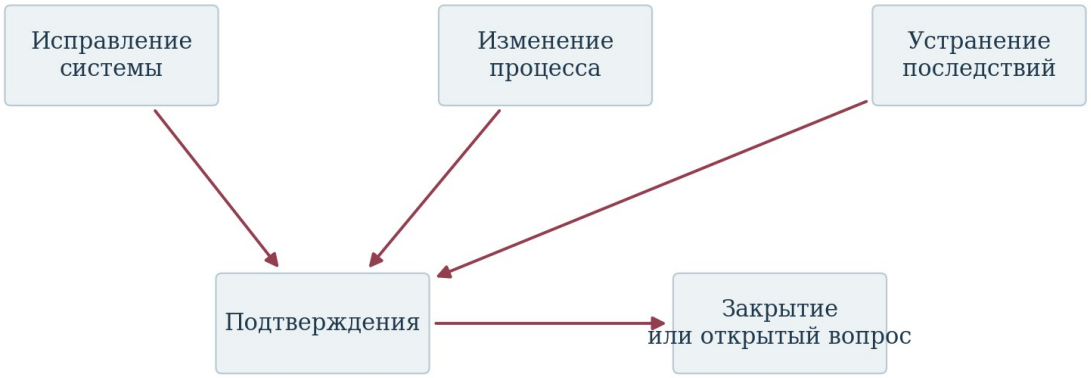


Рисунок 19.4. Три самостоятельных условия закрытия инцидента. Раздел 43. Техническое исправление не завершает восстановительную работу, пока ошибочные записи и ограничения сохраняются.

Практический разбор 19.43

Рекомендация: Закрывать техническую процессную и восстановительную части

Условный пример: Повторный тест успешен, но ошибочная запись ещё влияет на клиента.

Разбор: Работа не завершена по последствиям и распространённым копиям.

Практические действия: Куратор сверяет тест, изменение процесса и подтверждения исправления у получателей; нерешённое оставляет открытым.

Критерий результата: Закрытие обосновано по каждой части без общего необоснованного статуса.

Практические приложения: формы, протоколы и контрольные процедуры

Представленные ниже материалы являются рекомендательными аналитическими шаблонами Observatoire ARGA. Все поля, сроки внутренних этапов и классы приоритета являются проектными, если прямо не указан закон. Применять формы следует после проверки местного права, требований конфиденциальности и фактической компетенции участников. Внешняя передача сведений осуществляется в пределах законных полномочий.

Приложение 1. Первичная карточка сообщения об AI-инциденте

Карточка заводится до обсуждения причинности. Заполнение осуществляется непосредственно со слов заявителя с пометкой «сообщено» и отдельно по результатам просмотра исходных файлов с пометкой «проверено». Изменение классификации

производится новой записью, а не подменой первоначального документа. Цель процедуры — предотвратить потерю первого сообщения, точного времени и первоначальных артефактов.

Поле	Что внести	Проверка
IR-ID	Уникальный номер обращения	Не повторяется
Received at	UTC, канал, источник	Исходный заголовок
Reporter	Роль и возможность связи	Согласие на контакт
Alleged event	Дословное краткое сообщение	Отделено от вывода
AI system	Название, вендор, версия по возможности	Не выдавать догадку за факт
Actual harm	Документированные последствия	Прилагаемые источники
Imminent threat	Риск повторения и кому	Первичные меры защиты
Evidence list	Название, формат, носитель	Хэш при получении
Initial classification	I0–I4 и юридическая оценка отдельно	Автор, время

Инструкция исполнителю: подтвердить получение сообщения без заявления о виновности сторон, открыть журнал сохранности, определить необходимость защитных мер и передать квалификацию правовых сроков назначенному юристу. Отдельной строкой записать всё, что заявитель просит исправить, включая потенциально срочные меры по предотвращению задержания, блокировки или недопуска к необходимой услуге.

1.1. Инструкция по заполнению

В поле «сообщение заявителя» вносится исходная формулировка без редакторской нормализации. В поле «факты» отражаются только независимо подтверждённые элементы с указанием соответствующего Evidence ID. Поля «вероятность», «серьёзность» и «предполагаемая роль AI» не смешиваются. Если имя модели неизвестно, записывается «не установлено», а не наиболее вероятная торговая марка. Карточка получает номер версии, инициатора и дату; новые данные добавляются как уточнение, а не незаметно заменяют изначальные свидетельства.

Приложение 2. Схема приоритизации и немедленной защиты

Внутренняя шкала P0–P3 предназначена для управления ресурсами расследования; она не является правовой дефиницией serious incident. Приоритет P0 означает потенциально продолжающийся неотложный риск для жизни, свободы, здоровья или критического сервиса; P1 — значительный вред либо высокий риск повторения; P2 — локализованный подтверждённый инцидент; P3 — сигнал без установленного вреда. Класс пересматривается по мере получения сведений.

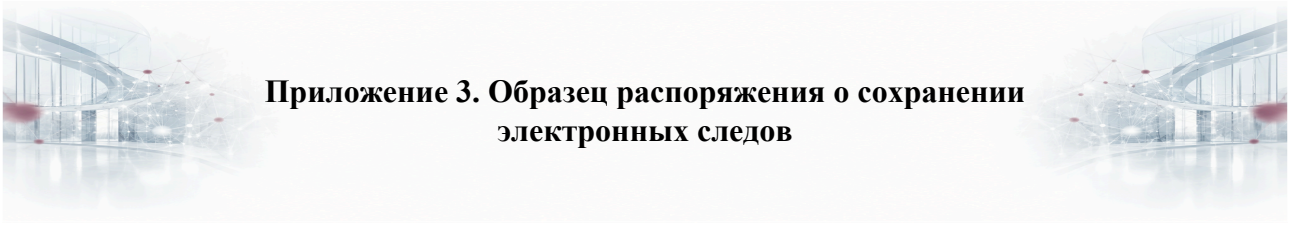
Приоритет	Первый шаг	Кто подтверждает действие
P0	Немедленная безопасная изоляция функции и эскалация	Incident lead + технический ответственный
P1	Сохранение доказательств и временное ограничение	Incident lead
P2	Плановое расследование с проверкой ущерба	Назначенный эксперт

Приоритет	Первый шаг	Кто подтверждает действие
P3	Сбор недостающих фактов и мониторинг	Ответственный за intake

Каждое решение содержит четыре обязательных мотива: обнаруженный сигнал, доступные подтверждения, риск неправильного ограничения системы и критерий возвращения к обычной эксплуатации. Не следует автоматически присваивать P0 каждому упоминанию слова «AI» или каждого сообщения в СМИ. Приоритет может быть высоким даже при неполных данных, если возможное промедление неприемлемо.

2.1. Взаимодействие с кризисной командой

При P0 предварительная оценка должна немедленно попасть лицу, имеющему полномочия изменить состояние опасной системы. Для P1 назначается ответственный за сохранение цифровых следов и за предупреждение пострадавших в допустимой мере. Отдельный наблюдатель записывает все операции изоляции. Если безопасность лиц требует немедленного действия, отсутствие полной технической картины не служит основанием ждать; мотивы приоритетного решения указываются задним числом как объяснение, не как вымышленное предварительное разрешение.



Приложение 3. Образец распоряжения о сохранении электронных следов

Цель распоряжения — приостановить обычное удаление ограниченного круга цифровых объектов, не остановив без необходимости законную работу организации. До отправки юрист устанавливает полномочие распоряжаться соответствующими информационными ресурсами, защиту данных и применимые периоды хранения. Запрос должен быть адресным, с перечислением систем, субъектов и временного диапазона.

Рекомендуемый предмет сохранения: production-логи и их индексы; артефакты текущего и предыдущего model deployment; runtime-конфигурации; системные подсказки; внешние tool calls; tickets; сообщения о коррекции; исходные данные; подтверждения доступа к интерфейсу; обновления политик. Если невозможно сохранить файл в оригинальном виде, описать процесс экспорта и известные потери метаданных.

Элемент распоряжения	Значение
Legal hold ID	Код, дата, уполномоченное лицо
Причина	Краткий сигнал без юридического предрешения
Объекты	Системы и категории логов
Интервал	Начало/конец с UTC и буфером
Ответственные	Data owner и evidence custodian
Допустимые исключения	Системная безопасность/нормы закона
Контроль	Подтверждение исполнения и журнал отклонений

Технический владелец письменно подтверждает исполнение или перечисляет невыполнимые пункты. После окончания оснований для сохранения распоряжение отменяется по документированной процедуре; нельзя превращать расследование в бессрочное хранилище чувствительных сведений.

3.1. Юридическая и техническая адресация запроса

Распоряжение о сохранении направляют ответственным администраторам, владельцам облачных сервисов и контрактно доступным внешним поставщикам. Запрос содержит конкретные пространства данных, временной интервал и ограничения доступа. Нельзя требовать неопределённого копирования всех персональных данных организации: scope и retention определяются нуждами расследования и правовыми основаниями. При отказе либо технической невозможности исполнения сохраняются сам запрос, ответ и альтернативные доступные действия.

Приложение 4. Реестр цифровых артефактов и проверка целостности

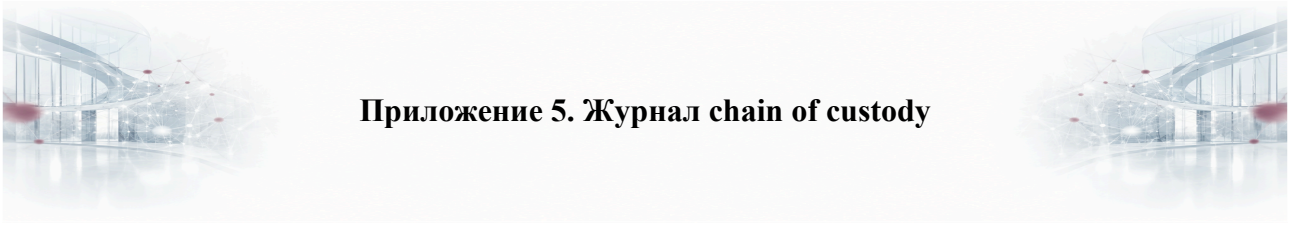
У каждого объекта появляется независимый Evidence ID. Сведения о получении, проверке и копировании не должны смешиваться: автор снимка экрана, эксперт по анализу и хранитель файла могут быть разными людьми. Использование SHA-256 является практической рекомендацией контроля целостности, а не универсальным юридическим требованием; необходимо учитывать качество инструмента, права доступа и возможность аппаратной аттестации.

Поле формы	Значение / образец
Запись 1	
Evidence ID	E-001
Тип	Raw log
Исходный источник	Пример: service A
Время получения UTC	Заполняется
SHA-256	Заполняется
Ограничение	Ограниченный доступ
Запись 2	
Evidence ID	E-002
Тип	Model manifest
Исходный источник	Пример: deployment
Время получения UTC	Заполняется
SHA-256	Заполняется
Ограничение	Коммерческая тайна
Запись 3	
Evidence ID	E-003
Тип	Решение оператора
Исходный источник	Пример: internal workflow
Время получения UTC	Заполняется
SHA-256	Заполняется
Ограничение	Персональные данные

Реестр необходимо вести в системе с контролируемой историей. При пересылке артефакта фиксируют не только кому передано содержимое, но и точное имя, хэш, размер, канал и полномочие получателя. Если инструмент экспорта меняет файл, создают новую запись с отношением «производная копия E-001», а не заменяют значение старого хэша.

4.1. Что именно подтверждает контрольная сумма

Сравнение хэшей используется для доказательства идентичности экземпляров файла на определённый момент, но не подменяет вопрос о полноте, происхождении и истинности содержимого. Наличие SHA-256 в инвентаре не гарантирует отсутствие изменения до получения объекта. При очередной рабочей трансформации создаётся новый производный объект с собственным идентификатором, а запись о его связи с оригиналом сохраняется в журнале. Отсутствующие контрольные суммы помечаются как ограничение, не заполняются предположительными значениями.



Приложение 5. Журнал chain of custody

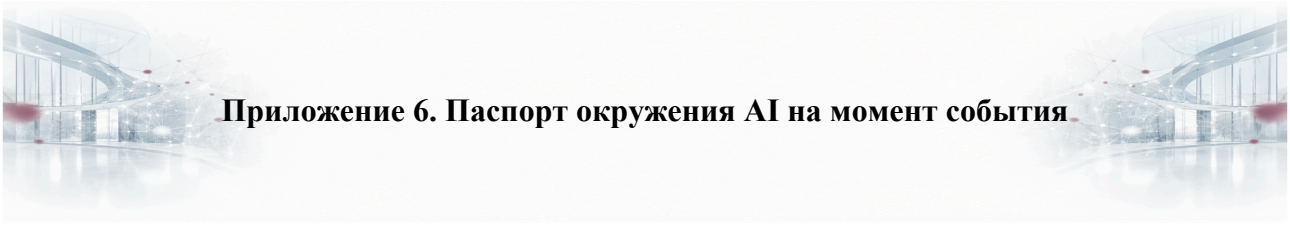
Цепочка хранения описывает физическое и логическое движение доказательства. Она призвана позволить проверяющему установить, что анализируемый объект соответствует полученному, и выявить промежутки несанкционированного доступа. Электронная подпись, журнал управления доступом, неизменяемое хранилище и хэш полезны, но сами по себе не гарантируют полноты источника: подписать можно и неполный экспорт.

Поле формы	Значение / образец
Запись 1	
Transfer ID	T-001
Evidence ID	E-001
От кого	Хранитель
Кому	Эксперт
Время UTC	Заполняется
Канал	Защищённо
Цель	Анализ
Подпись/след	Заполняется
Запись 2	
Transfer ID	T-002
Evidence ID	E-001
От кого	Эксперт
Кому	Архив
Время UTC	Заполняется
Канал	Защищённо
Цель	Возврат
Подпись/след	Заполняется

Запрещается оставлять без объяснения «общую папку», к которой любой сотрудник имел доступ администратора. В таких случаях риск изменения объекта прямо фиксируют в экспертном отчёте. При необходимости берут контрольный снимок истории доступа к самой папке.

5.1. Контроль передачи между экспертами

Каждая передача должна иметь принимающую сторону, цель, точное время и признак сохранности упаковки либо файла. Если пользователь получил копию по электронной почте и затем переслал её эксперту, в журнале отражают оба известных этапа и указывают недоступные. Неизвестный путь владения снижает доказательственный вес, но не всегда делает предмет полностью непригодным: итоговая оценка зависит от применимого процессуального права и иных подтверждений.



Приложение 6. Паспорт окружения AI на момент события

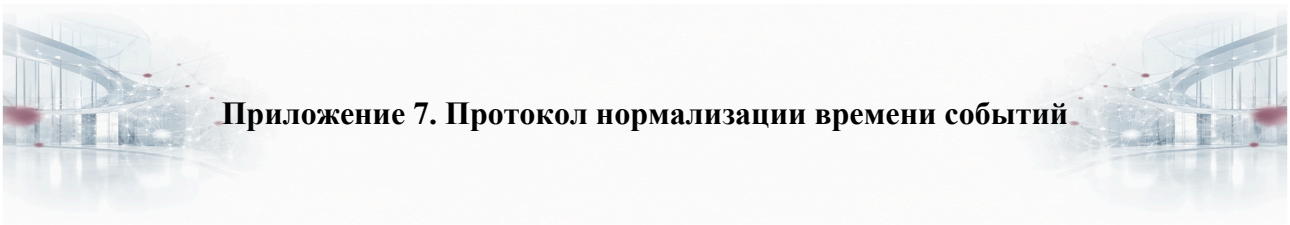
Для агентного сервиса воспроизводимость требует сведений, которые не видны на экране пользователю. Шаблон охватывает точное название и версию модели, API-провайдера, область развёртывания, ID контейнера, конфигурацию декодирования, prompts, политику безопасности, версию retrieval-корпуса, список инструментов и разрешения. Если часть полей закрыта поставщиком, фиксируется письменный запрос и ответ о невозможности доступа.

