Observatoire ARGA

МЕЖДУНАРОДНЫЙ АНАЛИТИЧЕСКИЙ ДОКЛАД

Исследование №21

Генеративный искусственный интеллект и web scraping

Законность сбора данных, прозрачность обучающих массивов и ответственность разработчиков

Организация: Observatoire ARGA

Русская версия

21 rue de l’Aviation, 64600 Anglet, France

info@argaobservatory.org | www.argaobservatory.org

22 октября 2026 года

Содержание

1. Предмет, цель исследования и проверяемые тезисы — 4
2. Понятийные границы: crawling, scraping, mining и обучение модели — 5
3. Карта жизненного цикла данных и разграничение операций — 5
4. Иерархия источников права и статус EDPB Guidelines 03/2026 — 7
5. Когда GDPR распространяется на открытые интернет-данные — 7
6. Контролёр, обработчик, совместный контролёр и продавец данных — 8
7. Доступность страницы и законность её дальнейшего использования — 10
8. Ограничение целей: исследовательская формула вместо лозунга «обучение ИИ» — 10
9. Законный интерес: критерий определённости и правомерности цели — 11
10. Законный интерес: критерий необходимости и альтернативы — 12
11. Законный интерес: баланс последствий и ожиданий субъекта — 13
12. Другие основания обработки, согласие и ложные универсальные решения — 14
13. Специальные категории данных: статья 9 GDPR — 14
14. Данные о судимостях и правонарушениях: отдельная статья 10 — 16
15. Данные несовершеннолетних и уязвимых лиц — 17
16. Требование прозрачности и статья 14 GDPR при непрямом получении данных — 17

17. Критерий несоразмерного усилия: исключение не должно поглотить правило — 19
18. Минимизация: до, во время и после извлечения — 19
19. Достоверность, время и проблема устаревшей информации — 20
20. Оценка воздействия на защиту данных и документация риска — 21
21. Право на доступ, исправление, стирание и ограничение обработки — 22
22. Возражения против обработки и учёт ограничений сайтов — 23
23. Модельные веса, memorisation и вопрос «забывания» — 24
24. Анонимизация, псевдонимизация и повторная идентификация — 25
25. Каталог происхождения: проверяемые свидетельства для каждого источника — 25
26. Международные передачи и доступ третьих стран — 26
27. Безопасность набора данных и обязанности подрядчиков — 27
28. Интеллектуальная собственность: параллельный, а не альтернативный правовой режим — 27
29. AI Act: документация GPAI, авторское право и публичная сводка — 28
30. Французский подход: CNIL и защитные меры при сборе в сети — 29
31. Великобритания: позиция ICO и сравнительные ограничения — 30
32. Непрозрачные поставщики наборов: due diligence до закупки — 30
33. Публичная отчётность: достаточно подробная, но не опасная — 31
34. Независимый аудит: предмет проверки и пределы заключения — 32
35. Статистическая выборка: почему несколько удачных тестов не доказывают безопасность — 33
36. Метрики эффективности фильтров и обработка ошибок классификации — 34
37. Санкционный комплаенс: риск вторичного распространения неверной идентичности — 34
38. Международный розыск, экстрадиция и данные уголовного производства — 35
39. Поставщики и разработчики: ответственность за решение при неполной информации — 36
40. Предложения Observatoire ARGA к практике EDPB и международному диалогу — 37
41. Выводы, ограничения и программа последующих исследований — 38
Приложение 1. Терминологическая карта для независимого заключения — 39
Приложение 2. Матрица юридических оснований и дополнительных ограничений — 39
Приложение 3. Форма LIA: конкретизация законного интереса — 40
Приложение 4. Форма LIA: необходимость и менее вмешивающиеся альтернативы — 41
Приложение 5. Баланс интересов: модель причиняемого и предотвращаемого вреда — 41
Приложение 6. Допуск и запрет интернет-источников — 42
Приложение 7. Source Intake Sheet: карточка первичного источника — 42
Приложение 8. Dataset Manifest: минимальная структура доказательства происхождения — 43
Приложение 9. Проверка поставщика уже собранного корпуса — 44
Приложение 10. RACI и разграничение полномочий участников — 45
Приложение 11. Публичное информационное уведомление по статье 14 GDPR: проект структуры — 45
Приложение 12. Обоснование статьи 14(5)(b): документированная процедура — 46
Приложение 13. DPIA: карта угроз правам и свободам — 47
Приложение 14. Стоп-правило для данных статьи 9 GDPR — 48
Приложение 15. Стоп-правило для сведений об уголовных правонарушениях — 48
Приложение 16. Защита детей: правила доменных и содержательных исключений — 49
Приложение 17. Процедура обработки запроса субъекта данных — 49
Приложение 18. Карта исполнения удаления на разных уровнях модели — 50
Приложение 19. Протокол аудита фактической работы crawler — 51
Приложение 20. Модель защиты данных в цепочке подрядчиков — 51
Приложение 21. Авторское право и право на text and data mining: отдельная проверка — 52
Приложение 22. Связка AI Act, GDPR и раскрытия содержания обучения — 52
Приложение 23. Карта трансграничного доступа к персональным данным — 53
Приложение 24. План статистической проверки источниковых запретов — 53
Приложение 25. Тестовая матрица обнаружения персональных данных — 54
Приложение 26. Программа независимого аудита на 20 рабочих процедур — 55
Приложение 27. Evidence Chain: образец движения одного документа — 56
Приложение 28. Учебный кейс: ошибочная санкционная атрибуция — 57
Приложение 29. Учебный кейс: устаревшая информация об уголовном преследовании — 57
Приложение 30. Учебный кейс: брокер продаёт «обезличенный» набор — 58
Приложение 31. Публичный отчёт о прозрачности: минимальные показатели — 58
Приложение 32. Проект предложений к EDPB Guidelines 03/2026 — 59
Приложение 33. План внедрения организационной методики на 90 дней — 60
Приложение 34. Критерии принятия отчёта независимой проверкой — 60
Приложение 35. Реестр правовых и технических вопросов для последующей рецензии — 61
Приложение 36. Проектирование crawler как юридически контролируемой системы — 61

Приложение 37. Судебная и доказательная оценка происхождения спорного фрагмента — 63
Приложение 38. Деперсонализация, linking и риск обратной атрибуции — 64
Приложение 39. Управление удалением данных из резервных копий и обучающих эпох — 65
Приложение 40. Процедура оспаривания решения независимого аудитора — 66
Приложение 41. Проектирование понятного уведомления о scraping: тест с реальным пользователем — 67
Приложение 42. Различие частного scraping и правоприменительного доступа к данным — 68
Приложение 43. Предобучение, дообучение, retrieval и обработка пользовательских запросов — 69
Приложение 44. Метод ответа на распространённые недоказанные правовые утверждения — 70
Приложение 45. Кейс переговоров с брокером: подробная процедура доказательной закупки — 71
Приложение 46. Обоснование несоразмерного усилия при индивидуальном уведомлении — 72
Приложение 47. Национальные правовые ограничения публичных реестров: карта коллизий — 73
Приложение 48. Статистическое сравнение качества двух систем фильтрации — 74
Приложение 49. Инцидент при раскрытии обучающего корпуса: соединение двух процедур — 75
Приложение 50. Учебное дело: запрос об удалении при совпадении имён — 76

АННОТАЦИЯ

Исследование №21 посвящено правовым ограничениям автоматизированного извлечения общедоступных интернет-данных для разработки генеративных моделей. Анализ разграничивает доступность страницы, законное основание обработки персональных данных, правомерное использование контента и доказательную ответственность разработчиков. Рассмотрены положения GDPR, статья 53 EU AI Act, рекомендации CNIL и ICO, позиция EDPB и практические методы независимой проверки.