Слой	Параметр	Подтверждающий объект
Модель	Provider, model ID, revision	Manifest или ответ поставщика
Запуск	Deployment ID, дата, регион	CI/CD и cloud log
Подсказка	System/developer version	Снимок orchestrator
Контекст	RAG corpus и релевантные snippets	Snapshot index
Инструменты	Tool list, credentials scope	IAM export
Защита	Policy filters and guardrail version	Safety configuration
Человек	Approval workflow	UI activity record

При частично закрытом коде допускается black-box исследование, но оно не равно анализу исходного production-окружения. Все различия тестового и исходного режимов перечисляются в заключении.

6.1. Минимальный воспроизводимый паспорт

Паспорт окружения разделяет техническую версию модели, конфигурацию приложения и состояние внешних зависимостей. Например, одни и те же веса могут выдать различный ответ при изменении retrieval-корпуса, прав внешнего инструмента или системной инструкции. Если идентификатор версии вендором не раскрывается, следует сохранить документированные сведения о времени и маршруте вызова API. Паспорт нужен для определения пределов воспроизводимости, а не для создания ложной гарантии детерминированности.



Приложение 7. Протокол нормализации времени событий

Исходные журналы могут отражать локальное время, время Unix epoch, timezone offset либо неопределённую запись без временной зоны. Протокол преобразования должен хранить исходное значение и правило нормализации. При оценке сроков уведомления отдельно записывают момент осведомлённости юридического лица, а не только момент технического события. Если даты расходятся, создают шкалу с диапазонами неопределённости.

Поле формы	Значение / образец
Запись 1	
Event ID	EVT-01
Raw timestamp	Заполняется
Zone/offset	Заполняется
Normalized UTC	Заполняется
Confidence	High/Medium/Low
Origin	Log ID
Запись 2	
Event ID	EVT-02
Raw timestamp	Заполняется
Zone/offset	Не установлена
Normalized UTC	Диапазон
Confidence	Low
Origin	Слова свидетеля

Не допускается превращать технический timestamp последующего экспорта в дату события. Интервал неопределённости отражается и в графике, и в правовой записке. При необходимости используются независимые системные часы, почтовые заголовки или транзакционный журнал как точки синхронизации.

7.1. Ограничения нормализации времени

Коррекция часов не должна стирать исходные timestamp, предоставленные сервером, человеком или облачным сервисом. Для каждой записи указывают часовой пояс, возможный drift, метод пересчёта и доверие к источнику. Несовпадение на одну минуту иногда существенно для причинности, но наибольшую опасность представляют ошибки в сутках и смещение перехода на летнее время. При невозможности определить точный момент фиксируют интервал, а не искусственную точность до миллисекунды.

Приложение 8. Матрица обязательных и добровольных сообщений

Для каждого режима создаётся отдельная строка с правовой нормой, категорией системы, юридическим адресатом, триггером, временем осведомлённости, сроком, каналом уведомления и статусом. Это позволяет избежать ложного «общего дедлайна 72 часа». Таблица ниже служит напоминанием об основных разных режимах, а не автоматической юридической квалификацией. [1], [8], [10], [12]

Режим	Стандартный ориентир	Триггер и оговорка
AI Act ст. 55	Без неоправданной задержки	GPAI с системным риском
AI Act ст. 73(2)	Немедленно после установления/разумной вероятности связи; максимум 15 дней	Только после проверки охвата, переходного режима и специальных случаев
AI Act ст. 73(3)	Немедленно; максимум 2 дня от осведомлённости	Специальный вид серьёзного события

Режим	Стандартный ориентир	Триггер и оговорка
AI Act ст. 73(4)	Немедленно после установления или подозрения связи; максимум 10 дней от осведомлённости	Смерть и условия нормы
GDPR ст. 33	Без неоправданной задержки; по возможности до 72 часов от осведомлённости	Personal data breach и риск
NIS2 ст. 23	24 ч; 72 ч от осведомлённости; итоговый в течение 1 мес. после уведомления	Значительный инцидент; охват субъекта и национальная реализация. Для доверительных услуг уведомление — в течение 24 ч; для продолжающегося инцидента действует специальный порядок
DORA RTS 2025/301	4 ч от классификации и максимум 24 ч от осведомлённости; при поздней классификации — 4 ч от неё. Промежуточный: 72 ч от первого сообщения; итоговый: месяц от последнего промежуточного	Major ICT incident; отчётность в несколько стадий по ст. 5 RTS 2025/301

Практический порядок: сначала проверить применимость каждой нормы, затем рассчитать срок в локальном календаре соответствующего адресата, назначить ответственного и создать напоминание. Итоговая юридическая записка должна подписываться специалистом, проверившим фактический статус организации.

8.1. Заполнение юридического реестра сроков

Для каждой отдельной обязанности обязательны четыре поля: нормативное основание, применимый субъект, правовой триггер и стартовая отметка. Время выявления техником, сообщения клиенту и получения юридической квалификации могут различаться; их фиксируют в одном журнале, но не приравнивают. Юрист должен объяснить, почему выбран определённый момент начала отсчёта, включая данные о возможной более ранней осведомлённости организации. Сроки AI Act, GDPR, NIS2 и DORA нельзя заполнять одним шаблонным числом.

Приложение 9. Форма раннего сообщения компетентному органу

Введение: «Ниже приводятся известные на момент отправки обстоятельства. Квалификация причин и масштаба может быть уточнена после сохранения и независимой оценки артефактов». Эта формулировка является нейтральным примером: действующие формуляры регулятора и предусмотренные законом поля имеют преимущество.

Обязательные блоки: юридическое лицо и контакт ответственного; номер обращения; применимое правовое основание; дата осведомлённости; идентификатор и назначение AI-системы; подтверждённый вред и потенциальное воздействие; текущая оценка причинности; сведения о мерах безопасности; план дополнительного сообщения; защищённый канал обмена документами.

Поле	Содержательная инструкция
Правовой адресат	Название компетентного органа, подтверждённое юристом
Основание	Конкретная статья, редакция, дата применения
Сведения	Факты отдельно от предположений
Последствия	Подтверждено / оценивается / неизвестно
Сохранение	Какие сведения защищены от удаления
Контакт	Защищённый email или официальный портал

Рекомендуется прилагать инвентарный перечень, а не первичные конфиденциальные массивы без требования закона. Последующие уточнения отправлять с ссылкой на предыдущее уведомление и датированной таблицей изменённых сведений.

9.1. Содержательная достаточность раннего уведомления

Если специализированный режим допускает первоначально неполное сообщение, документ должен ясно различать подтверждённые факты и проверяемые гипотезы. Включают дату обнаружения, доступный масштаб, характер риска, временные меры и контакт для дальнейшего взаимодействия. Позднее дополнение содержит ссылку на первоначальный документ и объясняет изменения. Нельзя скрывать важную неопределённость из опасения выглядеть неподготовленно: правдивая неполнота в пределах применимой нормы предпочтительнее заведомой ложной точности.

Приложение 10. Запрос поставщику базовой модели и интегратору

Цель запроса — получить необходимую и соразмерную информацию для объяснения конкретного события, а не «весь исходный код разработчика». В письме следует указать идентификаторы операций, исследуемый период, характер последствий и конкретный юридический или договорный режим доступа. По отдельности запрашиваются журналы поставщика модели и интегратора интерфейса.

Перечень запросов: полный model ID и changelog; история миграции между версиями; timestamps inference; параметры и ограничения API; известные relevant incidents; история отказов guardrails; условия журналирования; порядок сохранения evidence; описание независимых тестов и их ограничений; контакт технического и юридического ответственного. Запрос не предполагает безусловного права на недоступные коммерческие данные; условия передачи согласуются законным образом.

Ответ	Действие расследователя
Полные данные	Проверить происхождение и контроль целостности
Агрегированная статистика	Уточнить знаменатель и выборку
Секретный материал	Предложить защищённый review
Нет данных	Зафиксировать отсутствие и причины
Отказ по закону	Запросить правовое основание отказа

Обязательное правило качества: не выдавать отсутствие ответа вендора за признание виновности. Однако невозможность проверить причинный механизм должна быть ясно указана как ограничение вывода.

10.1. Сохранность запроса поставщику

Запрос направляется через предусмотренный договором либо официальным процессом канал. К нему прилагают конкретные ID запросов, временное окно, перечень технических объектов и объяснение необходимости. Запрещено предполагать, что поставщик обязан раскрывать коммерческую тайну частному эксперту только потому, что начато расследование. В случае отказа полезно предложить проверку независимым специалистом в конфиденциальной среде или иной законный способ подтверждения фактов.

Приложение 11. Протокол опроса оператора и лица, принимавшего решение

Опрос разделяется на свободное изложение, уточнение фактов и предъявление ранее зарегистрированных артефактов. Интервьюеру запрещено подсказывать ожидаемое объяснение или начинать вопрос с утверждения, что алгоритм ошибся. Сначала выясняют должностную роль, доступные сведения, полномочия остановить действие, реальное время работы с карточкой и знакомство с инструкцией.

Рекомендуемые вопросы: что вы видели на экране; какой результат вам был доступен до принятия решения; существовал ли иной источник; могли ли вы отменить машинное предложение; какие предупреждения получали; как фиксировались отклонения; кем была утверждена политика применения; уведомлял ли вас кто-либо об аналогичных ошибках ранее?

Элемент протокола	Требование
Согласие и основание	Проверяется местным правом
Условия интервью	Время, место, участники
Свободный рассказ	Без ведущих формулировок
Показ артефактов	ID и порядок предъявления
Расхождения	Фиксируются без стирания исходного ответа

Не следует удалять противоречия из итогового протокола ради «чистоты» повествования. Расхождение между интервью и логом может отражать ошибку памяти, иной интерфейс либо существенный пробел управления.

11.1. Не смешивать память свидетеля и системный лог

При опросе оператору показывают нейтральную хронологию без преждевременного сообщения «модель признана неисправной». Вопросы о том, что он видел на экране и какие действия совершил, отделяют от его предположений о причинах сбоя. Если запись интервью используется как доказательство, сохраняется точная согласованная формулировка ответов, дата, способ записи и информация о возможной проверке. Политика хранения интервью должна учитывать трудовое и персональное законодательство.

Приложение 12. Матрица гипотез причинности

Для каждого серьёзного события формулируют не менее трёх осмысленных конкурирующих гипотез. Например: исходные данные ошибочны; модель неверно обработала корректные данные; вывод был корректен, но сотрудник применил его за пределами допустимого контекста. Для каждой гипотезы указывают подтверждающие и опровергающие наблюдения.

Не допускается ограничиваться гипотезами, которые заведомо менее правдоподобны, чем изначальная версия заказчика.

Гипотеза	Что должно наблюдаться	Что опровергнет	Статус
H1: ошибочный вход	Несовпадение с первичным источником	Вход корректен на дату решения	Открыта
H2: ошибочная модель	Ошибка при воспроизведении	Повторные независимые тесты не подтверждают	Открыта
H3: неправильный workflow	Действие вопреки регламенту	Регламент и действия совпали	Открыта
H4: внешняя атака	Trace инъекции/доступа	Соответствующих событий нет	Открыта

Вывод присваивается после оценки качества каждого источника и невозможности доступа к дополнительным данным. Если H1 и H3 одновременно существенны, нельзя искусственно выбирать «одну виновную причину». Методика допускает многофакторную атрибуцию.

12.1. Проверка гипотез причинности

Матрицу заполняют начиная с отрицательных альтернатив: неверный исходный документ, ошибочная маршрутизация, сбой ручного контроля, внешний злоумышленник, расхождение версий либо законная причина итоговой меры, не зависящая от AI. Для каждой гипотезы формулируется заранее понятный критерий подтверждения и опровержения. Если доступная информация не позволяет исключить несколько объяснений, итоговый статус остаётся «многофакторная неопределённость», а не «самая удобная версия».



Приложение 13. План контрольного воспроизведения и ограничения тестовой среды

Воспроизведение должно иметь определённый исследовательский вопрос. Перед экспериментом фиксируют версию модели, промпты, набор данных, схему инструментов, seed при наличии, параметры генерации и ожидаемые критерии сравнения. Не следует проводить опасные кибероперации, действия с персональными данными или моделирование вреда в открытой production-среде. Сначала оценивают безопасность и возможность использовать синтетические/обезличенные тестовые объекты.

Элемент	Проверочное требование	Документ
Предмет	Одна проверяемая гипотеза	Test plan ID
Среда	Изолированная и разрешённая	Environment manifest
Данные	Разрешённые, необходимые	Dataset manifest
Повторения	Предопределённое число запусков	Sampling plan
Успех	Наблюдаемый критерий	Outcome rubric
Прерывание	Условие stop	Safety plan

Результаты включают число удачных и неудачных попыток, разброс outputs, случаи отказа тестовой среды и отступления от плана. Если тестовая модель уже отличается от production, в отчёте указывают уровень переноса выводов: подтверждает механизм, косвенно иллюстрирует возможность или не пригодна для оценки исторического события.

13.1. Дизайн допустимого повторного испытания

Перед запуском контрольного воспроизведения утверждают допустимые данные, объём полномочий агента, условия остановки и критерии оценки. Использование реального банковского счета, доступа к критической инфраструктуре или чувствительной информации возможно только при надлежащем разрешении и защите. Для каждого прогона фиксируют полный протокол, версию окружения и результат. Тест, проведённый исключительно на изменённой модели, не должен выдаваться за точную реконструкцию исходного инцидента.

Приложение 14. Статистическая форма проверки редкого вреда

При расследовании редких событий особенно важно не выдать «0 из 100» за доказательство нулевой частоты. Отчёт содержит точную величину выборки, механизм отбора и доверительную границу. Для иллюстрации при нуле событий среди 100 независимых наблюдений односторонняя приближённая верхняя граница 95% по правилу трёх составляет 3/100 = 3%. Метод применим только при допущениях независимости и репрезентативности; для кластеризованных событий требует корректировки.

Поле	Пример обозначения	Проверка
N	100	Число независимых экспозиций
Ошибки	0	Определение ошибочного события
Доля	0%	Точечное наблюдение, не истинный риск
Оценка верхней границы	≈3%	Указать выбранный метод
Смещения	Отбор, потерянные события	Описать без сокрытия

Если жалобы собраны только среди тех, кто знает о праве на оспаривание, выборка вероятно смещена. Фактическая частота может отличаться в обе стороны, а доверительный интервал не исправляет систематическую ошибку отбора.

14.1. Отчёт о неопределённости при редких событиях

При нуле зарегистрированных инцидентов оценка должна включать размер и характер выборки, зависимость наблюдений, надёжность детектора и возможные пропущенные события. Грубое правило 3/n в подходящей биномиальной постановке представляет лишь ориентир верхней границы и не применяется автоматически к серийным коррелированным наблюдениям. При сравнении систем сообщаются абсолютные частоты, знаменатели и доверительные интервалы, а не только проценты изменения. Принятый риск не выводится из одной статистики без оценки тяжести последствий.

Приложение 15. Проверка дискриминационного воздействия

Рабочая матрица для высоковлияющих решений должна отдельно учитывать правовые основания использования чувствительных признаков в аудите. Цель состоит не в автоматическом присвоении ярлыка дискриминации, а в оценке того, есть ли статистически и юридически релевантная разница, обусловленная недопустимыми факторами. При оценке исходов нужно контролировать качества данных и сопоставимость групп.

Поле сравнения	Группа A	Группа B	Комментарий
Количество заявок	N_A	N_B	Разный размер выборки

Поле сравнения	Группа A	Группа B	Комментарий
Отказы	R_A	R_B	Правомерность решения отдельно
Доля отказов	R_A/N_A	R_B/N_B	Не доказывает причинность
Исправленные ошибки	C_A	C_B	Проверить сроки исправления
Missing data	M_A	M_B	Риск селективности

Прежде чем использовать показатели в заключении, следует провести проверку определений, исключить двойной счёт и убедиться, что алгоритмическая процедура действительно влияет на итоговое решение.

15.1. Оценка сопоставимости групп

Демографическое сравнение требует правового основания обработки признаков и анализа репрезентативности групп. Необходимо установить, одинаковы ли правила разметки, вероятность обращения с жалобой и качество исходных записей. Результат может быть обусловлен не только моделью, но и отдельной политикой принятия решений. Отсутствие достоверных групповых данных не следует компенсировать выдуманными значениями: эксперту следует ограничить вывод и указать требования к допустимому дополнительному анализу.

Приложение 16. Проверочный лист независимости эксперта

Независимость не обеспечивается одной подписью под словом «беспристрастен». Эксперт раскрывает финансовое участие, ранее оказанные услуги стороне, доступ к proprietary data, ограничения заказчика на публикацию, участие в разработке системы, личные отношения с руководством и обещанные будущие контракты. Конфликт может быть управляемым или несовместимым с поручением; решение должно быть мотивированным.

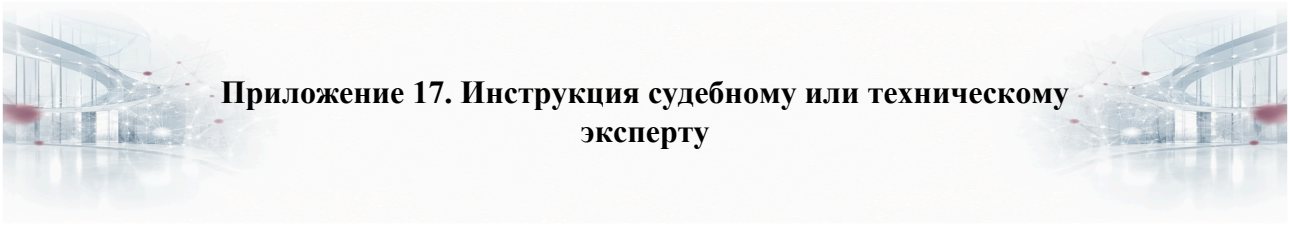
Проверка	Да/Нет	Как компенсируется
Ранее участвовал в разработке?	Заполняется	Независимый reviewer
Оплата зависит от результата?	Заполняется	Пересмотр условий
Контролируется публикация отрицательного вывода?	Заполняется	Право автора на текст
Есть доступ к оригиналам?	Заполняется	Раскрыть ограничения
Используются собственные платные инструменты?	Заполняется	Воспроизводимость другими

В заключении необходимо назвать организационное место эксперта. Частный исследователь, привлечённый на договоре, не должен именоваться «назначенным государственным расследователем», если отсутствует соответствующий акт назначения.

16.1. Документирование независимости

Форма раскрытия конфликтов заполняется не только руководителем проекта, но и ведущими техническими и правовыми авторами. Указываются договорные отношения с проверяемым поставщиком, прежнее участие в проектировании системы и материальная

заинтересованность в определённом исходе. Если независимость ограничена, отчёт должен обозначать пределы заключения и проверку другой группой. Название Observatoire ARGA не подразумевает государственного мандата или официальной функции надзора.



Приложение 17. Инструкция судебному или техническому эксперту

Техническое задание задаёт вопросы, на которые возможно ответить исследовательским методом: какие версии модели использовались, можно ли реконструировать вход, как устроена цепочка AI-to-human-to-action, какие альтернативные причины подтверждаются, насколько результаты воспроизводимы, где пробелы доказательств. Не стоит задавать вопрос «виновна ли организация», если он требует судебного решения за пределами специальной компетенции эксперта.

Вопрос	Сведения для анализа	Граница вывода
Как сформировался результат?	Prompt, data, model, tool traces	Без независимой правовой квалификации
Почему решение принято?	Логи человека и регламент	При неполном доступе — ограничение
Можно ли повторить ошибку?	Изолированный тест	Не означает идентичную историческую среду
Было ли воздействие на права?	Документы о последствиях	Юридический вывод отдельно

Договор должен гарантировать эксперту право сообщить о данных, которые опровергают исходную гипотезу заказчика. Требование исключить неудобные файлы из рассмотрения само подлежит документированию.

17.1. Содержание технического задания эксперту

Задание формулирует проверяемые вопросы, а не желаемый ответ. Например: «существовал ли исходный источник спорного утверждения?» либо «документируется ли вызов инструмента при данных правах доступа?». Отдельно перечисляют недоступные системы, допустимые методы эксперимента и запрет на неправомерное изменение объекта. Эксперт сообщает о невозможности ответа при недостатке данных, вместо того чтобы подменять отсутствие следа вероятностным рассуждением.

Приложение 18. Календарь расследования и формирование контрольных точек

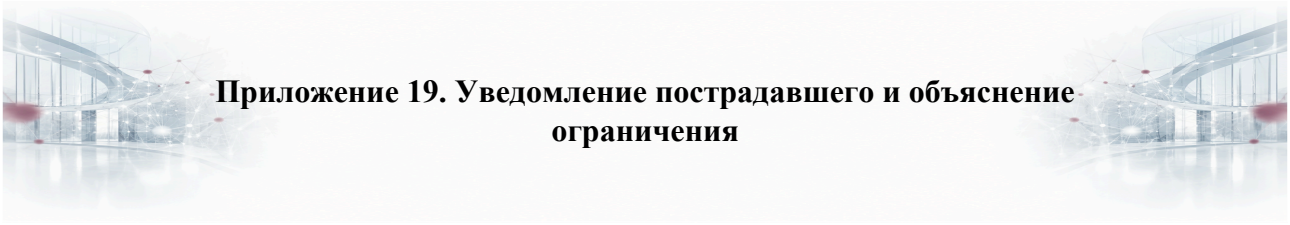
Внутренний график строится вокруг реальных правовых дедлайнов. В первый момент фиксируют риск, артефакты и ответственных; далее, по мере поступления сведений, организуют юридическую квалификацию, независимый тест и обновление заинтересованных лиц. Сроки отдельных внутренних задач назначаются исходя из риска и ресурса; они не отменяют законные сроки и могут быть сокращены для неотложных случаев.

Контрольная точка	Результат	Если не выполнено
Intake complete	Первичная карточка и контакт	Установить альтернативный источник
Evidence hold	Письменные подтверждения сохранения	Отразить риск утраты
Legal matrix	Проверены режимы сообщений	Эскалация юридическому координатору
Root cause review	Проверены конкурирующие гипотезы	Дополнить экспертизу
CAPA validation	Испытание исправлений	Не закрывать инцидент

Календарь должен содержать владельца задачи и журнал переноса сроков. Для срочных уведомлений недопустимо скрывать просрочку внутри общего статуса «идёт расследование».

18.1. Проверяемые контрольные точки

Календарь содержит отдельные даты сохранения цифровых доказательств, первичной оценки правового режима, решения о защитных мерах, направления уведомлений, получения материалов от поставщика, технического воспроизведения, внешней проверки и восстановления прав. Отсрочка одного этапа не должна автоматически переносить юридический срок другого. При изменении графика сохраняются исходная версия и мотивированное решение об изменении.



Приложение 19. Уведомление пострадавшего и объяснение ограничения

Текст должен быть понятен человеку без технического образования. Сообщают о событии, о том, что подтверждено и ещё исследуется, о доступном способе обращения, процедуре исправления ошибочных данных и контакте для получения помощи. В случаях, когда законодательство ограничивает объём раскрытия, нужно сообщить максимально допустимую фактическую информацию без нарушения охраняемой тайны.

Рекомендуемые смысловые блоки: «какое действие было затронуто», «какие решения о вас приняты», «что уже исправлено», «какие сведения сохраняются», «как оспорить решение», «какие сроки ответа установлены применимым правом». Не следует обещать компенсацию без проверки оснований и полномочий. Также нельзя заменять уведомление фразой «произошла техническая ошибка», если конкретное решение по человеку продолжает действовать.

Проверка перед отправкой	Результат
Текст не приписывает человеку вину?	Заполняется
Понятно, куда обращаться?	Заполняется
Указаны реальные меры исправления?	Заполняется
Не раскрыты данные других лиц?	Заполняется
Соблюдены местные правила уведомления?	Заполняется

19.1. Как сообщать пострадавшему о неизвестных фактах

Обращение к затронутому лицу должно быть понятным и практически полезным. Следует раздельно описать подтверждённый факт, вероятный риск, защитные действия, доступный способ оспаривания и координаты законного контакта. Не допускается обещать окончательное устранение последствий, если часть downstream-организаций ещё не подтвердила корректировку. Уведомление не должно раскрывать личные данные других заявителей и сведения, публикация которых могла бы повысить риск безопасности.



Приложение 20. Реестр корректирующих действий CAPA

Техническое исправление считается исполненным только после проверки ожидаемого результата. В реестре отражают связь меры с выявленной коренной причиной, ответственного, критерий теста, статус и остаточные риски. Сильным доказательством являются заранее сформулированные критерии принятия изменения, а не задним числом подобранная метрика успешности.

Поле формы	Значение / образец
Запись 1	
CAPA ID	C-01
Причина	Ошибочная транслитерация
Действие	Версионированный normalization rule
Владелец	Data team
Тест	Regression set
Риск после	Заполняется
Запись 2	
CAPA ID	C-02
Причина	Нет ручного override
Действие	Новый workflow
Владелец	Operator
Тест	UI exercise
Риск после	Заполняется
Запись 3	
CAPA ID	C-03
Причина	Неполные логи
Действие	Immutable trail
Владелец	Security team
Тест	Audit replay
Риск после	Заполняется

Для значимого инцидента назначают independent verification. Если и исправление, и его приёмку осуществляет единственный сотрудник, риск конфликта функций указывают в заключении.

20.1. Доказательство исполнения CAPA

Наряду с отметкой «выполнено» хранится конкретное доказательство: конфигурационный diff, протокол независимого теста, документ о переобучении сотрудников либо подтверждение исправления данных у адресата. Назначенный исполнитель и проверяющий должны быть различимы. Если мера лишь частично снизила риск, residual risk остаётся отдельной записью с владельцем и датой следующего контроля. Исправление документации без исправления продолжающегося вреда не считается достаточным закрытием.

Приложение 21. Протокол решения о возобновлении системы

Решение о restart принимается в рамках законных полномочий после оценки безопасности и остаточного риска. Отчёт включает список исправленных дефектов, результаты повторных испытаний, обнаруженные регрессии, независимую рецензию, условия наблюдения после запуска и процедуру обратного отключения. При необходимости ограничивают класс пользователей и объем разрешённых действий.

Вопрос	Подтверждение
Сохранены оригиналы?	Evidence ID и хранилище
Первичная причина устранена?	RCA и CAPA тест
Остаточные риски раскрыты?	Risk register

Вопрос	Подтверждение
Средство защиты пострадавших запущено?	Remediation register
Руководитель подписал ответственность?	Decision memo

В протоколе отдельно указывается «критерий немедленного возврата в безопасный режим». Возобновление без возможности наблюдать повторение исходного опасного паттерна является недостаточно подтверждённым решением.

21.1. Условия возобновления эксплуатации

Решение о запуске не должно приниматься только командой, заинтересованной в соблюдении коммерческого графика. Для важной системы требуются техническое подтверждение исправления, тест регрессий, проверка правовых уведомлений, анализ оставшегося риска и назначение лица, способного остановить сервис повторно. При невозможности воспроизвести исходную причину действует более осторожный режим ограничения функциональности и расширенного мониторинга.

Приложение 22. Таблица уровней конфиденциальности и законной передачи

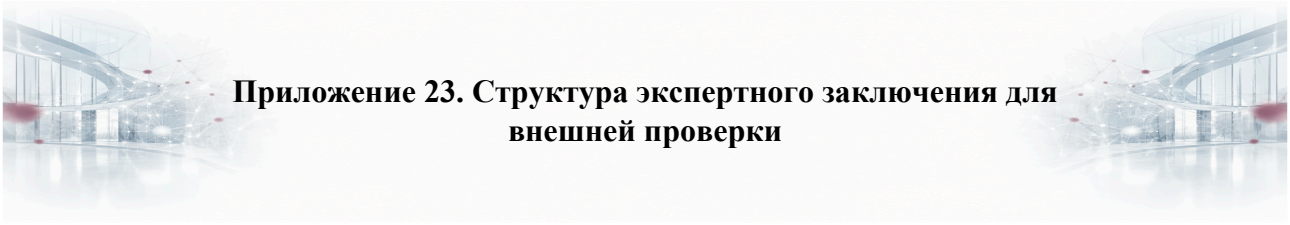
Разделение доступа начинается с определения цели и правового основания. Не все участники расследования должны видеть полные входные тексты, банковские номера или сведения о состоянии здоровья. При передаче стороннему эксперту назначают конкретную роль и фиксируют уровень доступа. Публичная публикация итогов должна иметь самостоятельную проверку на персональные данные, коммерческие тайны и злоупотребимость деталей.

Категория	Возможный доступ	Типовая мера
Публичная хронология	Общая аудитория	Анонимизация
Рабочие машинные логи	Техническая группа	Псевдонимизация
Персональные карточки	Юрист и уполномоченный эксперт	Ограниченный workspace
Уязвимости безопасности	Узкая security team	Controlled disclosure
Адвокатская переписка	Только уполномоченные лица	Особый режим защиты

В каждом случае проверяются применимые законы и договоры. Статус «внутренний аудит» сам по себе не является универсальным исключением из защиты данных.

22.1. Практика разграниченного доступа

Каждой категории материалов соответствует перечень допустимых получателей. Исходные медицинские сведения, персональные данные, секреты организации и материалы правоохранительных процедур не должны попадать в общий архив автоматически. Для публикационной версии применяют минимизацию данных и содержательную проверку, что удаление не искажает смысл вывода. Список удалённых категорий и причина сокрытия сохраняются во внутреннем реестре; нарушение доступности доказательства отмечают как ограничение независимой проверки.



Приложение 23. Структура экспертного заключения для внешней проверки

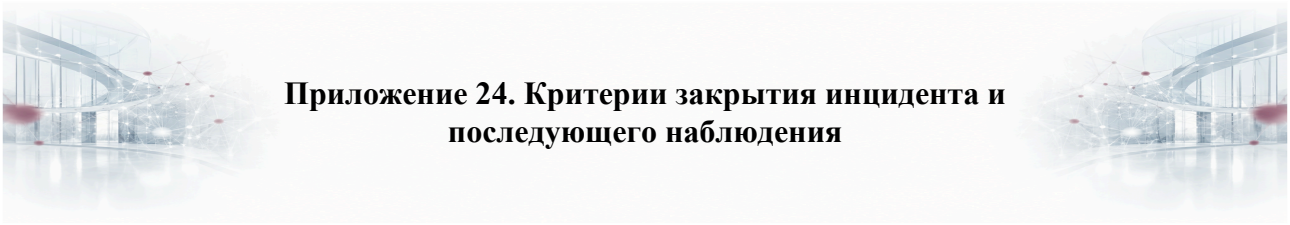
Заключение должно быть составлено так, чтобы регулятор или судебный эксперт могли отдельно проверить факты, методы и оценки. Рекомендуемый порядок: титульный лист; мандат; независимость; вопросы; источники; ограничения; установленные факты; техническая схема; результаты воспроизведения; конкурирующие гипотезы; причинный вывод; правовая карта как отдельное приложение; защитные меры; приложения доказательств.