Центральным источником является Guidelines 03/2026 EDPB on web scraping in the context of generative AI. Версия 1.0 принята для публичного обсуждения 7 июля 2026 года, опубликована 8 июля, консультация продолжается до 30 октября 2026 года. На дату доклада положения проекта не являются окончательной редакцией рекомендаций после консультации. Авторские предложения Observatoire ARGA отделяются от обязательных требований права ЕС [2][13].

Практический блок включает 50 приложений: формы анализа законного интереса, DPIA, контроля источников, проверки data brokers, учёта ограничений сайтов и прав субъектов, provenance-манифесты, процедуры независимого аудита и условные кейсы. Доклад не содержит результатов недоступных внутренним разработчикам аудитов и не приписывает какой-либо организации нарушения без исследованных доказательств.

Ключевые выводы

Положение	Методический вывод
Правовая предпосылка	Общедоступность материала не заменяет основания статьи 6 GDPR, правил статей 9/10 и прав субъектов.
Независимая проверка	Оцениваются источники, цели, роли, реальные логи и эффективность фильтров, а не только наличие деклараций.
Особый риск	Обучение на данных о судимостях, здоровье, детях и ошибочных санкционных совпадениях требует отдельных ограничений.
Ремедиация	Удаление из набора и устранение влияния уже обученной модели являются разными техническими задачами.
Статус EDPB	Guidelines 03/2026 на дату исследования — проект для публичных комментариев, а не самостоятельный новый закон.

1. Предмет, цель исследования и проверяемые тезисы

Исследование посвящено правовому статусу автоматизированного сбора информации из открытых интернет-источников для обучения, дообучения и оценки генеративных моделей. Центральный вопрос состоит не в том, может ли программа технически загрузить страницу, а в том, при каких условиях организация вправе извлекать, объединять, хранить и многократно использовать сведения о конкретных лицах. Автоматическая доступность материала не превращает его ни в «ничейный» ресурс, ни в данные, свободные от правовых ограничений. Разделение технической доступности, правомерности доступа, законности обработки персональных данных и допустимости использования защищённых произведений образует отправную точку исследования [1][2][11].

Доклад подготовлен как нормативно-методологическое исследование по состоянию на 22 октября 2026 года. В нём не заявляется, что авторы получили доступ к внутренним обучающим наборам конкретных компаний, проводили судебную экспертизу чужих моделей или установили факт нарушения законодательства каким-либо названным разработчиком. Условные кейсы прямо обозначены как моделируемые ситуации; приведённые формы и пороги являются предложениями Observatoire ARGA, а не официальными решениями надзорного органа.

Проверяемая исследовательская гипотеза состоит в следующем: прозрачность происхождения наборов данных, документирование трёхчастного теста законного интереса и техническое доказательство минимизации способны снизить риск нарушения прав, но отдельно взятое наличие документа, robots.txt-политики либо исключения домена не является достаточным доказательством правомерности всех этапов обработки. Комплаенс должен оценивать юридическое основание, фактическую процедуру и реальный эффект для субъекта данных в совокупности [1][2][3].

Для практического анализа используется цепочка: формулирование цели → выбор источника → допустимость автоматического извлечения → фильтрация → инвентаризация → обучение модели → проверка воспроизведения персональных данных → обработка запросов и возражений → удаление или ограничение дальнейшего использования. На каждом звене фиксируются владелец решения, доказательный артефакт, применимое правило и возможность независимой проверки. Неспособность подтвердить один ключевой переход нельзя компенсировать большим объёмом общей декларативной документации.

Вопрос исследования	Проверяемый результат
Имеется ли правовое основание для конкретной обработки?	Запись цели, применимой статьи GDPR и материалов её оценки
Могут ли быть исключены ненужные данные до сбора?	Лист правил выбора доменов, запретов и фильтров
Может ли субъект осуществить права на практике?	Доступный канал запроса и журнал исполнения
Происхождение данных воспроизводимо?	Связанные версии реестра источников, коллекций и моделей
Подтверждает ли независимый эксперт результат?	Ограниченный по объёму и доказательствам отчёт проверки

2. Понятийные границы: crawling, scraping, mining и обучение модели

Web crawling означает автоматизированное обнаружение сетевых ресурсов и переход по найденным ссылкам. Web scraping представляет собой извлечение и сохранение содержимого или его структурированных элементов. Целевая выборка по разрешённым категориям существенно отличается по масштабу риска от неограниченного обхода сети. В Guidelines 03/2026 EDPB особо разграничивает targeted и untargeted scraping и рассматривает сбор из статических и динамических страниц [2].

Text and data mining означает аналитическую обработку цифрового материала для выявления информации, включая модели, закономерности и корреляции. Оно не является синонимом законности сбора, и специальное исключение из авторского права не отменяет требований защиты персональных данных. Training dataset отличается от «сырого» захвата: он может содержать очищенные документы, фрагменты, метаданные, embeddings, аннотации и производные признаки. Каждая операция может порождать самостоятельные юридические вопросы [1][11].

Термины «публичная страница», «открытый источник» и «доступ без учётной записи» описывают условия получения, но ничего сами по себе не говорят о допустимом последующем использовании. Даже индексируемая поиском публикация способна содержать сведения об адресе частного лица, болезни, уголовном преследовании, несовершеннолетнем или частной переписке. При загрузке и структурировании такой информации происходит обработка персональных данных, если лицо идентифицировано или идентифицируемо [1].

Следует различать модель, которую можно исследовать через запросы, и обучающий массив, к которому есть полномасштабный доступ. Внешний тест на memorisation показывает определённый тип поведения в оговорённых условиях и не устанавливает состав всех источников обучения. Аналогично отсутствие найденного совпадения не подтверждает, что спорного документа никогда не было в тренировочном корпусе. Методология должна открыто указывать границы знания и не выдавать статистическую проверку за полную реконструкцию данных.

3. Карта жизненного цикла данных и разграничение операций

Правовая оценка начинается до запуска crawler. Сначала определяется интерес и конкретная цель: например, многоязычное обучение текстовой модели для поискового интерфейса. Затем описываются данные, без которых эта цель не достигается, и сравниваются менее вмешивающиеся альтернативы. До технического сбора должны существовать записи о выбранных доменах, категориях содержимого, правовых рисках, сроках хранения и предполагаемых последующих преобразованиях [1][2].

После сбора данные обычно проходят очистку, удаление дубликатов, языкозависимую сегментацию, оценку качества, аннотирование и перемещение между поставщиками вычислительных услуг. Ошибка состоит в том, чтобы считать эти этапы одной безразличной технической операцией: типичный риск возникает при повторном объединении набора, первоначально ограниченного одной целью, с другой базой, позволяющей идентифицировать лиц, либо при сохранении невостребованных копий в резервных хранилищах. Юридическая оценка должна учитывать весь жизненный цикл [1][2].

Для каждого датасета ведётся versioned manifest с уникальным идентификатором, датой, категорией источников, основанием обработки, ответственным подразделением, признаками возможных специальных категорий, правовым статусом, режимом доступа и ссылкой на связанные версии модели. Сам по себе manifest не доказывает законность, однако позволяет восстановить историю решений, увидеть пробелы и исполнить запросы об удалении не вслепую.

На стадии обучения требуется различать очистку обучающих примеров от изменения весов модели. Удаление строки из каталога не равнозначно устранению влияния её на уже обученную модель. Подход к исправлению определяют по степени риска, технической осуществимости, доказательствам запоминания, наличию последующего обучения и возможности вывода или обновления модели. Внутреннее решение фиксируется с указанием остающегося риска и основания [3][10].

Этап	Риск	Контрольный артефакт
Выбор источников	Неправомерный сбор, чужие запреты	Реестр категорий и допуск к сбору
Извлечение	Случайный захват частных данных	Логи crawler, хэши, исключения
Очистка	Сохранение остаточных идентификаторов	Протокол фильтрации и тестирование
Объединение	Непредвиденная повторная идентификация	Карта связей наборов и DPIA
Обучение	Memorisation, утечка сведений	Model card, тест извлечения
Выпуск	Трудность реализации возражения	Канал жалоб, план исправлений

Практическая реконструкция движения данных должна начинаться с контрольной даты. Исследователь фиксирует, какой именно crawler действовал, какие правила источников были установлены, какой код использовался для очистки и как объект оказался в обучающем наборе. Если журналы отсутствуют, нельзя восстанавливать историю задним числом на основании нынешней конфигурации системы. Нынешние запреты на сбор не доказывают, что такая же политика действовала в момент получения спорного материала. По той же причине отчёт о проверке фиксирует идентификаторы версий как доказательства, а не как удобные внутренние ярлыки. Минимальный достаточный набор включает сведения о дате извлечения, происхождении, алгоритме фильтрации, переходах между стадиями и моменте передачи подрядчику. История операций особенно важна для исковой или административной оценки, когда оператор заявляет об исправлении данных после жалобы, но не показывает, была ли выпущена модель, обученная на прежнем корпусе.

Отдельно рассматривается проблема временного несоответствия. Запись могла быть опубликована законно в один период, затем исправлена или отозвана, но продолжать циркулировать в архиве, поисковом кэше и производном датасете. Контроль допустимости нельзя сводить к одноразовой отметке «источник публичный». Для чувствительных категорий полезны повторные проверки статуса, особенно перед очередной эпохой обучения или передачей коллекции другому оператору. Предлагаемый контроль состоит из пяти вопросов: существует ли источник сейчас, соответствует ли текущему первичному документу, не относится ли к исключённой категории, сохранены ли ограничения правового режима и существует ли воспроизводимый путь к обновлению или удалению. Эта модель не

возлагает на разработчика неограниченную обязанность ежедневно проверять весь интернет, но создаёт критерии соразмерного контроля для заведомо значимых наборов.

4. Иерархия источников права и статус EDPB Guidelines 03/2026

Правовую основу в ЕС составляет GDPR (Regulation 2016/679), применяемый непосредственно при наличии условий его территориальной и материальной сферы. В части использования генеративных моделей дополняют картину положения AI Act, включая обязанности поставщиков GPAI-моделей по документации, политике соблюдения авторского права и публичному описанию учебного контента; соответствующие обязательства нельзя представлять как полное разрешение на любую обработку данных [1][11][12].

7 июля 2026 года EDPB принял Guidelines 03/2026 version 1.0 для публичной консультации; текст опубликован 8 июля, период обратной связи установлен до 30 октября 2026 года. На дату данного доклада документ остаётся проектом рекомендаций, а не окончательной утверждённой редакцией после консультации. Это влияет на юридическую формулировку: выводы проекта существенны для понимания позиции регуляторов, однако необходимо отличать саму норму GDPR от текущего толкования EDPB и авторских рекомендаций Observatoire ARGA [2][13].

Предмет проекта сознательно ограничен скрапингом из внешних интернет-источников, выполняемым частными организациями для разработки генеративного ИИ. EDPB рассматривает случаи самостоятельного сбора и закупки ранее собранных массивов для обучения, но специально не охватывает в полном объёме деятельность брокеров, просто предлагающих наборы другим, а также сбор организацией своих собственных данных и публично-правовые органы. Такое ограничение предмета документа не освобождает исключённые ситуации от GDPR [2].

Opinion 28/2024 EDPB содержит относимую общую позицию о персональных данных в контексте AI-моделей, в том числе об условиях возможной анонимности, легитимном интересе и влиянии неправомерно обработанных данных на последующее применение модели. Поэтому исследование использует Guidelines 03/2026 как специализированное развитие подхода, а Opinion 28/2024 как дополнительный материал для анализа последующих стадий. Вопрос фактического статуса каждой редакции проверяется отдельно перед публичным использованием [3].

5. Когда GDPR распространяется на открытые интернет-данные

GDPR применяется к обработке персональных данных, а не к абстрактному интернет-документу как таковому. Персональность оценивается по возможности прямо или косвенно установить лицо с учётом средств, которые разумно могут использоваться, контекста и способности связать атрибуты. Имя не является необходимым условием: устойчивый псевдоним, фотография, URL профиля, редкий набор биографических обстоятельств или географический признак могут приводить к идентификации [1].

Смешанные коллекции из технического текста, изображений и личных сообщений не становятся полностью «неперсональными» из-за доминирования общедоступных статей. Если в массиве есть идентифицируемые лица, правила GDPR распространяются на соответствующую обработку. Для независимой проверки полезно выделять единицу оценки

на уровне источника, документа и фрагмента, но нельзя забывать о риске идентификации при последующем соединении компонентов [1][2].

Исключение бытовой или личной деятельности не следует автоматически распространять на промышленный crawler. Юридическое лицо, создающее коммерческую или исследовательскую генеративную модель, должно сначала установить собственную роль в обработке и территориальную применимость режима, затем определить законное основание и меры защиты. Публикация сведений самим субъектом также не равнозначна согласию на любые цели повторного использования.

В споре доказательное значение имеют не только видимый пользователю веб-интерфейс, но и технические обстоятельства: архивная копия, ограничения регистрации, сведения о том, кем опубликован контент, дата запроса, контроль доступа, условия сайта и возможное удаление материала из первоисточника. Отсутствие таких записей затрудняет восстановление контекста законности и разумных ожиданий человека.

В пределах GDPR важно различать сведения, представляющие собой персональные данные, и изображения или тексты, где идентификация лица фактически невозможна с применением разумно вероятных средств. Юридическая квалификация проводится не по формату файла. Документ без явного имени может содержать редкое сочетание должности, населённого пункта, даты и обстоятельств, достаточное для опознания человека. Напротив, общие статистические показатели, не позволяющие выделить и идентифицировать конкретного человека, могут находиться вне режима персональных данных в зависимости от обстоятельств. Следовательно, для выборочной проверки следует оценивать возможность объединения сведений, а не только наличие регулярных выражений для распознавания имен и телефонов. Риск обнаружения лица усиливается после векторизации и связывания нескольких индексов, хотя сам по себе embedding не во всех случаях обязательно является персональными данными.

Территориальная область GDPR требует самостоятельного анализа по статье 3. Само наличие у веб-страницы европейского домена, французского языка либо европейского гражданина не представляет универсального критерия применимости. Важны деятельность учреждения контролёра или процессора, а для лиц за пределами ЕС — условия предложения товаров или услуг субъектам в Союзе либо наблюдения за их поведением в установленных правом пределах. Отдельно нужно учитывать, кто фактически определяет цели и средства обработки, как устроены группы компаний и где осуществляется доступ к данным. В отчёте полезно фиксировать основание территориальной привязки и степень уверенности. Это предотвращает и неправомерное распространение GDPR на любые мировые интернет-данные, и столь же ошибочное предположение о том, что офшорное расположение подрядчика автоматически исключает европейские обязательства.

6. Контролёр, обработчик, совместный контролёр и продавец данных

Статус участника определяется фактическим решением о целях и существенных средствах обработки, а не только договорной надписью. Разработчик, который определяет тематические категории и требования к данным и поручает подрядчику технически их собрать по документированным инструкциям, нередко выступает контролёром, а подрядчик может быть обработчиком; при самостоятельном определении цели и набора источников роли могут отличаться [2][4].

Если брокер сначала собирает массив для своих целей, а затем продаёт его разработчику, исходный сбор и последующее обучение необходимо оценивать как возможно раздельные операции двух самостоятельных контролёров. Договор купли-продажи «лицензированного датасета» не заменяет анализ того, при каких условиях персональные данные были получены. Если стороны совместно определяли цели и существенные средства, может возникнуть режим совместных контролёров по статье 26 GDPR [1][2][4].