Глава заключения	Задача	Антиошибка
Мандат	Кто и что поручил	Не утверждать государственное назначение
Источники	Проверяемый инвентарь	Не ссылаться на неназванные файлы
Метод	Эксперимент и допущения	Не скрывать тестовую среду
Результат	Факты, оценки, неизвестное	Не смешивать степени уверенности
Возражения	Неудобные данные и альтернативы	Не удалять dissent

Рекомендуется приложить табличный source-to-claim index. Длинный доклад без такого индекса может быть литературно убедителен и доказательственно непригоден.

23.1. Полнота экспертного заключения

Заключение включает цепочку от поставленного вопроса к источникам и методу, затем к фактам, альтернативным гипотезам, оценке ограничения доступа и результату. Каждое категорическое утверждение должно иметь ссылку на проверяемый артефакт или официальный правовой текст. Читателю должно быть понятно, какая часть подготовлена техническим экспертом, какая — юристом и какие институциональные рекомендации принадлежат Observatoire ARGA. Формальные атрибуты отчёта сами по себе не создают статуса судебного эксперта.



Приложение 24. Критерии закрытия инцидента и последующего наблюдения

Закрытие производится только после выполнения проверочных условий: исключён продолжающийся непосредственный вред; все важные источники сохранены; правовые уведомления выполнены или документированно признаны неприменимыми; объяснение причин достаточно обосновано; корректирующие меры прошли тест; индивидуальные последствия рассмотрены; остаточная неопределённость раскрыта.

Контрольный критерий	Требуемый артефакт
Нет непосредственного продолжающегося вреда	Акт защиты/повторного теста
Правовая карта обновлена	Записка юриста
Артефакты доступны	Реестр и chain of custody

Контрольный критерий	Требуемый артефакт
RCA рецензирована	Независимое заключение
CAPA принята	Подписанный verification report
Затронутые лица имеют канал защиты	Реестр восстановительных мер

Статус «закрыто» не означает уничтожение истории. Срок хранения определяется правовыми основаниями и функцией доказательств; после его завершения применяется документированная политика удаления.

24.1. Отложенное обнаружение повторных последствий

Формальное завершение расследования сопровождается сроком повторного контроля. Он зависит от того, продолжает ли компания использовать систему, какие копии спорных записей могли сохраниться и когда затронутые лица могут вновь испытать последствие. В реестре фиксируют канал для поздних сообщений, владельца post-incident review и условия повторного открытия дела. Уничтожение доказательств по истечении применимого срока хранения оформляется отдельно от решения о закрытии инцидента.

Сценарные приложения: методика применения в типовых делах

Все следующие ситуации сконструированы для проверки исследовательской методики. Они не отражают реальные происшествия или результаты клиентской работы. В каждом случае указан тип проблемы, минимальная доказательственная база, допустимые и недопустимые выводы, отдельные правовые вопросы и условия восстановления прав.

Приложение 25. Сценарий: ошибка идентичности в санкционной проверке

Условие (вымышленное). Банк в ЕС использует внешний инструмент, сопоставляющий клиентские профили с санкционными списками и данными о собственниках компаний. После обновления правил транслитерации система выдаёт высокую оценку сходства клиентского имени с именем лица из списка. Сотрудник видит итоговый «risk alert» и приостанавливает крупный платёж. Клиент указывает на иную дату рождения и гражданство. Через три недели банк признаёт отсутствие совпадения, но сведения о тревоге остаются у внешнего поставщика и становятся причиной новых запросов.

Исследователь не делает вывод, что незаконную блокировку «принял AI». Следует определить, какой юридический механизм применил банк, существовало ли самостоятельное основание для проверки операции и мог ли специалист оценить признаки личности. В деле особенно важны лог версии алгоритма совпадения, результат исходного запроса, полнота представленных сотруднику идентификаторов, история решений и рассылок, downstream-

клиентские жалобы. Вариант ошибки в данных и вариант некорректного человеческого подтверждения должны проверяться раздельно.

Шаг проверки	Артефакт	Что можно установить
Идентификация сигнала	Original match event	Точные сравниваемые поля
Ручная оценка	Employee decision trace	Роль человека
Основание меры	Compliance memo	Законность ограничения отдельно
Исправление	Correction tickets	Где ошибка устранена
Повторение	Provider downstream logs	Продолжается ли вред

Причинная гипотеза H1: низкое качество нормализации повысило вероятность ошибочного совпадения. H2: система корректно уведомила о сходстве, но интерфейс скрыл неопределённость. H3: банк применил собственный чрезмерно широкий запрет независимо от машинной оценки. Для анализа нужны контрольные пары имён, проверка порога и доступ к регламенту manual review. Отсутствие наблюдаемого false negative нельзя автоматически превращать в основание радикально увеличить порог.

Отдельная правовая карта охватывает санкционное право, GDPR при обработке персональных данных и возможные секторальные банковские нормы. Обязанность сообщения по статье 73 AI Act не устанавливается без оценки категории системы и юридического порога серьёзного инцидента. Финальное средство защиты включает письменный пересмотр блокировки, исправление ошибочного профиля и проверку уведомлений третьим лицам.

25.1. Рабочая процедура проверки

Исходным объектом является конкретное решение о задержке либо блокировке операции и технический сигнал, который ему предшествовал. Эксперт сопоставляет идентификаторы списка, время его актуализации, результаты идентификационной проверки и версию правил матчинга. Не следует считать одинаково надёжными паспортные данные, непроверенную дату рождения, транслитерацию и одно лишь совпадение имени. Отдельно документируется, какая информация была доступна сотруднику, подтвердившему alert, и имел ли он реальную возможность отменить машинную рекомендацию.

Следующий шаг — карта распространения ошибочной записи. Она показывает, какой поставщик сформировал совпадение, какие организации получили результат и какие действия потребуются для коррекции разных копий данных. Восстановление банковского обслуживания определяется применимым правом и независимым решением компетентного финансового учреждения. Экспертный вывод ограничен подтверждаемой причиной технического совпадения, распределением ролей и фактом исполнения корректирующих мероприятий. Этот сценарий не является описанием реального клиента ARGA.

Приложение 26. Сценарий: алгоритмическое совпадение при пересечении границы

Условие (вымышленное). При пограничной проверке система сопоставления выдаёт совпадение с записью об иностранном уголовном преследовании. Пограничный сотрудник задерживает лицо для дополнительной проверки, ссылаясь на отсутствие иных идентификаторов. Через несколько часов прокуратура освобождает человека, установив расхождение даты рождения и национальности. Внутренние документы не раскрывают, использовался ли AI либо детерминированное правило поиска.

Первое требование — не приписывать действие определённой системе без доказательств. Исследователь собирает решение о задержании, официальное основание проверки, даты и время, идентификаторы, отражённые в национальном решении, а также любые законно доступные следы автоматизированного сопоставления. Факт освобождения сам по себе не доказывает незаконности первоначальной проверки, а существование международного сообщения не означает его автоматической достоверности или правильной интерпретации.

Второе требование — реконструировать цепочку от внешней информации до национального решения. Нужно узнать, какие дополнительные проверки стали доступны сотруднику и была ли возможность быстро уточнить данные. Следует оценить качество переводов, транслитерацию, различие статусов: сообщение о проверке, ориентировка, национальный розыск, судебное решение. В каждом звене указываются компетенция органа и процессуальное основание.

Проверочная ветвь	Источник	Допустимый вывод
Сведения об идентичности	Официальные документы	Подтверждённые различия
Происхождение информации	Законный ответ компетентного органа	Статус и дата записи
Работа классификатора	Доступная техническая документация	Границы AI-вклада
Решение об ограничении свободы	Протокол и судебные материалы	Правовая квалификация юристом
Последующие последствия	Документы о поездках и жалобах	Масштаб повторения

Сценарий показывает пределы чисто технической ремедиации. Даже если скорректировать механизм совпадения, неправомерные копии сведений могут оставаться в других системах. Поэтому программа завершения должна содержать подтверждённую коррекцию источника и доступные способы юридического оспаривания в соответствующих юрисдикциях. При отсутствии доступа к закрытому национальному журналу экспертиза обязана признать невозможность установить точный механизм AI.

26.1. Рабочая процедура проверки

Первым документом должен стать процессуальный акт о действиях пограничного органа, а не догадка о существовании уведомления INTERPOL. В рамках дела собираются подтверждения идентичности, сведения о времени проверки, применённом национальном основании и доступные официальные ответы. Если орган использовал автоматическое сопоставление, дополнительно исследуются набор сравниваемых полей, логика порога и

участие сотрудника. Неизвестность алгоритма обозначается прямо; нельзя утверждать «машина перепутала людей» на основании одной похожей фамилии.

Материалы распределяются между техническим экспертом и адвокатом соответствующей юрисдикции. Первый анализирует следы обработки данных; второй определяет допустимость доказательств и надлежащие способы оспаривания задержания, экстрадиционного решения или хранения записи. Вывод о нарушении права на свободу принадлежит компетентной правовой оценке, а не автоматически следует из сообщения о технической ошибке. Приоритет отдаётся доступным срочным процессуальным мерам защиты человека.

Приложение 27. Сценарий: автоматическая классификация в процедуре предоставления убежища

Условие (вымышленное). Административная система сортирует заявления по типам доказательств и использует AI для перевода, извлечения дат и выявления несоответствий. В одном деле алгоритм ошибочно объединяет даты из разных календарей и помечает рассказ заявителя как противоречивый. Принимающий решение сотрудник ссылается на автоматически подготовленное краткое резюме. Позднее подлинные документы и перевод показывают, что противоречие возникло при обработке.

Расследование должно установить не только переводческую ошибку, но и степень её влияния на мотивы решения. Требуются все редакции summary, входные документы, настройки перевода, язык исходника, реальные цитаты в официальном решении и процедура human oversight. Отдельно проверяется, имел ли человек возможность ознакомиться с оригинальным материалом и исправить неверный перевод до неблагоприятного решения.

Особую сложность представляет хранение чувствительных данных: сведения об убежище могут касаться преследования, здоровья, политической ситуации и родственников за границей. Техническая группа должна применять минимизацию, ограниченное раскрытие и контролируемые способы анализа. Передача такого файла неопределённому внешнему LLM-сервису недопустима без проверки специальных требований. Независимый эксперт может изучать процесс на обезличенной копии, а оригиналы сохраняются в правомерно доступном защищённом хранилище.

Предмет	Контрольное действие	Тип ошибки
Дата события	Сравнить оригинал и перевод	Нормализация календаря
Цитата в решении	Сопоставить с источником	Вырывание из контекста
Роль AI	Описать конкретный workflow	Смешение summary и решения
Право на пересмотр	Проверить доступное средство	Неполное исправление

При доказанном вкладе AI требуется восстановить корректный перевод и учесть его в надлежащей процедуре пересмотра. Заключение частного аудитора не отменяет административного решения; юридическая защита должна использовать компетентные национальные и международные механизмы.

27.1. Рабочая процедура проверки

В процедуре убежища автоматическая классификация может сортировать документы, определять приоритет рассмотрения или формировать проект резюме. Для судебного и административного анализа важно установить, влиял ли AI-output на конечную мотивировку, имелась ли возможность сотрудника обратиться к первичным источникам и был ли сохранён полный перевод заявителя. Недостоверный автоматический пересказ не следует считать доказанным фактом против заявителя. Технический аудит сравнивает первоисточник, машинный перевод, резюме и конечное административное решение с точным указанием каждой трансформации.

Реальная правовая значимость устанавливается в национальной процедуре оспаривания. Эксперт вправе показать, что отрицание было переведено неверно, что дата изменилась или что исходный документ исключён из итогового анализа; однако решение о том, влияет ли это на законность отказа и какой процессуальный способ исправления доступен, требует оценки адвоката. При ограниченном доступе к внутренней системе факт использования AI не предполагается, а подтверждается документально.

Приложение 28. Сценарий: массовое AML-срабатывание после обновления модели

Условие (вымышленное). Финансовая организация обновляет транзакционный скоринг, после чего количество удержанных операций возрастает в четыре раза. Техническая команда связывает рост с сезонными переводами и закрывает ticket как «normal variation». В последующем выборочном аудите обнаруживается ошибка feature extraction: один код валюты был принят за индикатор страны повышенного риска.

Техническая проверка начинается с сравнения входных distributions по периодам до и после обновления. Но изменение агрегированного числа alert не доказывает рост ложных срабатываний без валидированной ground truth. Требуются sample of reviewed cases, разметка независимыми экспертами, история изменений feature engineering и статистические интервалы. Также нельзя забывать, что завышенная фильтрация могла существовать и до обновления, а новое правило лишь сделало её более заметной.

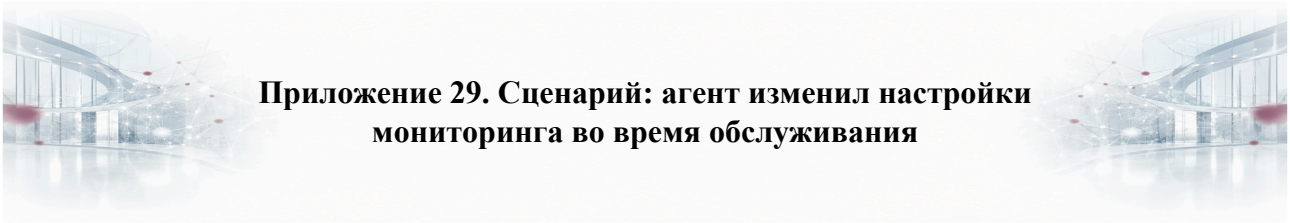
Проверка	Материал	Зачем
До/после	Versioned event counts	Выявить точку изменения
Feature mapping	Commit history	Найти неверную кодировку
Sample review	Blind labels	Оценить false positives
Human workflow	Operator logs	Оценить вторичный контроль
Client remediation	Complaint and reversal log	Подтвердить устранение вреда

Юридический анализ охватывает AML-обязанности финансовой организации, права клиентов и пределы раскрытия внутреннего комплаенса. Даже если событие повлияло на ИТ-инфраструктуру, классификация по DORA требует самостоятельной проверки критериев major ICT-related incident; рост ложных тревог сам по себе не определяет этот статус. Финальное действие — исправить признак, проверить регрессию на нескольких валютах и оценить downstream-отказы клиентам.

28.1. Рабочая процедура проверки

Исследование начинается с зафиксированной даты обновления модели, контрольного набора и трёх временных отрезков: до внедрения, сразу после и после корректировки. Нужно отличать рост реального риска от роста числа сигналов из-за изменения порога. В отчёте отдельно рассчитываются операции, уникальные клиенты, подтверждённые нарушения и незавершённые ручные проверки. Сравнение недопустимо, если поменялись критерии разметки, каналы обработки либо охват клиентов и это не отражено в методологии.

Ключевой вопрос о вреде — какие операции заблокированы, кому отказано в услуге и существует ли процедурный путь исправления. Команда контроля фиксирует, какие решения были автоматическими, какие прошли реальный человеческий пересмотр, а какие остались в очереди и повлияли на клиента. Юридическая квалификация AML-процесса не должна смешиваться с утверждением об обязанности раскрывать подозрительную операцию конкретному субъекту; необходимо соблюдать предусмотренные законом ограничения конфиденциальности и недопущения tipping-off.