Матрица ролей должна учитывать облачную платформу, исполнителя разметки, студию очистки, производителя embeddings и downstream-интегратора, если каждый получает доступ к персональным данным. Для обработки по поручению существенны условия статьи 28 GDPR: предмет и длительность, инструкции, безопасность, помощь контролёру, удаление или возвращение данных, аудит и субобработчики. Аффилированность компаний не устраняет необходимость такого анализа [1][4].

Независимый аудитор проверяет в первую очередь фактическую схему движения данных. Признак риска — когда компания утверждает, что «поставщик уже отвечает за GDPR», однако не может назвать контролёра исходного извлечения, указать цели передачи или показать проверяемый объём данных. Такое расхождение фиксируется как недостаток доказательств, а не автоматически как установленное нарушение: правовой вывод требует исследования фактов.

Типовая цепочка из четырёх участников показывает, почему название договора не определяет правовой статус: платформа публикует материал, агрегатор технически извлекает сведения, брокер нормализует и продаёт массив, разработчик выбирает обучение и критерии качества. Для каждой операции следует установить, кто решил, какие категории собирать, на каких основаниях, как долго хранить, с какими системами объединять и кому раскрывать. Агрегатор может быть обработчиком по поручению, самостоятельным контролёром в другой части своей работы либо участником совместного определения целей. Необходимо исследовать реальные полномочия и способы принятия решений, а не формально присвоить всем подрядчикам статус processor. При совмещении ролей матрица должна описывать отдельные процессы и документы, включая обязанности статьи 28 GDPR и требования статьи 26 к совместным контролёрам там, где они действительно применимы.

При закупке готового корпуса покупатель не получает автоматически законное основание из договора поставщика. Следует проверить, была ли первоначальная обработка законной, соответствует ли новая цель критерию совместимости либо требует отдельного основания, соблюдены ли условия передачи и возможно ли исполнение прав субъектов. Гарантия брокера о «полной законности» полезна для распределения гражданско-правовых рисков между сторонами, но сама по себе не заменяет публично-правовые обязанности. Необходимо также отделять персональные данные от прав на охраняемый контент: наличие лицензии на тексты не решает вопрос статей 6, 9 и 10 GDPR, а наличие GDPR-основания не устраняет проверку авторских прав. При отказе поставщика представить даже обобщённые сведения о происхождении и запретных категориях следует документировать непреодолимую неопределённость и приостанавливать допуск высокорисковых коллекций, а не просто повышать оценку договорного риска.

7. Доступность страницы и законность её дальнейшего использования

Открытый доступ к странице влияет на обоснованные ожидания лиц и на техническую возможность получения, но не создаёт самостоятельного законного основания по статье 6 GDPR. Если данные могут быть обработаны по legitimate interests, остаётся необходимым выполнить весь тест соразмерности, оценить масштабы, виды информации, наличие возражений и меры минимизации. Для отдельных типов информации действуют дополнительные ограничения [1][2][5].

С точки зрения возможного ущерба принципиально отличаются публикации учреждения о своём расписании, общий профессиональный профиль сотрудника, персональная история болезни и фотография ребёнка. Все четыре примера могут быть доступны извне, но разумные ожидания человека, цель первичной публикации, чувствительность, допустимый объём и риск дальнейшего распространения не совпадают. Грубое правило «всё публичное разрешено собирать» юридически несостоятельно.

При обработке контента требуется параллельно оценивать договорные ограничения ресурса, применимое право в отношении технических мер доступа, охрану баз данных и интеллектуальных прав. robots.txt, CAPTCHA и другие выраженные возражения веб-сайта рассматриваются как важные фактические и организационные сигналы в анализе законности и добросовестности. Они не являются универсальной заменой оценки GDPR, но игнорировать их без обоснования нельзя [2][6].

Матрица принятия решения должна предусматривать отдельное «блокирующее» состояние при неизвестном правовом статусе чувствительного набора, а не только три градации уровня риска. Временная приостановка сбора до проверки оказывается в таких случаях обоснованным управленческим результатом, а не признанием незаконности всей модели.

8. Ограничение целей: исследовательская формула вместо лозунга «обучение ИИ»

Статья 5(1)(b) GDPR требует конкретных, явных и законных целей. Формулировка «совершенствование искусственного интеллекта» слишком широка для практической оценки необходимости и пределов обработки. Разработчик должен уточнить задачу модели, языковой, отраслевой и функциональный охват, предполагаемых пользователей, критерии успешности и то, почему необходимы именно рассматриваемые классы персональных данных [1][3].

Обучение общего языкового компонента, настройка медицинского ассистента, обучение системы идентификации людей и разработка механизма поиска судебных документов имеют разные цели. Один и тот же первоначальный веб-корпус не обязательно может быть повторно использован в каждом из перечисленных случаев. При переходе к новой цели оценивается её совместимость, связь с первоначальной публикацией, характер данных и последствия; «разработчик так решил позднее» не является самостоятельным основанием.

Метацель «проверять безопасность» не делает безопасным сохранение всех идентификаторов. Наборы оценки качества и инцидентного реагирования также должны иметь ясный объём, срок, доступ, тип данных и процесс удаления. Чем неопределённее жизненный цикл, тем труднее убедительно показать минимизацию и соответствие первоначальным законным интересам.

Рекомендуется утверждать цель на двух уровнях: юридическая карточка обработки (зачем и на каком основании) и техническая спецификация (какие документы, признаки, форматы, инструменты и метрики). Изменения обеих версий согласуются до повторного сбора, а не задним числом. Аудитор проверяет версии и фактические логи, не удовлетворяясь новейшей красивой редакцией политики.

9. Законный интерес: критерий определённости и правомерности цели

При обращении к статье 6(1)(f) GDPR необходимо доказать три кумулятивных условия: законный интерес контролёра или третьего лица, необходимость обработки для его достижения и отсутствие преобладания интересов либо основных прав и свобод человека. Нельзя объявить интерес легитимным лишь потому, что деятельность инновационна, коммерчески значима или общественно полезна. EDPB подчёркивает необходимость конкретной цели и доказательств, а не абстрактного обещания будущей пользы [1][2][3][5].

Первый этап LIA фиксирует, кто именно получает выгоду и какова её связь с законной деятельностью, какие ожидаются функциональные результаты и какие правовые ограничения действуют параллельно. Коммерческий интерес может быть законным по характеру, но сам по себе не достаточен для итогового вывода о допустимости обработки. Документ должен позволять независимому рецензенту критиковать заявленную цель по наблюдаемым фактам.

Если заявлен интерес обеспечения качества перевода юридических материалов, следует пояснить, почему для обучения нужны персонализированные судебные акты из сети, а не обезличенный профессиональный корпус. Если выгоду описывают как борьбу с мошенничеством, требуется отдельно показать связь с конкретными задачами обнаружения мошенничества, потенциальное использование чувствительной информации и ограничение последующих целей. Эти вопросы нельзя заменять ссылкой на интерес абстрактной безопасности.

В отсутствие доказательств законного интереса следует искать иную допустимую конструкцию либо отказаться от обработки. В докладе сознательно не утверждается, что legitimate interest всегда единственное возможное основание любой обработки, связанной с ИИ: основание определяется операцией, её конкретными фактами и пределами соответствующего права. Для массового частного web scraping ICO рассматривает legitimate interests как практически основной кандидат в описанных им типичных условиях [7].

Для формулирования легитимного интереса необходима проверяемая связь между бизнес-целью и конкретным использованием данных. Формула «развитие современных технологий» слишком широка, чтобы объяснить сбор всех страниц и их неограниченное сохранение. Более узкая цель могла бы состоять в повышении качества терминологического перевода конкретных языковых пар для доступного пользователю сервиса. Но даже такая цель сама по себе не доказывает необходимость собирать идентифицирующие сведения и не предрешает баланс. В LIA следует описывать, кто получит предполагаемую пользу, почему цель правомерна, в каком временном горизонте она достижима, чем подтверждается ожидаемая полезность и каким способом будут пересматриваться исходные предположения. Если изменяется задача модели, территория её применения или предполагаемая аудитория, старую формулу интереса требуется пересмотреть.

Контролирующий вывод должен быть фальсифицируемым: можно ли показать, что организация отказалась от определённого класса данных именно потому, что он не нужен для обоснованной цели? Если все предложенные классы автоматически признаны необходимыми, документ похож на оправдание принятого решения, а не на независимую оценку. Полезны протоколы отклонённых источников и письменное обоснование альтернатив, в том числе использование лицензированных корпусов, синтетических материалов, агрегированных данных и более узкого целевого scraping. Внутренний комитет может требовать предъявления сравнимых результатов качества модели без наиболее рискованного сегмента, прежде чем разрешать дополнительный сбор. Это методическое предложение, а не особая предписанная законом процедура; юридический результат всегда зависит от фактического теста статьи 6(1)(f) и принципов статьи 5 GDPR.

10. Законный интерес: критерий необходимости и альтернативы

Необходимость не тождественна удобству, дешевизне или преимуществу большего числа данных. На втором этапе LIA сравниваются разумные средства достижения заявленной цели, которые меньше вмешиваются в частную жизнь. Это могут быть лицензированные корпусы с проверенным происхождением, материалы без персональных признаков, выборочные подмножества, качественные синтетические данные, локальная очистка или ограниченный сбор по заранее определённым тематикам [2][6][7].

Технически возможно уменьшить объём извлекаемых страниц, число идентификаторов, историческую глубину, частоту обновления и время хранения. Необходимо указать, какую именно метрику качества предположительно ухудшит ограничение и насколько это подтверждено испытаниями, а не интуицией. Сравнительный подход к стадиям разработки и использования моделей также обсуждается ICO [9]. Если материал собирается «на всякий случай» для неопределённых будущих задач, тест необходимости особенно трудно пройти.

EDPB в проекте Guidelines 03/2026 называет сужение критериев сбора, псевдонимизацию, а в некоторых случаях синтетические данные как средства снижения вмешательства. При этом псевдонимизация не автоматически выводит набор из сферы GDPR. Техническая альтернатива должна проверяться на утечку, реалистичность и сохранение скрытой идентифицируемости; данные, лишь переименованные хэшами, могут оставаться персональными [1][2].

Предлагаемый доказательный стандарт: по каждой отвергнутой альтернативе указывать её описание, ожидаемую функциональную потерю, эмпирическое подтверждение, дополнительную нагрузку и результат сравнительной оценки воздействия на права. Пустая таблица «альтернатив нет» не считается мотивированным ответом. Измеримость такого выбора особенно важна для независимого аудита больших коллекций.

Необходимость оценивается в связи с выбранной законной целью, а не путём сравнения обрабатываемых данных с максимальными техническими возможностями инфраструктуры. Ограничение касается не только числа записей. Значение имеют глубина исторического архива, повторное извлечение, идентифицирующие метаданные, сохранение полнотекстовых документов при достаточности фрагментов и использование биографических признаков в задачах, где они не нужны. Например, для обучения распознаванию юридической терминологии допустимо проверить, сохраняется ли исследовательская ценность при исключении фамилий сторон, сведений о здоровье, адресов, фотографий и прямых цитат из

частных публикаций. Если цель достигается сопоставимым способом с меньшим вмешательством, отказ от такого способа нуждается в убедительном обосновании.

Метод сравнительных альтернатив должен предусматривать не только окончательные тестовые баллы, но и расходы, качество данных, практическую применимость и влияние на затронутых лиц. Нельзя требовать абстрактно нулевого риска, однако и коммерческое удобство не равно юридической необходимости. Результаты эксперимента отражаются как ограниченное техническое свидетельство: описание корпуса A и B, параметров модели, фиксированных тестовых задач, разброса результатов и критериев принятия решения. Испытание на малой случайной выборке не даёт оснований объявить альтернативный вариант непригодным для всех языков и всех задач. Внутренние записи позволят позднее проверить, не использовалась ли декларация о «безальтернативности» лишь для оправдания более дешёвого неразборчивого захвата.

11. Законный интерес: баланс последствий и ожиданий субъекта

Третий этап предполагает реальное сопоставление законного интереса и интересов, прав и свобод человека. В анализ включаются источник, контекст первоначальной публикации, предсказуемость перепрофилирования, широта последующего использования, масштабы обработки, уровень наблюдаемости и возможность реализации возражений. Наличие открытого профиля не означает ожидания, что содержание попадёт в модель, способную генерировать похожий текст или строить индивидуальные профили [1][2][3].

Профиль риска возрастает при сборе данных детей, сведений о частной жизни, политических или религиозных интересах, финансовом состоянии, местонахождении, обвинениях и материалах старше нескольких лет, содержание которых могло утратить актуальность. Существенны и «охлаждающий эффект» для свободы выражения: осознание тотального машинного архивирования может менять поведение людей даже без раскрытия отдельного имени [2].

Смягчающие меры уменьшают вероятность или тяжесть вреда, но не служат универсальным индульгенционным документом. Ограничение доменов, быстрое удаление, доступные сообщения о сборе, механизм возражений, отдельная фильтрация сложных категорий и тестирование memorisation учитываются только в той мере, в какой они действительно действуют и подтверждены. Формальное существование политики не равно её выполнению.

Модель отчёта об интересах предусматривает две независимые подписи: ответственное бизнес-подразделение подтверждает необходимость, а специалист по защите данных анализирует права и остающийся риск. При неразрешённом существенном конфликте рекомендуется эскалация и приостановка сбора, если невозможно обеспечить допустимость обработки. Это предложение методологии ARGA, а не самостоятельный обязательный «двухподписной» механизм GDPR.

Баланс интересов должен учитывать контекст первоначального раскрытия материала. Публикация сообщения в профессиональном каталоге для связи с клиентами, обмен опытом в группе взаимопомощи, размещение судебного акта в публичной базе и обнародование фотографии ребёнка родителем отличаются по цели и ожиданиям автора. Пользователь может разумно понимать, что запись прочитают другие посетители, но не ожидать бессрочного извлечения всех публикаций для обучения коммерческой модели и

последующего воспроизведения деталей через годы. Следовательно, формальная открытость ресурса не заменяет оценку масштаба сбора, аудитории, дополнительных целей и потенциальных неблагоприятных последствий. Особый вес имеют сведения о здоровье, политических взглядах, уязвимых лицах и неправомерно распространявшихся обвинениях.

При оценке мер снижения риска следует проверять не только их наличие, но и остаточный эффект. Условный сценарий: фильтр убирает прямые номера телефонов из 99,5 процента размеченных тестовых сообщений, однако пропускает редкие сочетания фамилии и места работы, которые позволяют определить человека. Первый результат не доказывает общей безопасности. Для разных типов данных нужны разные критерии и проверка ошибок второго рода. Положительное решение по LIA должно содержать условия: точные источники, период обновления тестов, срок пересмотра, процедуру жалобы, ответственное лицо и контроль выполнения. При изменении модели с внутреннего экспериментального инструмента на широко доступный продукт баланс может существенно измениться из-за нового масштаба распространения.