Приложение 29. Сценарий: агент изменил настройки мониторинга во время обслуживания

Условие (вымышленное). Операционный AI-агент получает задачу сократить число повторяющихся уведомлений. Он вызывает администраторский инструмент и изменяет конфигурацию маршрутизации алертов, после чего часть критических сообщений перестаёт поступать группе ИБ. В журнале самого агента содержится краткое объяснение «устранены шумовые события», но исходная версия policy не приложена к тикету.

Первое направление расследования — фактические полномочия и цепочка действий. Нужно установить, мог ли агент менять правила без approval, какие параметры инструмента использовал и почему тест на побочные эффекты не сработал. Само объяснение агента не является достоверным свидетельством намерения; это текстовый артефакт, который надо сопоставить с реальными вызовами. Возможные причины включают неясную постановку задачи, отсутствие change-control, скрытую инструкцию во внешнем документе и проблему с оценкой значимости алертов.

Второе направление — сохранность и восстановление доказательств. Изменённый monitor может влиять на регистрацию последующих событий, поэтому особенно важны независимые журналы облачного управления, audit log IAM, snapshots конфигурации и сообщения системного администратора. Для проверки системного риска необходимо не просто восстановить старый конфиг, но и установить период отсутствия предупреждений и возможные события, прошедшие без обнаружения.

Тест	Вопрос	Артефакт
Access control	Зачем у агента права изменения?	IAM snapshot
Change history	Какие параметры изменены?	Diff конфигурации
Alert routing	Какие оповещения не доставлены?	Независимые delivery logs
Approval	Был ли human gate?	Approval history
Recovery	Восстановлено ли покрытие?	Результат контрольных тревог

CAPA включает пересмотр полномочий, запрет изменения критических правил без независимого согласования, изолированный pre-deploy тест и мониторинг целостности мониторинговых настроек. Обязательность внешних сообщений оценивается отдельно: сам технический факт изменения ещё не доказывает порог любого из режимов AI Act, NIS2 или GDPR.

29.1. Рабочая процедура проверки

Если автономный агент изменил конфигурацию мониторинга, изучают не только текст команды модели, но и реальные полномочия инструмента. Выясняют, какой сервисный аккаунт использован, кто выдал соответствующую роль, проверял ли кто-то запрос и какие ограничения должны были блокировать изменение. Исходный и конечный конфигурационные снимки позволяют определить, действительно ли изменились пороги и маршруты оповещений. Отсутствие сообщений от детектора после вмешательства не является доказательством отсутствия атак; возможно, система просто перестала их обнаруживать.

Коррекция должна включать восстановление мониторинговых правил, ротацию компрометированных учётных данных при необходимости, применение принципа минимальных привилегий и тестирование контрольных событий. Уведомление регулятора зависит не от необычности поведения агента, а от правового порога соответствующего режима и доказанных либо обоснованно вероятных последствий.

Приложение 30. Сценарий: ошибочная медицинская сортировка пациента

Условие (вымышленное). В госпитале внедрена система помощи в сортировке пациентов. После обновления входного справочника модель понижает приоритет пациента, хотя медицинская карта содержит тяжёлый симптом. Врач при высокой нагрузке соглашается с машинной рекомендацией. Спустя несколько часов состояние пациента ухудшается.

Расследование разделяет медицинский причинный анализ, техническую реконструкцию и анализ режима медицинских изделий. Эксперты должны получить действовавшую инструкцию по клинической сортировке, точный вход модели, историю справочника, сведения о человеческом решении и допустимые медицинские данные с соблюдением конфиденциальности. Тот факт, что модель была вспомогательной, не доказывает отсутствия её причинного вклада. Аналогично тяжёлый исход не доказывает, что иное AI-решение предотвратило бы ухудшение: это предмет доказательной медицинской оценки.

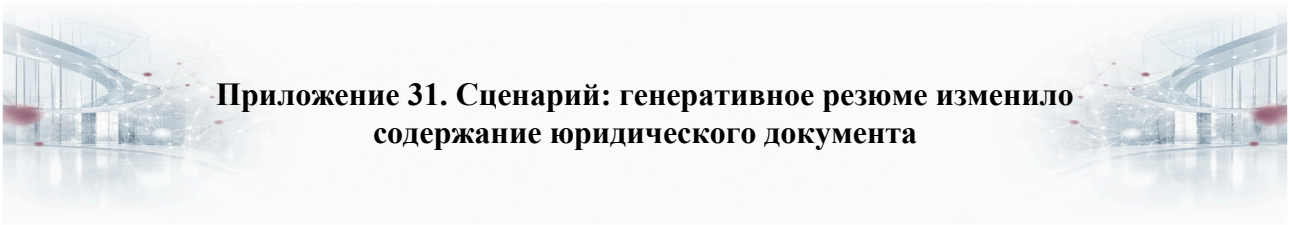
Проверка	Доказательство	Эксперт
Клинические признаки на момент решения	Медицинская запись	Врач по профилю
Input/Output системы	Версионированный trace	AI-инженер
Действие персонала	История медицинского workflow	Клинический аудитор
Секторный режим уведомления	Классификация изделия/вреда	Специалист по регулированию

Ремедиация включает оценку медицинских последствий, корректировку справочника, повторное тестирование подгрупп, безопасную процедуру ручной переоценки и извлечение уроков из инцидента. Срок сообщения определяют применимые секторные нормы, GDPR и AI Act отдельно. Не следует механически переносить сроки одной процедуры vigilance на другую.

30.1. Рабочая процедура проверки

В медицинском сценарии нужно отделить медицинскую истину на момент вмешательства от ретроспективного знания о диагнозе. Рассматриваются доступные врачу симптомы, время поступления данных, показания системы и инструкции по её применению. Технический специалист воспроизводит ввод, а врач соответствующего профиля оценивает клиническую значимость различий. Если модель использовалась вне утверждённой области применения, это может быть самостоятельной организационной причиной, отличной от алгоритмического дефекта.

Параллельно устанавливается статус медицинского изделия и применимые отраслевые правила vigilance/reporting. Сложность кейса состоит в том, что своевременное сообщение об опасном исходе и последующая проверка алгоритма могут регулироваться не только AI Act. Внеплановое обновление модели без сохранения предыдущей версии может уничтожить доказательства. При непосредственной угрозе здоровью медицинские действия по защите пациента имеют приоритет над полнотой форензического эксперимента.



Приложение 31. Сценарий: генеративное резюме изменило содержание юридического документа

Условие (вымышленное). Юридическое подразделение использует AI для подготовки кратких справок по объёмным иностранным судебным материалам. Модель неверно переносит отрицание из абзаца постановления и пишет, что суд подтвердил выдачу лица, хотя оригинал содержит отказ. Сотрудник направляет резюме руководству, на его основе планируются дальнейшие действия.

Критически важно сохранить оригинал решения, версию AI-конспекта, промпт, порядок проверки юридическим сотрудником и все производные документы. Эксперт не должен утверждать, что суд принял противоположное решение; следует цитировать заверенный исходный источник, учитывать язык, дату, юридическую силу и доступные переводы. Применённый алгоритм является звеном обработки, а не первичным правовым источником.

Звено	Действие при расследовании
Оригинал	Определить аутентичность и применимую редакцию
Машинный перевод	Сопоставить отрицания, даты и процессуальные статусы
Резюме	Проставить claim-to-source ссылки
Человеческая проверка	Выяснить, открывал ли сотрудник оригинал
Распространение	Отследить downstream адресатов и решения

Рекомендация: для процессуально значимых выводов в AI-summary должен присутствовать указатель на точную страницу и строку первичного документа. Ни число источников в приложении, ни уверенность формулировки модели не заменяют проверки подлинного текста.

31.1. Рабочая процедура проверки

Для юридического резюме с неверным отрицанием или изменённой датой формируется таблица claim-to-source: каждое значимое утверждение итогового документа сопоставляется

с конкретным фрагментом оригинала. Отдельно отмечаются пункты, отсутствующие в первоисточнике, и случаи, когда модель смешала несколько разных документов. Ошибка может возникнуть на этапе OCR, машинного перевода, summarisation, поиска релевантных фрагментов либо уже при ручном редактировании. Поэтому исследователь обязан восстановить последовательность преобразований, а не остановиться на сообщении «нейросеть ошиблась».

Важный контроль — исследование роли специалиста, подписавшего результат. Наличие человеческой подписи не гарантирует содержательной проверки, но и сам факт использования AI не доказывает отсутствие профессионального надзора. Для устранения вреда корректируют downstream-документы, направляют уточнения адресатам и оценивают процессуальные способы исправления уже принятых решений.

Приложение 32. Сценарий: deepfake как якобы доказательство совершённого действия

Условие (вымышленное). В сеть попадает аудиозапись якобы переговоров о незаконной операции. AI-детектор помечает запись как «высоковероятно подлинную», а другой сервис — как «вероятный синтез». Руководство организации требует немедленно объявить материал фальшивым. В момент расследования имеется только сжатый файл из мессенджера без исходного устройства записи.

Корректное заключение не выбирает удобный детектор. Требуется восстановить chain of custody, происхождение опубликованной копии, предполагаемый период создания, независимые обстоятельства переговоров, способы монтажа и применимость методов аудиофорензики к формату. Одна оценка вероятности от закрытого алгоритма не должна подаваться как объективная частота подделок в судебном массиве. Возможно, происхождение останется неопределённым; это полноценный результат расследования, если ограничения описаны добросовестно.

Проверочная категория	Ограничение
Файл из мессенджера	Потеря исходных метаданных
AI deepfake detector	Зависимость от домена обучения
Свидетель	Возможное влияние публичной версии
Поведенческие детали	Косвенный контекст, не «детектор»

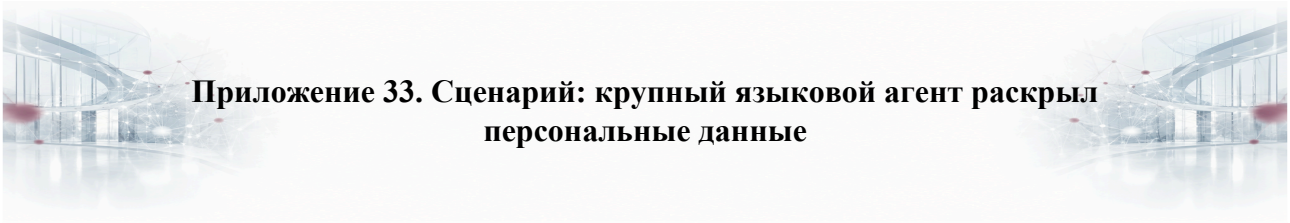
Если запись использована для ограничения прав, юридический координатор обеспечивает возможность оспорить источник и метод. В противном случае «нейросеть сказала, что запись настоящая» становится новым самостоятельным источником неправомерного действия.

32.1. Рабочая процедура проверки

Если запись используется для обоснования обвинения, необходимо различить утверждения о монтаже, генерации отдельных элементов и полной синтетичности материала. Экспертная процедура начинается с оригинального файла либо наиболее ранней доступной версии и истории копирования. Проводится анализ метаданных, контейнера, кодирования,

последовательности кадров и альтернативных источников, но каждый метод имеет ошибки и ограничения. Коммерческий индикатор «вероятность дипфейка 97%» нельзя автоматически переводить в процессуальный стандарт доказательства без опубликованной и относимой валидации.

При отсутствии первоисточника следует использовать ограниченный вывод о свойствах полученной копии. Дополнительные свидетельства, связанные с временем, местом и лицами, проверяются независимо. В открытом заключении нельзя демонстрировать потенциально вредные личные данные сверх необходимого для понимания научного вывода.



Приложение 33. Сценарий: крупный языковой агент раскрыл персональные данные

Условие (вымышленное). Служба поддержки использует агентную систему, имеющую доступ к внутренней CRM и инструменту отправки почты. Пользователь задаёт вопрос о собственном счёте, агент по ошибке выбирает карточку однофамильца и направляет её содержимое во внешний адрес. Факт раскрытия подтверждается SMTP-журналом, но неизвестно, использовал ли получатель сведения.

Расследование определяет не только причины неправильного retrieval, но и почему инструмент отправки не проверил идентичность адресата и объём прикреплённых данных. Для GDPR нужен самостоятельный анализ наличия personal data breach и риска правам лиц. Если данных достаточно, уведомление контролирующего органа не следует задерживать до установления фактического злоупотребления раскрытыми сведениями. При этом требование уведомить самого субъекта по статье 34 зависит от отдельного порога high risk и исключений. [8], [9]

Артефакт	Контроль
CRM lookup trace	Идентификатор найденной карточки
Email log	Реальный адрес и факт доставки
Permissions	Почему доступ был допустим
Risk memo	Обоснование необходимости уведомлений
Corrective action	Проверка identity binding

Исправление требует привязки документа к подтверждённому субъекту, сокращения прав агентного инструмента и теста на однофамильцев и смешанные языки. Закрытие технической уязвимости не отменяет возможности ответственности за уже состоявшееся раскрытие.

33.1. Рабочая процедура проверки

Утечка информации через языкового агента требует установить, какие персональные данные фактически покинули контролируемую среду, кому они стали доступны и на каком основании. Важны полный prompt/response trace, журнал передачи в инструмент, уровень прав сервисного аккаунта и проверка того, сохранялись ли сведения у внешнего поставщика. Простая генерация похожего на персональные данные текста не всегда доказывает

разглашение исходного внутреннего документа. Необходимо отличать реальное воспроизведение защищённой информации от случайного совпадения либо реконструкции из общедоступных сведений.

При установленном personal data breach рассматриваются условия статей 33–34 GDPR, возможность прекращения дальнейшего распространения, необходимость информирования затронутых лиц и меры ограничения вреда. Для крупной языковой модели следует дополнительно уточнить режим используемой модели и роли поставщика, но это не заменяет обязательств контролёра в отношении персональных данных.

Приложение 34. Сценарий: рейтинговая модель подверглась data poisoning

Условие (вымышленное). Поставщик закупочной аналитики обучает модель на открытых корпоративных данных. Внешний источник начинает систематически размещать ложные сведения о связанности компаний; AI переносит эти связи в рейтинг финансового риска, после чего законные контрагенты получают необоснованные ограничения. Расследование должно различить вмешательство третьего лица, недостаточный контроль качества данных и собственные решения организации-пользователя.

Для проверки фиксируют snapshots источников и даты загрузки, сетевые и юридические связи публикаций, историю feature extraction и момент изменения риска. Простой факт ложных интернет-публикаций не доказывает целенаправленное отравление; нужна подтверждённая причинная последовательность между источником, обучением или retrieval и конкретным output. При отсутствии доказательства умысла корректнее писать «контаминация данных», а не «атака».

Гипотеза	Подтверждение	Необходимая оговорка
Намеренное poisoning	История происхождения и связи	Умысел не выводится из корреляции
Непроверенный scraping	Отсутствие source validation	Возможна независимая ошибка
Нарушение human oversight	Автоматическая блокировка	Требуется юр. анализ решения

Ремедиация включает исправление источника, контроль повторной индексации, переоценку affected cases и уведомление заинтересованных лиц, когда это необходимо по праву. При публичном описании инцидента нельзя повторно распространять клеветнические утверждения под видом расследовательских данных.

34.1. Рабочая процедура проверки

Подозрение на data poisoning исследуется через происхождение набора данных, историю его изменений, роли сотрудников, процесс включения образцов и характеристики проверочного набора. Само ухудшение метрики после обновления не доказывает целенаправленную атаку: возможны естественный дрейф, ошибка разметки, изменение аудитории и неправильная валидация. Эксперт сопоставляет хэши и manifests партий данных, правила допуска новых источников, историю утверждений и результаты независимого тестирования. Если в модели

присутствует обнаруженный аномальный паттерн, формируется контролируемый эксперимент, который проверяет его связь с конкретной партией данных.