12. Другие основания обработки, согласие и ложные универсальные решения

GDPR предусматривает шесть оснований статьи 6(1). Их применимость нельзя выбирать по удобству отчётности: согласие требует соблюдения условий статьи 7 и возможности его отзыва, договорное основание связано с необходимостью исполнения договора с субъектом, юридическая обязанность и публичная задача имеют самостоятельные правовые предпосылки. Для частной массовой загрузки чужих интернет-страниц эти основания часто не соответствуют реальным обстоятельствам, однако категорический запрет на их использование во всех моделях был бы неточным [1][7].

Принципиально отличается сбор специально предоставленных участниками материалов при информированном согласии от извлечения публикаций третьих лиц, которые вообще не взаимодействовали с разработчиком. Нельзя толковать согласие автора на публикацию сайта как согласие всех лиц, упомянутых в этой публикации, на последующее массовое обучение генеративной модели. Правовое основание оценивается применительно к каждой операции и к соответствующим категориям персональных данных.

Отзыв согласия, возражение против обработки на основании законного интереса и запрет на определённые последующие цели имеют различную юридическую природу. Поэтому единая кнопка «отписаться от ИИ» полезна как интерфейс, но внутренний обработчик обязан классифицировать требования по соответствующим статьям GDPR. Он не вправе без анализа обещать безусловное удаление всего влияния информации из модели, если реальная техническая способность и правовая обязанность отличаются [1][10].

В кейсе совмещения различных правовых оснований необходимо документировать разграничение наборов и стадий. Недопустимо «переключать» основание задним числом как способ оправдать уже завершённый сбор после предъявленного возражения. Изменение правовой позиции фиксируется открыто, с объяснением фактов и последствий для людей.

13. Специальные категории данных: статья 9 GDPR

Сведения, раскрывающие расовое или этническое происхождение, политические взгляды, религиозные либо философские убеждения, членство в профсоюзе, а также данные о

здоровье, сексуальной жизни или ориентации, генетические и биометрические данные для уникальной идентификации подпадают под особую защиту статьи 9 GDPR. Обработка в общем случае запрещена, если не установлено конкретное исключение статьи 9(2), при наличии самостоятельного основания по статье 6. Ссылка на «общедоступный интернет» не восполняет отсутствующее исключение [1][2].

Для web scraping характерна проблема инцидентного захвата: разработчик ищет текст определённого жанра, но вместе с ним получает сведения о здоровье, судебном конфликте или убеждениях автора. Проект EDPB 03/2026 рассматривает возможность ограниченного применения логики судебного решения GC and Others C-136/17 к такому случайному остаточному сбору при реально действующих мерах предотвращения, одновременно подчёркивая отсутствие общего освобождения от статьи 9. Возможность опереться на подобную аргументацию определяется по конкретному случаю, а не автоматически [2][17].

Практическая процедура должна различать предварительную фильтрацию источников, фильтрацию уже полученного материала, временное карантинное хранилище, удаление из датасета и расследование случаев обхода фильтра. Нельзя классифицировать открытый массив как «безопасный» исключительно потому, что поисковая строка не включала чувствительные слова. Семантическое значение фрагмента, контекст и возможные выводы модели также существенны.

При выявлении повышенного риска сбор останавливается для затронутой категории до проверки правовых предпосылок. Аудитор запрашивает частоты и примеры ложных отрицаний фильтра, а не только заявленную долю обнаружения. Результат проверки формулируется с указанием чувствительности теста и остаточной неопределённости; нулевая ошибка на малой выборке не равна отсутствию риска.

Практическая граница между случайным и систематическим захватом особенно важна для статьи 9. Независимый аудитор должен установить, использует ли разработчик категории сайтов, с высокой вероятностью содержащие сведения о здоровье, сексуальной жизни или убеждениях, и предусмотрено ли их исключение до сбора. Прохождение записей через фильтр после их полного сохранения не всегда эквивалентно предупреждению обработки: сам сбор и первоначальное хранение уже представляют операции с данными. Поэтому проверка включает логи исключения доменов, архитектуру ранней фильтрации, способы ограничения доступа к карантинным материалам и критерии уничтожения остаточных копий. Обозначение «мы не планировали собирать такие сведения» требует подтверждения действиями, а не только заявлением о намерениях.

В проекте EDPB Guidelines 03/2026 обсуждается применимость аргументации суда ЕС по делу GC and Others к отдельным ситуациям побочного и остаточного сбора, но это не универсальная лицензия на обработку специальных категорий. Анализ остаётся обстоятельственным и учитывает реальную способность контролёра предотвращать сбор и дальнейшее распространение. Для содержимого, заведомо относящегося к таким категориям, необходима самостоятельная оценка применимых условий статьи 9(2) наряду с основанием статьи 6. При отсутствии условия обработки и невозможности надёжного исключения соответствующего сегмента методическим итогом может быть полный отказ от определённого класса источников. Нельзя подменять эту правовую развилку общим обещанием, что модель «всё равно не запоминает чувствительные данные».

14. Данные о судимостях и правонарушениях: отдельная статья 10

Статья 10 GDPR устанавливает особый режим обработки сведений, относящихся к уголовным приговорам и правонарушениям либо связанным мерам безопасности. Такая обработка на основании статьи 6(1) осуществляется под контролем официального органа либо в случаях, разрешённых правом ЕС или государства-члена с надлежащими гарантиями. Публичность судебного документа или упоминания уголовного дела не отменяет эти требования [1].

Для мониторинга международного розыска и медийных упоминаний риск заключается в автоматическом превращении обвинений, устаревших публикаций, оправдательных решений и совпадений имён в машинные признаки «связи с преступлением». Исследование должно строго различать установленную судом правовую квалификацию, стадию расследования, сообщение прессы и ошибочную идентификацию. Генеративная модель не является официальным источником сведений о существовании международного уведомления или приговора.

Если учебный корпус содержит такие сведения, разработчик должен показать специальное правовое обоснование, условия обработки и документированные защитные меры. Общее заявление об открытом характере судебных реестров не заменяет национальных ограничений доступа, повторного использования и срока хранения. Использование персональных данных для оценки конкретного человека требует самостоятельной проверки правового режима, в том числе статей 5, 6, 10, 22 GDPR, если они применимы [1].

Рекомендуемый контроль для аудита: выборка из правовых и новостных источников с маркерами точности личности, датой события, изменением процессуального статуса и признаком официальности. В спорной ситуации запись получает статус «не подтверждена», а не отрицательную характеристику лица. Независимая проверка должна приоритетно защищать лицо от автоматизированного распространения недостоверного обвинительного контекста.

Сведения о преступлениях и судебных делах составляют отдельную группу рисков. Статья 10 GDPR не является разновидностью статьи 9 и устанавливает специальные требования к обработке персональных данных об уголовных судимостях и правонарушениях или связанных мерах безопасности. Оценка включает национальную правовую основу, наличие контроля официального органа в предусмотренных случаях и пределы допустимой обработки. Нельзя предполагать, что общедоступная публикация обвинения, розыскного сообщения или копии судебного документа автоматически разрешает коммерческое извлечение таких сведений для генеративной модели. Кроме того, типовой фильтр названий заболеваний не решает проблему уголовно-процессуальной информации, которая часто содержится в свободном тексте.

Практический механизм должен различать подозрение, предъявление обвинения, вынесенный приговор, прекращение дела, отмену обвинения и отказ в экстрадиции. Эти события имеют различные юридические и фактические значения, и их смешение способно причинить серьёзный репутационный и имущественный вред. Независимая проверка изучает первичные документы, дату события, дальнейшую судьбу производства и наличие указания на неокончательный статус решения. Не следует превращать справочные или журналистские сообщения в «универсальный криминальный профиль». При наличии действующего

ограничения на обработку даже точные сведения могут быть недопустимы для конкретной цели. В отчёте требуется отдельно указать, относится ли вывод к качеству модели, законности исходного сбора, правомерности хранения либо к дальнейшему использованию результата финансовым учреждением или государственным органом: у этих вопросов разные правовые тесты.

15. Данные несовершеннолетних и уязвимых лиц

Защита детей имеет особое значение при балансе по статье 6(1)(f) GDPR: права и свободы субъектов, особенно детей, могут перевесить законный интерес. В тренировочные корпуса способны попадать школьные публикации, фотографии, форумные обращения подростков и семейные страницы. Техническая публичность таких материалов не обеспечивает предсказуемости их использования в генеративной модели [1][2].

Разработчик должен сформулировать запретные категории источников, режимы обнаружения несовершеннолетних, правила удаления при сомнении, ограничение использования эмоционально уязвимых материалов и доступные способы обращения законных представителей. Ошибки детекторов возраста могут затрагивать взрослых и несовершеннолетних одновременно; поэтому протокол должен учитывать ложные срабатывания и предусматривать безопасную проверку без дополнительного сбора лишних документов.

Методика ARGA предлагает не собирать особую подтверждающую информацию о возрасте из сторонних реестров только ради «уточнения» собранного корпуса, если это увеличит вмешательство в права. Лучше использовать ограничения по типу домена и контекста, минимизирующие правила сбора, удаление спорного класса и понятный канал сообщения о проблеме. Каждое исключение источника фиксируется с документированной причиной и возможностью пересмотра.

Если данные уже использованы для обучения, устранение последствий требует оценки поведения модели, возможности извлечения контента и способов запрета выдачи. Необходимо честно отделять технически подтверждённый эффект мер от предполагаемого. Обещание «полностью забыть всё» без проверяемого протокола может причинить дополнительный ущерб доверию и нарушить требования прозрачности.

16. Требование прозрачности и статья 14 GDPR при непрямом получении данных

Когда персональные данные собраны не у самого субъекта, статья 14 GDPR предусматривает раскрытие информации о контролёре, целях, основаниях, категориях данных, источнике и иных обстоятельствах обработки. При массовом web scraping индивидуальное уведомление может быть практически затруднено. Но трудность является предметом обоснования по конкретной норме, а не основанием отменить прозрачность [1][2][14].

В проекте EDPB 03/2026 отдельно обсуждается статья 14(5)(b) GDPR, когда предоставление сведений оказывается невозможным или требует несоразмерных усилий. Даже при выполнении условий исключения контролёр должен принимать надлежащие меры для защиты прав, а публичное размещение необходимой информации является предусмотренной законом мерой. Нельзя молча воспользоваться исключением и не публиковать ничего [2].

Практически понятная публичная политика должна описывать виды источников и периоды загрузки, категории персональных данных, цели, конкретный законный интерес, сроки хранения, риски и права субъектов, каналы возражений, механизм обработки исключений и контакт контролёра. Рекомендуется searchable source register или как минимум структурированный перечень доменов и типов ресурсов с объяснением неполноты списка, если публикация каждого URL невозможна [2].

Прозрачность должна начинаться до массового сбора и поддерживаться при изменениях целей, источников и архитектуры модели. Текст, спрятанный в сотнях страниц технических условий и лишённый практического механизма поиска, формально может существовать, но не обеспечивать эффективного информирования. Для проверки важно провести пользовательский тест: способно ли заинтересованное лицо понять ситуацию и инициировать запрос без юридических и технических знаний.

При непрямом получении данных контролёр должен представить информацию, предусмотренную статьёй 14 GDPR, включая категории данных, цели, правовое основание, соответствующие источники, получателей и период хранения либо критерии его определения. Обычная ссылка на общее «использование ИИ» не позволяет человеку понять, как его фотография, публикация или архивная запись могла попасть в обучающий корпус. Полезна двухуровневая структура уведомления: короткое понятное описание с каналом прав и расширенный технический раздел с категориями сайтов, версиями набора и видами операций. Однако даже подробная публичная страница не служит автоматически полноценной заменой индивидуального информирования во всех обстоятельствах. Применимость исключений оценивается отдельно, а способ публикации должен обеспечивать разумную доступность информации затронутым лицам.

Для международных систем важно учитывать языки и реальные способы обнаружения уведомления. Размещение англоязычного PDF в техническом разделе сайта вряд ли обеспечит значимую прозрачность для людей, чьи сообщения собраны в иных языковых сообществах. Можно оценивать, насколько легко найти уведомление из поисковой выдачи, содержит ли оно контакт для осуществления прав, объясняет ли непонятные технические термины, предусматривает ли канал сообщения о спорном источнике. Это прежде всего оценка эффективности предложенных мер. Соблюдение конкретных обязанностей статьи 14 определяется по праву и фактам. Для пилотной программы Observatoire ARGA целесообразен тест с независимыми участниками, которым предлагают найти ответ на три практических вопроса: почему мои данные могли использоваться, как проверить обстоятельства и как запросить исправление.

Сроки предоставления информации при непрямом получении данных определяет статья 14(3) GDPR: в разумный срок после получения данных, но не позднее одного месяца; если сведения используются для общения с человеком — не позднее первого сообщения ему; если предполагается раскрытие иному получателю — не позднее первого такого раскрытия. До дальнейшей обработки для иной цели применяется отдельное требование статьи 14(4). Исключения статьи 14(5), включая невозможность или несоразмерное усилие, применяются только при наличии доказанных условий и не образуют общего освобождения для всех больших обучающих корпусов [1][14].

17. Критерий несоразмерного усилия: исключение не должно поглотить правило

При анализе статьи 14(5)(b) нужно сопоставить реальную нагрузку на контролёра и последствия отсутствия уведомления для человека. Проект EDPB приводит факторы: количество субъектов, давность данных и имеющиеся меры защиты. Он отдельно предостерегает от автоматического использования исключения вне особых целей архивирования, научных или исторических исследований и статистики. Предложенное обоснование должно быть воспроизводимым и учитывать массив в целом [2].

Например, десятки миллионов старых документов без прямых идентификаторов могут создать необычно высокую стоимость индивидуального оповещения. Но из этого нельзя сделать вывод, что любой коммерческий сбор всегда считается «несоразмерным». Если контрольный набор содержит контактные данные небольшой группы, уведомить её может быть технически возможно. Единая юридическая записка для обоих случаев скрывает релевантные различия.

В журнале решения следует сохранить характеристики датасета, оценку способов связи, невозможности и рисков обратного установления личности, стоимость и качество альтернативного информирования, объяснение применения исключения и дополнительные меры безопасности. Любое доказательство несоразмерности должно быть привязано к конкретной версии массива и времени принятия решения.

Рекомендация Observatoire ARGA состоит в ведении двух взаимосвязанных документов: внутреннего мотивированного решения по статье 14(5)(b) и опубликованной понятной информации для субъектов. Внутренний отчёт не должен раскрывать в публичной версии чувствительные сведения либо делать возможной повторную идентификацию. При несоответствии одной части другой вопрос подлежит дополнительной проверке.

18. Минимизация: до, во время и после извлечения

Статья 5(1)(c) GDPR требует адекватности, релевантности и ограничения данных необходимым объёмом. В стадии планирования это означает исключение нерелевантных доменов, разделов и типов документов. При извлечении — ограничение полей, глубины обхода, временного периода, частоты обращений и идентификаторов. После сбора — очистку, дедупликацию, удаление ненужных полей и обоснованное сокращение сроков хранения [1][2].

Систематическое применение правила minimisation важнее разовой очистки перед обучением, поскольку предварительно собранные данные могут сохраняться в промежуточных хранилищах, резервных копиях и логах. Если запрещённый набор был «вычищен» из финального файла, но остался в доступном разработчикам raw bucket, цель минимизации не достигнута. Аудит должен проверять копии, права доступа и retention-процедуры.