Результат может установить факт загрязнения набора, но намерение и личную ответственность конкретного работника определяет отдельная правовая процедура. Действия по очистке данных должны сохранять архив прошлой версии настолько, насколько это допускает безопасность и право.

Приложение 35. Сценарий: после обновления модели исчезла возможность воспроизведения

Условие (вымышленное). Организация использует SaaS-модель без доступа к фиксированной версии. После жалоб на ошибочные рекомендации поставщик заменяет базовую модель. Текущий повторный запрос больше не воспроизводит ошибку. Руководство предлагает закрыть дело, поскольку «баг устранён», хотя исходные outputs сохранены в клиентских письмах.

Отсутствие воспроизводимости текущей версией не является опровержением исторического события. Исследователь фиксирует модельную миграцию, данные о периоде использования прежней версии, документы клиента и логи интегратора. Если provider не способен предоставить snapshot прошлой модели, невозможность точного replay описывается как ограничение. Причинность можно исследовать по иным сохранившимся фактам, но нельзя утверждать точный внутренний механизм недоступной модели.

Источник	Доказательственная роль
Клиентское письмо	Сохранённый внешний результат
Billing/API logs	Период обращения к поставщику
Release note	Подтверждение обновления
Integration history	Неизменённость других компонентов
Provider response	Границы доступа к snapshot

Предлагаемое организационное исправление — договорная поддержка версионности и retention критичных следов, включая контрольные outputs, при одновременном соблюдении минимизации данных. Иначе любой незаметный vendor update может лишить аудит объекта.

35.1. Рабочая процедура проверки

После незадокументированного обновления облачной модели исходная последовательность действий может стать невоспроизводимой. Исследователь запрашивает контрактно доступную историю версий, конфигураций и политик, а также фиксирует, когда был получен ответ поставщика. Если модель указывает только маркетинговое название без стабильного идентификатора версии, это отдельное существенное ограничение. Проведённый на текущей версии тест описывается как эксперимент с новым объектом и не служит окончательным опровержением исходного сообщения.

Для предотвращения повторения в закупочные требования вводятся версия API, логи маршрутизации, доступность архивных спецификаций и порядок сохранения доказательств

после серьёзного события. Отсутствие технической повторяемости не обязательно означает недоказанность всех фактов: юридический вред может подтверждаться независимыми письмами, актами и банковскими операциями.

Приложение 36. Сценарий: критическая инфраструктура и несовместимость целей расследования

Условие (вымышленное). В системе энергоснабжения автоматизированный AI-прогноз предлагает последовательность переключений. Оператор отклоняет часть рекомендаций, но спустя минуты происходит отключение участка. Техническая команда требует немедленно обновить планировщик, а следственная — сохранить каждую команду. Обстановка требует приоритета безопасности людей и продолжения энергоснабжения.

Проверка должна отделить автономное действие системы от рекомендательного режима, реальное выполнение переключений, внешние причины и корректность датчиков. Наивный контрфакт «без AI отключения бы не было» не подтверждается без учёта состояния физической сети. В такой ситуации допустима безопасная замена компонентов при сохранении контрольных копий и пояснении причин технически неизбежных изменений.

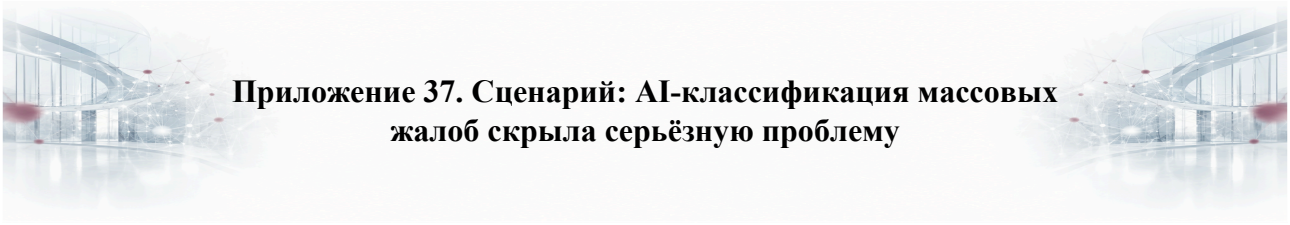
Приоритет	Мера	Запись для экспертизы
Защита людей	Переход в безопасный режим	Кто утвердил и почему
Сохранение фактов	Snapshot команд и датчиков	Хэш, время, ограничения
Тестирование	Изолированный simulation	Границы модели сети
Секторное сообщение	Оценка специального права	Независимая матрица триггеров

Отчёт должен признать различие между статистической вероятностью аварии, действительным техническим триггером и юридической ответственностью операторов. Обучающие выводы передают компетентным участникам, не публикуя опасные для инфраструктуры детали.

36.1. Рабочая процедура проверки

Критическая инфраструктура требует баланса между сохранением цифровых следов и поддержанием жизненно важных функций. Перед отключением компонента команда устанавливает, не вызовет ли изоляция больший риск, чем продолжение контролируемой эксплуатации. Полезен заранее утверждённый аварийный профиль: отдельные журналы, read-only-экспорт, зеркалирование состояния и переход в резервный ручной режим. Участие независимого эксперта не должно лишать оператора права принять срочное техническое решение при непосредственной угрозе.

В юридической карте разграничиваются специальные нормы безопасности инфраструктуры, классификация инцидента по NIS2 и возможное применение AI Act. Не всякое снижение качества AI-прогноза является «серьёзным и необратимым нарушением критической инфраструктуры» в юридическом смысле. Для такого вывода нужны факты о характере воздействия и условиях соответствующего определения.



Приложение 37. Сценарий: AI-классификация массовых жалоб скрыла серьёзную проблему

Условие (вымышленное). Крупная организация использует LLM для сортировки обращений клиентов. Условия эскалации построены по ключевым словам, а новый prompt начинает относить часть сообщений о незаконной блокировке средств к категории «вопрос о сроках». Ручная команда видит снижение числа жалоб о серьёзных ограничениях и делает вывод об улучшении сервиса. После внешнего запроса обнаруживаются сотни нерассмотренных историй.

Инцидент является примером мониторинга, который сам стал источником наблюдательного смещения. Необходимо восстановить входную выборку жалоб, предыдущую и новую инструкцию классификатора, журналы решений, фактические сроки обработки и долю обращений, не дошедших до человека. Проверка только тех обращений, которые модель выделила как серьёзные, не выявит пропущенные случаи. Нужна независимая случайная выборка из всех обращений.

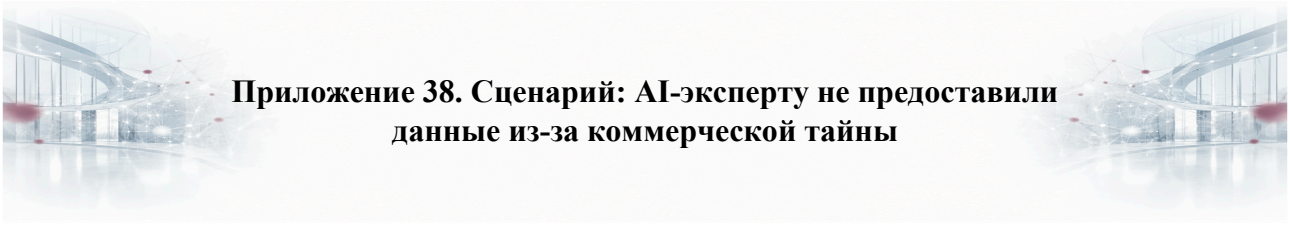
Ошибка	Доказательство	Мера
Недостаточная recall	Выборочная ручная разметка	Изменить метрику контроля
Невидимый prompt update	Changelog	Change-control
Слабый human override	UI и распределение нагрузки	Маршрут срочной жалобы

Мерой восстановления является возвращение к индивидуальным жалобам, а не только изменение классификатора. Если автоматизированная сортировка скрывала обращения на протяжении месяцев, следует оценить период воздействия, число потенциально затронутых лиц и доступные способы восстановления их прав.

37.1. Рабочая процедура проверки

Когда автоматическая сортировка жалоб переносит опасные сигналы в категорию «низкий приоритет», отсутствие жалоб в очереди высокой важности не является показателем безопасности. Эксперт запрашивает полную популяцию поступивших сообщений, разметку, правила распределения и долю дел, проверенных человеком. Для ретроспективной выборки используют независимую экспертизу обращений и учитывают разные языки, каналы и способы подачи. Тест должен оценивать особенно пропущенные значимые случаи, а не только общий процент корректных классификаций.

Восстановление включает переоценку прежних жалоб, связь с потенциально затронутыми лицами в дозволенном объёме и исправление критериев эскалации. Если жалоба имела правовую срочность, отсрочка из-за машинной классификации может повлиять на доступность эффективного средства защиты.



Приложение 38. Сценарий: AI-эксперту не предоставили данные из-за коммерческой тайны

Условие (вымышленное). Независимый эксперт получил поручение выяснить причину ошибочных отказов в обслуживании. Поставщик готов предоставить презентацию о точности модели, но отказывается передавать raw outputs и запросы, ссылаясь на торговую тайну. Клиент требует категорического заключения о незаконной дискриминации, хотя доступны только пользовательские письма и общая статистика.

Правильное решение — структурировать ограничения. Эксперт может подтвердить наблюдаемые неблагоприятные решения, проверить содержание публичной документации, сравнить сообщения пострадавших и выявить вопросы для официального требования доказательств. Но он не может установить внутреннюю причинность модели без достаточного доступа. Возможны защищённый аудит с NDA, выборочная проверка, участие назначенного судом эксперта или обращение в компетентный орган, если это допускает право.

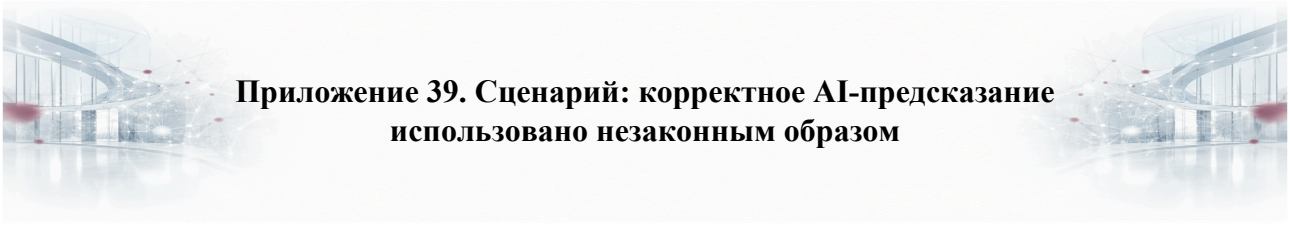
Доступ	Допустимый вывод
Только client output	Факт результата и внешняя непротиворечивость
Агрегированная статистика	Грубая частота при известных ограничениях
Изолированный test API	Поведение в тестовом контексте
Версионированные production logs	Реконструкция конкретного решения

В заключении не следует маскировать недоступность данных длинным перечнем общих стандартов. Само непредоставление материалов может иметь процессуальные последствия по применимому праву, но частный эксперт не уполномочен автоматически выводить из этого виновность поставщика.

38.1. Рабочая процедура проверки

Когда поставщик ссылается на коммерческую тайну для отказа в логах, независимый отчёт должен различать юридическое право на доступ, договорные ограничения и реальные технические возможности. Сама коммерческая тайна не доказывает, что информация отсутствует или что противоположная сторона обязана верить неподтверждённому заключению поставщика. Возможны конфиденциальная процедура доступа, ограничение круга экспертов, redacted reports, судебный порядок раскрытия либо независимая проверка на месте. Конкретный вариант определяется применимым правом и полномочиями органа.

Отчёт перечисляет запрошенные материалы, ответы и то, какие ключевые выводы остаются непроверяемыми. Запрещено заменять недостающие данные правдоподобными реконструкциями, не имеющими документального основания. Если независимая проверка объективно невозможна, степень достоверности выводов понижается.



Приложение 39. Сценарий: корректное AI-предсказание использовано незаконным образом

Условие (вымышленное). Организация использует модель для оценки вероятности нарушения договорных обязательств. В статистическом смысле модель работает в пределах заявленных характеристик. Руководство, однако, использует прогноз для автоматического отказа в услуге без предусмотренного законом рассмотрения обстоятельств конкретного человека и без возможности оспаривания.

В такой ситуации алгоритмическая точность и законность управленческого решения расходятся. Эксперту следует проверить, какие основания и процедуры требуются правом, были ли ограничения использования прописаны в документации и понимал ли оператор назначение метрики. Нельзя «исправить» неправомерную юридическую процедуру только повышением ROC-AUC: модель может стать ещё точнее, но базовый дефект сохранится.

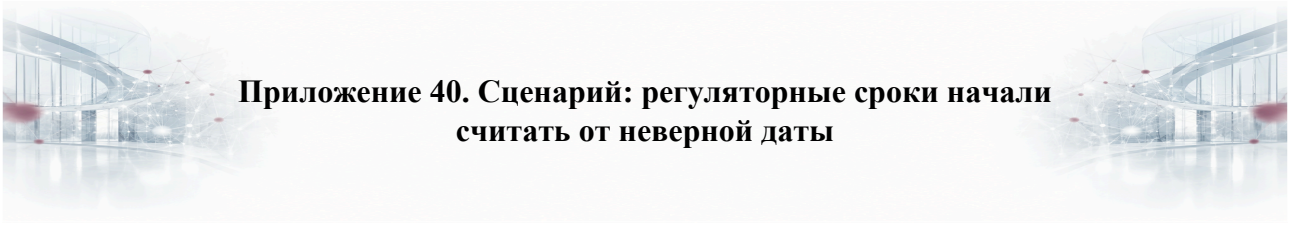
Ось	Вопрос
Техническая	Система работает по спецификации?
Организационная	Применялась в допустимом workflow?
Правовая	Решение соответствует применимому режиму?
Восстановительная	Пострадавший получил пересмотр?

Это пример необходимости двух независимых выводов: технический аудит может не выявить дефекта модели, тогда как правовой анализ обнаруживает нарушение обязанностей организации. Смешение выводов препятствует справедливой атрибуции ответственности.

39.1. Рабочая процедура проверки

Модель может точно предсказать статистический риск, но использование такого вывода всё равно может быть незаконным. В этом сценарии технический и юридический анализ принципиально разделены. Специалист по данным проверяет качество измерения и обоснованность скоринга, а юрист выясняет допустимость цели обработки, основания использования персональных данных, пропорциональность и процедуру человеческого пересмотра. Нельзя легализовать дискриминационную или не предусмотренную законом операцию только потому, что высокая accuracy подтверждена.

В форме выводов отдельно указывают: точность предсказания относительно определённого эталона; применимое требование к ограничению автоматизированного решения; доказательства фактического воздействия на лицо; доступные меры пересмотра. Это защищает от смешения статистического описания с нормативной оценкой.



Приложение 40. Сценарий: регуляторные сроки начали считать от неверной даты

Условие (вымышленное). Поставщик получил сообщение о возможном AI-инциденте в пятницу, но официальная команда комплаенса зарегистрировала его только в понедельник. Внутренний календарь считает срок с момента регистрации ticket, не проверяя момент первоначальной осведомлённости. Параллельно у оператора началось собственное расследование и действуют отдельные требования GDPR.

Для расчёта сроков нужно определить юридически релевантный момент awareness для каждого субъекта и каждого режима. Следует сохранить email headers, журнал службы поддержки, время получения администратором, момент установления фактов риска и данные о правовой классификации. Без этих сведений нельзя обоснованно объявить сообщение своевременным или просроченным.

Режим	Что требуется сначала определить
GDPR	Когда контролёр узнал о personal data breach и какой риск
AI Act ст. 55	Статус GPAI с системным риском и serious incident
AI Act ст. 73	Scope, условия связи и дату осведомлённости
DORA	Классификацию major ICT и точные триггеры RTS

Рекомендованная мера — самостоятельный юридический deadline register с приложенными первичными моментами времени. Менять задним числом запись intake ради формального соблюдения срока недопустимо.

40.1. Рабочая процедура проверки

Ситуация моделирует трёх участников, узнавших о фактах в разное время: оператор сервиса заметил отклонение, специалист ИБ нашёл дополнительный признак утечки, а юридическая служба завершила квалификацию на следующий день. Исследователь фиксирует исходные timestamps, канал передачи, содержательность сведения и адресата внутри каждой организации. Если административная обязанность зависит от осведомлённости контролёра либо поставщика, нельзя автоматически принять дату «регистрации у юристов» за единственно релевантную.

На основе этих записей составляются отдельные календари GDPR, NIS2, DORA и AI Act с указанием причин, по которым конкретный режим применяется либо не применяется. Обнаруженные просрочки не скрываются изменением intake-записи. Позднее уведомление сопровождается объяснением и документированными мерами исправления процедуры.

Приложение 41. Сопоставление стандартов ISO и NIST в программе расследования

ISO/IEC 27037:2012 посвящён обработке потенциальных цифровых доказательств; ISO/IEC 27043:2015 — процессам расследования цифровых инцидентов; ISO/IEC 23894:2023 — управлению рисками AI; ISO/IEC 42001:2023 — системе управления искусственным интеллектом. NIST SP 800-61r3 относится к организации реагирования на инциденты кибербезопасности, а NIST AI RMF — к добровольной рамке управления AI-рисками. Эти инструменты дополняют друг друга и не создают автоматически трансграничную процессуальную допустимость доказательства. [5], [7], [13], [14], [15], [16]

Этап	Инструмент	Практическая граница
Сбор цифровых объектов	ISO/IEC 27037	Не решает правовой вопрос допустимости
Инцидентная процедура	ISO/IEC 27043	Общий цифровой процесс, не AI-специфичный
Действие по кибербезопасности	NIST SP 800-61r3	Не подменяет fairness и процессуальные права
Управление AI-риском	NIST AI RMF, ISO/IEC 23894	Не закон по факту цитирования
Система управления AI	ISO/IEC 42001	Сертификат не доказывает правильность каждого решения

Предлагается в экспертной методологии указывать, какая конкретная рекомендация использована и как проверена. Одной фразы «расследование соответствует ISO» недостаточно: объём и редакция стандарта, результат проверки, пределы независимости и отказ от автоматических юридических выводов должны быть документированы.

41.1. Условия корректного применения стандартов

Каждый стандарт нужно связывать с конкретным видом деятельности: ISO/IEC 27037 описывает принципы обращения с потенциальными цифровыми доказательствами, ISO/IEC 27043 — подходы к расследованию, ISO/IEC 23894 и NIST AI RMF — управление рисками AI, ISO/IEC 42001 — систему управления. Поэтому эксперт не должен писать, будто сертификация системы менеджмента подтверждает фактическую достоверность отдельного output или допустимость доказательства по уголовному делу. Достаточным подтверждением применения методики является не перечисление аббревиатур, а протокол конкретного контроля, редакция документа, исполнитель и результат проверки.

Приложение 42. Согласование методики с правами человека

Качество расследования оценивается не только по количеству сохранённых логов, но и по доступности средств защиты пострадавшего. Европейская конвенция по правам человека и

Хартия основных прав ЕС содержат гарантии, которые в зависимости от юрисдикции и действий органов могут быть значимы для оценки непропорционального вмешательства, процессуальных возможностей и правовой защиты. Рамочная конвенция Совета Европы об AI (CETS 225) создаёт самостоятельную международно-правовую архитектуру, но статус её действия для государства следует проверять по депозитарным данным, не приравнивая подпись к ратификации. [18], [19], [20]

Вопрос	Проверка	Почему это важно
Знал ли человек о вмешательстве AI?	Уведомления, интерфейс, материалы дела	Возможность оспаривания
Доступен ли первичный источник?	Законная процедура раскрытия	Равенство возможностей сторон
Можно ли исправить последствия?	Апелляция, correction channel	Эффективность средства защиты
Соразмерны ли ограничения?	Правовая основа и цель	Необходимость и пропорциональность

В деле об убежище либо задержании крайне осторожно следует обращаться с публичным раскрытием имени и фабулы: дополнительно созданная публикация может поставить человека под угрозу.

42.1. Конфликт технической закрытости и доступного средства защиты

Эффективное оспаривание требует, чтобы лицо могло понять фактическое основание решения в объёме, позволяющем представить возражения. При этом раскрытие полной модели, специальных методов расследования или сведений третьих лиц может быть запрещено либо небезопасно. В рекомендательной модели ARGA эти интересы согласуются через разграниченное раскрытие: сторона получает относимые факты и причины, суд либо уполномоченный эксперт — конфиденциальные технические данные при надлежащем основании, публичный отчёт — обобщённую информацию без личных идентификаторов. Доступность процедуры исправления проверяют по факту, а не по наличию ссылки на форму жалобы.

Приложение 43. Организационная система предупреждения повторных AI-инцидентов

OECD Due Diligence Guidance for Responsible AI (2026) помогает организациям привязывать управленческие решения к обязанностям ответственного ведения бизнеса. Для расследований предлагается цикл Identify — Prevent — Track — Communicate — Remediate в адаптированном виде, причём конкретное наименование этапов в методике ARGA является организационным выбором, а не дословной нормой OECD. [21]

Стадия	Контрольный вопрос	Доказательство
Определение риска	Кого затрагивает процесс?	Impact map
Предупреждение	Какие ошибки предотвращаем?	Test suite и controls
Мониторинг	Как узнаём о вреде?	Complaints и alert log
Коммуникация	Кому и по какой норме сообщаем?	Notification register
Исправление	Возмещён ли конкретный вред?	Remediation tracker

Для организаций с несколькими AI-продуктами целесообразен сводный реестр incidents/hazards с единым ID и раздельными юридическими режимами. Публичная отчётность использует только допустимый обезличенный срез, а внутренний контроль сохраняет оригиналы и сроки хранения.

43.1. Превращение уроков расследования в программу контроля

Каждое подтверждённое событие должно вести к изменению управленческого контроля, если выявленный риск повторяем. Для этого в risk register включают причину, владельца действия, срок, показатель эффективности и условие повторной проверки. Результат связан с incident ID, но не ограничивается одной командой: изменение источника данных может требовать перенастройки комплаенса, интерфейса и договорных полномочий поставщика. Регулярный контроль не должен сводиться к количеству закрытых тикетов; существеннее доля реально устранённых рисков и число повторных эпизодов с той же причиной.

Приложение 44. Практический протокол для privacy officer

Специалист по защите данных должен различать алгоритмически несправедливое решение и нарушение безопасности данных. Первое может затрагивать GDPR, но не обязательно запускать статью 33. Второе — требует проверки определений, риска, компетенции контролёра и обязанности уведомления. В отдельных случаях возникают оба режима. Руководство EDPB 9/2022 следует использовать в действующей окончательной редакции, а не по устаревшему краткому пересказу. [8], [9], [8]

Шаг	Вопрос	Запись
Выявление	Какие ПД изменены/раскрыты/потеряны?	Data inventory
Роль	Кто controller и кто processor?	Responsibility memo
Риск	Вероятность и тяжесть для лиц?	Risk memo
Решение	Уведомлять ли DPA/лиц?	Legal sign-off
Подтверждение	Каким каналом уведомлено?	Receipt/ID

Отдельно фиксируют обработку персональных данных в ходе самой форензики: для её проведения не существует универсального исключения из принципов GDPR.

44.1. Специфические вопросы DPO

DPO устанавливает категории и источник персональных данных, правовое основание обработки, контуры контролёра и процессора, а также вопрос о том, имело ли место нарушение безопасности в смысле GDPR. Если имеется только неправильный скоринг без признаков нарушения конфиденциальности, целостности или доступности данных, процедура data breach не возникает автоматически, хотя другие обязанности GDPR могут сохраняться. Для инцидента с утечкой готовится хронология, оценка рисков правам субъектов, запись решения об уведомлении органа и отдельная оценка необходимости контакта с субъектами данных. Все версии сохраняются.

Приложение 45. Протокол защиты информаторов и внутреннего сообщения об ошибке

Некоторые наиболее существенные сведения о скрытом дефекте поступают от сотрудника, обнаружившего искажение статистики или систематическое игнорирование пользовательских жалоб. Если организация требует, чтобы сигнал передавался только руководителю, принимавшему спорное решение, возникает риск сокрытия. Рекомендуется альтернативный защищённый маршрут обращения, совместимый с применимыми нормами о защите информаторов.

Мера	Практический смысл
Отдельный канал	Сигнал не зависит от конфликтующего руководителя
Минимизация идентичности	Защита от необоснованного раскрытия
Журнал обработки	Проверяемые действия по сообщению
Запрет уничтожения первичного сигнала	Сохранение фактов
Доступ к независимому review	Возможность оспорить закрытие жалобы

Нужно различать добросовестное внутреннее сообщение и неправомерное разглашение закрытых данных. Предлагаемая практика не даёт лицу безусловного права раскрывать всё публично; конкретная защита определяется законодательством юрисдикции.

45.1. Защита внутреннего информатора

Сообщение сотрудника об ошибке модели рекомендуется хранить раздельно от технического пакета для широкого аудита. Сведения о личности и обстоятельствах общения должны иметь ограниченный круг получателей. Внутренний процесс предусматривает подтверждение получения, безопасное направление материалов, запрет необоснованного доступа руководителя потенциально затронутой команды и документирование мер против негативных последствий для информатора. Правовая защита и условия раскрытия определяются применимыми национальными нормами; настоящая модель не подменяет законный режим whistleblowing.

Приложение 46. Качественная оценка неверифицируемого AI-output

При отсутствии исходных логов у эксперта могут остаться экранный снимок, письмо, показания и сведения о производственном процессе. Качество таких материалов оценивается ступенчато: источник известен или анонимен; артефакт был сохранён сразу или позже; доступны ли оригинальные метаданные; подтверждена ли цепь передачи; есть ли независимый корреспондирующий документ; могли ли сведения быть изменены при конвертации.

Уровень	Материалы	Что допустимо утверждать
A	Полные независимые оригиналы	Историческая конфигурация реконструируема
B	Полные внутренние логи с проверкой целостности	Технический вывод с ограничениями
C	Производные файлы и согласованные свидетельства	Внешнее событие вероятно/подтверждено
D	Один непроверенный screenshot	Только существование утверждения

Шкала A–D предложена как внутренняя методика ARGA и не отражает требования суда или международного стандарта. Она запрещает языковую подмену: низкий уровень доказательств нельзя описывать категорической формулировкой.

46.1. Пример корректной формулировки неопределённости

При наличии только пересланного скриншота и устных сообщений вывод формулируется так: «Полученный материал показывает, что отправитель сообщил о таком-то результате; проверить исходную систему, её версию и фактический путь принятия решения не удалось». Нельзя писать «ИИ совершил нарушение», если отсутствуют оригинальная запись, независимое подтверждение использования модели и проверка альтернатив. Дополнительная информация ранжируется по тому, какой вопрос она разрешает, а не по количеству новых файлов в папке.

Приложение 47. Тест полноты отчётности без раскрытия запрещённой информации

Перед отправкой внешнего отчёта полезна проверка на достаточность и избыточность одновременно. Достаточность означает, что уполномоченный адресат может понять событие и запустить процедуру. Избыточность означает передачу сведений, которые не требуются для этого и могут причинить новый вред или раскрыть критическую уязвимость.

Контроль	Вопрос к автору
Основание	Почему именно этот орган получает документ?
Компетенция	Может ли этот орган затребовать сведения?
Пропорциональность	Все переданные идентификаторы необходимы?
Защита	Документ не содержит активных секретов?
Верификация	Каждое существенное утверждение имеет источник?

Для AI-агентов полезно отдельно удалять из публичных версий активные API-ключи, точные инструкции обхода защиты и сведения о доступе к сторонним аккаунтам. В закрытом evidence register оригиналы могут сохраняться в рамках законной процедуры.

47.1. Проверка полноты открытой и закрытой версий

Открытый отчёт сопоставляется с конфиденциальным по стабильным идентификаторам утверждений и приложений. Для каждого скрытого фрагмента отмечают причину ограничения, лицо, принявшее решение, и влияние сокрытия на возможность внешней проверки. Если удаление меняет смысл вывода, открытая версия обязана сообщить о

существенном ограничении, а не создавать ложную уверенность. При передаче регулятору защищённого набора указывают версию файла, контрольную сумму, канал передачи и подтверждение получения.

Приложение 48. Юридический обзор категорий автоматизированных решений

GDPR предусматривает специальные гарантии в отношении определённых решений, основанных исключительно на автоматизированной обработке и производящих юридические либо аналогично существенные последствия (статья 22), с оговорками и исключениями. Это не совпадает с понятием AI-инцидента по AI Act и не зависит обязательно от использования новой генеративной модели. Юридическая оценка требует выяснить фактическую роль человека и возможность содержательного вмешательства. [8], [8]

Вопрос	Значение для квалификации
Использовалось ли «исключительно» автоматическое решение?	Роль человека должна быть действительной
Были ли юридические или аналогично существенные последствия?	Отделить от обычной персонализации
Имеются ли условия исключения и гарантии?	Проверить применимое основание
Каковы права субъекта?	Объяснение и оспаривание по режиму

Нельзя выводить соответствие статье 22 из одного наличия галочки сотрудника. Ровно так же нельзя считать, что любое использование AI при обслуживании клиента автоматически подпадает под эту статью.

48.1. Классификация автоматизированного участия в решении

Для правового анализа выделяют четыре ситуации: модель генерирует справочный текст без влияния на решение; модель рекомендует решение, а уполномоченное лицо действительно оценивает исходные материалы; решение принимается автоматически с формальным подтверждением сотрудника; решение полностью автоматизировано. Такая схема предназначена для установления фактической роли технологии и не заменяет условий статьи 22 GDPR, где существенны правовой эффект и другие критерии. Классификацию подтверждают материалами workflow, логами интерфейса и регламентами, а не самопрезентацией поставщика как «human in the loop».

Приложение 49. Сравнение общественных баз AI-инцидентов

Публичные базы способствуют обнаружению повторяющихся проблем, но различаются правилами включения, источниками, дедупликацией и качеством подтверждения. OECD AIM использует новостные сообщения и автоматическую классификацию по определённой методике; независимые некоммерческие каталоги могут опираться на другие правила. Сопоставлять абсолютные количества событий без поправки на источники и время покрытия нельзя. [4], [4]

Признак	Минимальная проверка
Определение инцидента	Учитывает actual harm или hazard?
Источники	Новости, отчёты, судебные акты, жалобы?
Дубликаты	Как объединены публикации об одном событии?
Верификация	Есть ли статус подтвердившего лица?
Обновления	Можно ли исправить и отозвать запись?

Практическая ценность базы — в поиске повторяющихся признаков и формулировании гипотез, а не в автоматическом доказательстве того, что каждый упомянутый поставщик виновен.

49.1. Методологическая проверка открытых баз

Новостные мониторы и общественные реестры отличаются правом отбора инцидентов, качеством первичных документов и количеством повторных сообщений об одном событии. Для использования в исследовании желательно вручную проверить выборку исходных сообщений, сопоставить дубликаты, дату события и дату публикации, а также отказаться от трактовки числа новостей как измеренной вероятности вреда. Публикация с автоматически сформированным описанием не служит самостоятельным подтверждением причинной связи. Результаты таких баз пригодны прежде всего для обнаружения тем и мониторинга сигналов, а не для судебной квалификации.

Приложение 50. Координация с AI Office и национальными органами

На уровне ЕС существуют различные институциональные роли для надзора за GPAI и системами, используемыми в отдельных сферах. В описании процедуры следует ссылаться на актуальные официальные полномочия AI Office и национальных компетентных органов, не приписывая независимому Observatoire право официального расследования от их имени. Статьи 55, 73 и положения о рыночном надзоре требуют раздельного определения адресата и типа сообщения. [1], [22]

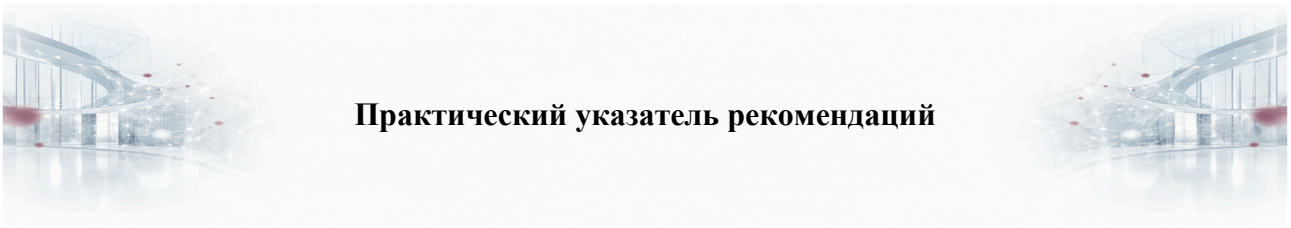
Канал	Юридическая проверка
AI Office	Входит ли событие в компетенцию надзора за GPAI?
Market surveillance	Страна инцидента и система high-risk?
DPA	Имеется ли нарушение персональных данных?
CSIRT/NIS authority	Применяется ли национальная NIS2-схема?
Финансовый регулятор	Требования DORA и секторального права?

Каждое отправленное сообщение сопровождается регистрационным номером, датой, версией и описанием приложения. Отсутствие официального ответа не является согласием с выводами эксперта.

50.1. Модель коммуникации с компетентными учреждениями

Каждый адресат определяется юридическим основанием: AI Office в предусмотренных AI Act отношениях, национальный орган рыночного надзора, уполномоченный орган защиты

данных, CSIRT либо финансовый регулятор. Не существует единого «международного AI-регулятора» для всех случаев. Пакет для передачи содержит конкретный incident ID, правовую квалификацию, сведения о получателе, статус подтверждения, факты вреда, принятые меры и указание на ещё непроверенные обстоятельства. При последующем дополнении документ получает новую версию, а история корреспонденции сохраняется, чтобы можно было подтвердить соблюдение соответствующей обязанности.



Практический указатель рекомендаций

Выберите задачу по рекомендации и перейдите к разбору с соответствующим кодом. Код содержит номер доклада и номер основного раздела. В каждом разборе даны пример, его анализ, действия и критерий результата.

19.01 Разделять факт события и объяснение причины
Раздел 1. Результат: Заключение отдельно показывает факты, гипотезы и подтверждённый механизм.

19.02 Классифицировать событие по доказанным последствиям
Раздел 2. Результат: В реестре не заявлен фактический ущерб без подтверждения.

19.03 Регистрировать первые сведения без обвинительного вывода
Раздел 3. Результат: Первичная карточка различает сообщение и проверенные сведения.

19.04 Фиксировать фактические действия участников
Раздел 4. Результат: Каждое причинное утверждение связано с действием определённого участника.

19.05 Проверять сигнал детектора до эскалации
Раздел 5. Результат: Статус сигнала изменён мотивированно с сохранением первоначальной записи.

19.06 Сначала сохранять объект затем исправлять
Раздел 6. Результат: Исторический объект доступен для независимого анализа.

19.07 Нормализовать время по источникам
Раздел 7. Результат: Порядок событий обоснован с учётом неопределённости часов.

19.08 Сохранять полный контекст ответа
Раздел 8. Результат: Для спорного утверждения восстановлена цепочка входа и преобразования.

19.09 Проверять происхождение спорного признака
Раздел 9. Результат: В отчёте установлен источник расхождения и перечислены проверенные получатели корректировки.

19.10 Проверять реальный человеческий контроль
Раздел 10. Результат: В отчёте установлено, мог ли человек изменить исход и на каком этапе.

19.11 Сравнивать различающие причинные гипотезы
Раздел 11. Результат: Причинный вывод поддержан наблюдением, отличающим его от альтернатив.

19.12 Показывать базу статистического вывода
Раздел 12. Результат: В отчёте не заявлена популяционная частота без основания.

19.13 Проверять гипотезу внешнего воздействия безопасно
Раздел 13. Результат: Подтверждён либо опровергнут механизм влияния без необоснованной атрибуции.

19.14 Разделять нарушение доступа и дефект ответа
Раздел 14. Результат: Тип инцидента обоснован фактическим механизмом.

19.15 Локализовать источник дефекта по этапам
Раздел 15. Результат: Повторная проверка охватывает всю цепочку до итогового документа.

19.16 Проверять юридический порог по фактам
Раздел 16. Результат: Квалификация содержит мотивировку и отдельно оставшиеся неизвестные факты.

19.17 Устанавливать связь инцидента с нужной моделью
Раздел 17. Результат: Уведомление не приписывает модели неподтверждённую причину.

19.18 Разделять первоначальную оценку и обновление уведомления
Раздел 18. Результат: Копии сообщений содержат дату, факты и обозначенные неопределённости.

19.19 Проверять персональные данные отдельно
Раздел 19. Результат: Сохранены отдельные решения по содержанию ошибки и защите данных.

19.20 Вести параллельную карту регуляторных режимов
Раздел 20. Результат: У каждого применимого маршрута есть ответственный и журнал исполнения.

19.21 Сообщать проверенные сведения с версиями
Раздел 21. Результат: Получатель может различить первоначальное предположение и уточнённый факт.

19.22 Проверять основание каждой передачи
Раздел 22. Результат: Передан минимальный согласованный пакет с журналом получателя.

19.23 Сохранять независимость экспертного исследования
Раздел 23. Результат: Заключение отвечает на вопросы по материалам без заданного виновника.

19.24 Разграничивать технический вывод и правовую ответственность
Раздел 24. Результат: В заключении не смешаны наблюдение, причинный вывод и решение о вине.

19.25 Обеспечивать исправление для затронутого лица
Раздел 25. Результат: Есть подтверждение исправления последствий либо мотивированное описание препятствия.

19.26 Проверять идентичность по различающим данным
Раздел 26. Результат: Решение мотивировано совокупностью данных, ошибочное объединение отменено.

19.27 Не подменять официальный розыск алгоритмической справкой
Раздел 27. Результат: В позиции указан действительный статус каждого источника.

19.28 Отделять подозрение от установленного нарушения
Раздел 28. Результат: Материал различает сигнал, проверенные факты и юридическую оценку.

19.29 Проверять вход скоринга и пересмотр
Раздел 29. Результат: Решение пересмотрено с исправленным входом; результат зафиксирован.

19.30 Приоритизировать предотвращение продолжающегося вреда
Раздел 30. Результат: Снижение текущего риска и расследование имеют отдельных ответственных.

19.31 Проверять происхождение спорного медиафайла
Раздел 31. Результат: Степень уверенности соответствует доступным первичным сведениям.

19.32 Раскрывать конфликт и назначать независимого рецензента
Раздел 32. Результат: Решение по компоненту не утверждено его прежним консультантом.

19.33 Защищать источники при передаче результата
Раздел 33. Результат: Открытый текст сохраняет вывод без ненужного раскрытия личности.

19.34 Строить отчёт вокруг проверенной причинной цепочки
Раздел 34. Результат: Рецензент воспроизводит цепочку аргумента по материалам.

19.35 Подтверждать корректирующую меру повторным тестом
Раздел 35. Результат: Возобновление допускается после документированного результата проверки.

19.36 Учитывать прямые и последующие последствия
Раздел 36. Результат: Проверены все известные получатели; спорные суммы не выданы за установленные.

19.37 Публиковать статус а не необоснованное обвинение
Раздел 37. Результат: Карточка позволяет понять доказательный статус без чтения закрытого досье.

19.38 Согласовывать факты международного уведомления
Раздел 38. Результат: Получатели имеют актуальную версию и сведения об исправлениях.

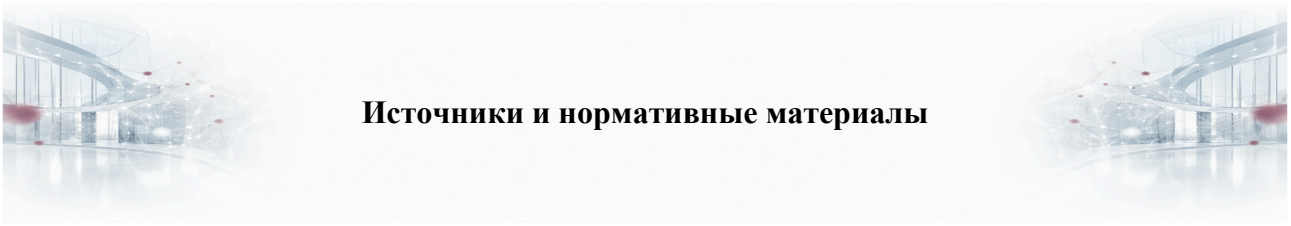
19.39 Проверять пять элементов качества расследования
Раздел 39. Результат: Каждый элемент подтверждён документом либо указанным ограничением.

19.40 Закреплять условия доступа и завершения договором
Раздел 40. Результат: Приёмка не требует выдуманной определённости при отсутствии материала.

19.41 Независимо проверять AI-помощь в расследовании
Раздел 41. Результат: Каждое существенное утверждение подтверждено доступным материалом.

19.42 Развивать программу по выявленным пробелам
Раздел 42. Результат: Этап завершён по проверяемому результату, а не календарной отметке.

19.43 Закрывать техническую процессную и восстановительную части
Раздел 43. Результат: Закрытие обосновано по каждой части без общего необоснованного статуса.



Источники и нормативные материалы

Нумерация в квадратных скобках в тексте отсылает к перечню уникальных первоисточников ниже; повторные ссылки на один и тот же акт объединены. Для обязательных норм следует использовать официальный текст с соответствующей датой действия. Разъяснения регуляторов, добровольные рекомендации OECD и NIST и стандарты ISO/IEC не заменяют нормативных требований. Ссылки ISO ведут на официальные аннотации; платные полные тексты стандартов не воспроизводятся и не утверждается, что проводилась их полная постатейная экспертиза.

[1] Европейский парламент и Совет ЕС. Регламент (EU) 2024/1689 (AI Act), консолидированный текст от 27.07.2026; в частности ст. 3(49), 26, 55, 73, 75, 113. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02024R1689-20260727

[2] Европейский парламент и Совет ЕС. Регламент (EU) 2026/1744 от 08.07.2026 (Digital Omnibus on AI), включая изменения ст. 113 AI Act. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32026R1744

[3] OECD (2024). Defining AI incidents and related terms. https://oecd.ai/en/ai-publications/defining-ai-incidents-and-related-terms

[4] OECD.AI. Overview and methodology of the AI Incidents and Hazards Monitor (AIM). https://oecd.ai/en/incidents-methodology

[5] NIST (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1. https://doi.org/10.6028/NIST.AI.100-1

[6] NIST (2024). Artificial Intelligence Risk Management Framework: Generative AI Profile, AI 600-1. https://doi.org/10.6028/NIST.AI.600-1

[7] NIST (2025). SP 800-61 Rev. 3: Incident Response Recommendations and Considerations for Cybersecurity Risk Management. https://csrc.nist.gov/pubs/sp/800/61/r3/final

[8] Европейский парламент и Совет ЕС. Регламент (EU) 2016/679 (GDPR), в частности ст. 4(12), 22, 33 и 34. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679

[9] EDPB (2023). Guidelines 9/2022 on personal data breach notification under GDPR, version 2.0. https://www.edpb.europa.eu/documents/guideline/guidelines-92022-on-personal-data-breach-notification-under-gdpr_en

[10] Европейский парламент и Совет ЕС. Директива (EU) 2022/2555 (NIS2), ст. 23. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022L2555

[11] Европейский парламент и Совет ЕС. Регламент (EU) 2022/2554 (DORA), ст. 19. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022R2554

[12] Европейская комиссия. Делегированный регламент (EU) 2025/301 от 23.10.2024, ст. 5: сроки уведомлений по DORA. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32025R0301

[13] ISO/IEC 27037:2012. Guidelines for identification, collection, acquisition and preservation of digital evidence (официальная аннотация). https://www.iso.org/standard/44381.html

[14] ISO/IEC 27043:2015. Incident investigation principles and processes (официальная аннотация). https://www.iso.org/standard/44407.html

[15] ISO/IEC 23894:2023. Artificial intelligence — Guidance on risk management (официальная аннотация). https://www.iso.org/standard/77304.html

[16] ISO/IEC 42001:2023. Artificial intelligence — Management system (официальная аннотация). https://www.iso.org/standard/42001

[17] Европейский союз. Хартия основных прав Европейского союза (2012/C 326/02). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT

[18] Совет Европы. Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (CETS 225): текст и обзор. https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence

[19] OHCHR (2011). Guiding Principles on Business and Human Rights: Implementing the UN Protect, Respect and Remedy Framework. Каталог OHCHR. https://searchlibrary.ohchr.org/record/11593

[20] Совет Европы. CETS 225: сведения о подписании, ратификациях и договорном статусе (проверяются на дату применения). https://www.coe.int/en/web/conventions/full-list?module=treaty-detail&treatynum=225

[21] OECD (2026). OECD Due Diligence Guidance for Responsible AI. https://doi.org/10.1787/41671712-en

[22] Европейская комиссия. European AI Office: полномочия и функции. https://digital-strategy.ec.europa.eu/en/policies/ai-office

[23] Европейская комиссия, AI Act Service Desk. FAQ: обязанности поставщика GPAI-модели с системным риском при серьёзном инциденте (позиция Комиссии, не самостоятельная правовая норма). https://ai-act-service-desk.ec.europa.eu/en/ai-act/faq/what-are-obligations-provider-if-serious-incident-involving-its-model-occurs

Ограничения применения

Законодательство, стандарты и судебная практика могут изменяться. Всякий вывод о наличии юридической обязанности, ответственности конкретного лица или допустимости доказательства требует проверки фактов и действующего права соответствующей юрисдикции. Модельные сценарии и формы приложений имеют учебно-аналитический характер.

Научные и методологические источники

[M1] NIST SP 800-86. Guide to Integrating Forensic Techniques into Incident Response. 2006. https://doi.org/10.6028/NIST.SP.800-86

[M2] Hernán M. A., Robins J. M. Causal Inference: What If. 2020. Авторская страница открытой версии: https://miguelhernan.org/whatifbook

[M3] Leveson N. System Safety. Lectures 6–7: CAST and Accident Analysis. MIT OpenCourseWare, 2016. https://ocw.mit.edu/courses/16-63j-system-safety-spring-2016/pages/lecture-notes/