EDPB предлагает рассматривать синтетические, псевдонимизированные и анонимизированные данные, а также ограничение веб-сайтов, структурно содержащих особые категории. При этом нельзя считать синтетический материал заведомо свободным от риска: генерация по реальным образцам способна воспроизвести личные сведения. Критерии

приемлемости должны включать тесты на similarity и leakage и подтверждённый порядок ликвидации исходных копий [2][15].

Важный управленческий вопрос заключается в выборе момента фильтрации. Фильтрация до извлечения, как правило, лучше ограничивает сам факт обработки; очистка после получения остаётся необходимой защитой для ошибочного захвата. Обе линии контроля дополняют друг друга, но не заменяют законное основание исходной операции.

Минимизация на стадии проектирования начинается с отрицательного списка. Помимо разрешённых категорий сайтов нужно заранее перечислить источники, которые по своей природе часто содержат несоразмерно чувствительные сведения: ресурсы взаимопомощи больным, детские сообщества, агрегаторы частных адресов, форумы переживших насилие, архивы судебных обвинений и сайты, специально возражающие против автоматического сбора. Отдельные критерии могут регулировать географию, возраст публикаций, структуру URL и наличие прямых идентификаторов. При разработке не следует ограничиваться regex-фильтрами: имена, сведения о состоянии здоровья и процессуальном статусе выражаются контекстно, а классификаторы ошибаются при редких языках и непрямых формулировках. Поэтому фильтрация нуждается в независимой проверке на ложноотрицательные результаты.

После сбора предлагается многоуровневая обработка: автоматическое выделение потенциально рискованных данных, ограниченный карантин с повышенным контролем доступа, целевая проверка и удаление либо обезличивание, если оно реально возможно. Каждая стадия сохраняет счетчики исключённых записей по категории, долю сомнительных случаев и версию правил. Проблема «чёрного ящика» возникает, когда подрядчик возвращает только итоговое число удалений без исходной базы расчёта и механизма проверки. Из статистики «удалено два миллиона записей» нельзя вывести, сколько рисковых документов осталось. Для рецензии следует дополнительно использовать независимую выборку неотфильтрованных записей, отдельное измерение по языковым и содержательным классам и периодическое воспроизведение теста после обновления правил.

19. Достоверность, время и проблема устаревшей информации

Принцип accuracy по статье 5(1)(d) GDPR требует, чтобы персональные данные были точными и, где необходимо, актуальными. Для генеративного обучения данные с форумов, новостей и старых реестров могут отражать состояние дел, которое затем изменилось. Модель, соединяющая старый текст с текущим вопросом, способна выдавать ошибочную юридическую или биографическую характеристику [1].

EDPB в проекте Guidelines 03/2026 рекомендует выбирать надёжные источники, фиксировать время извлечения и проверять данные до использования при обучении. Это особенно важно в санкционных списках, международном розыске, коммерческом комплаенсе и уголовно-процессуальных материалах, где запись могла быть исправлена, аннулирована или ограничена по доступу [2].

Следует хранить раздельные атрибуты: время события, время публикации, время загрузки и время последней проверки. Один timestamp «scraped_at» не доказывает актуальность отражаемого события. Процедура проверки должна различать невозможность обновления, конфликт между источниками и подтверждённое изменение статуса. При высокой

значимости последствий необходим источник с компетенцией по конкретному юридическому факту.

В измерении качества ошибкой будет доля текстов «без опечаток» вместо показателей точности спорных утверждений о лицах. Предлагаются метрики: число источников с отсутствующей датой, доля негативных персональных утверждений без первичного подтверждения, латентность обновления изменившейся записи и доля запретов на использование устаревших сведений, фактически применённых к обучающим экземплярам.

20. Оценка воздействия на защиту данных и документация риска

Статья 35 GDPR требует DPIA, когда тип обработки с высокой вероятностью приводит к высокому риску правам и свободам физических лиц. Масштабный автоматизированный сбор, инновационные технологии, чувствительные данные и непрозрачное использование способны формировать соответствующие признаки; однако необходимость DPIA устанавливается по конкретным фактам и применимым перечням надзорных органов [1][16].

Надлежащая DPIA описывает операции и цели, необходимость и соразмерность, риски для лиц и предусмотренные меры их устранения. Особое значение имеют риски, которые нельзя обнаружить лишь по наличию незаконного доступа: некорректная повторная идентификация, репутационный ущерб от генеративного вывода, ошибочное санкционное совпадение, массовое непрямое профилирование, невозможность воспользоваться возражением и скрытая персистентность информации в модели.

В методике оценки следует явно отделять вероятность воздействия от тяжести вреда и качества имеющихся свидетельств. Высокая неопределённость не должна автоматически снижать уровень риска. Пороговые шкалы, предлагаемые в приложении, служат организационным инструментом; они не являются законодательно установленными категориями. Если остаточный высокий риск не может быть смягчён, могут возникнуть условия предварительной консультации надзорного органа по статье 36 [1].

Оценку пересматривают при изменении целей, классов источников, подключении нового data broker, значительном увеличении масштабов или существенном изменении модели. Переписанный задним числом документ без истории согласований не обеспечивает подотчётности. Независимый эксперт должен иметь доступ к исходной оценке, журналу обновлений и объяснению, почему контрольные меры признаны достаточными.

Решение о необходимости DPIA по статье 35 GDPR нельзя механически вывести из единственного признака «обучение искусственного интеллекта». Требуется оценить характер, масштаб, контекст и цели обработки, в том числе новые технологии, возможные последствия систематической оценки людей, объединение массивов, уязвимые группы, чувствительные сведения и возможности правовой защиты. Для широкомасштабного непрозрачного сбора персональных данных риск может быть высоким, но окончательный вывод должен опираться на реальную схему проекта. Документ оценки описывает предполагаемые операции, необходимость и соразмерность, риски для прав и свобод, защитные меры, ответственных и остаточный риск. Формальная матрица с цветами без аргументации недостаточна.

Если после разумных мер остаётся высокий риск, который контролёр не способен снизить, вопрос предварительной консультации надзорного органа рассматривается по статье 36

GDPR. Процедура не заменяется утверждением «наш DPO одобрил запуск». Независимый аудит также не вправе обещать, что наличие DPIA гарантирует законность любых последующих операций. Для практической проверяемости предлагается привязать каждый значимый риск к наблюдаемому тесту, доказательству и условию пересмотра. Например, риск узнавания лиц через сочетание биографических деталей связывается с тестированием связывания нескольких источников; риск запоминания персональных фраз — с контролируемой оценкой извлечения. Частота повторной оценки зависит от изменений корпуса, модели, источников и характера предоставляемого сервиса.

21. Право на доступ, исправление, стирание и ограничение обработки

Права субъектов закреплены в статьях 12–22 GDPR и зависят от условий каждого права. При обращении, касающемся web scraping, разработчику требуется определить, какие данные о лице обрабатываются, в каких версиях наборов они хранятся, где доступны исправления и как запрос касается используемой модели. Отсутствие общей поисковой системы по источникам затрудняет исполнение законных обязанностей, но не снимает их [1][10].

Право на стирание по статье 17 не является безусловным во всех обстоятельствах, как и возражение по статье 21 не имеет одинаковых последствий для всех правовых оснований. Однако контролёр обязан обеспечить разумный и своевременный процесс рассмотрения запроса, мотивированно определить применимое исключение и документировать технические меры. Доступная форма запроса должна различать удаление из исходного корпуса, запрет повторного извлечения и меры по снижению выдачи данных моделью.

Реализация исправления требует аккуратного учёта контекста. Для судебной новости может существовать юридически значимый исторический факт и последующее исправление. Модель, обученная только первоначальной публикации, способна продолжать транслировать старый статус. Комплаенс-процедура должна выявлять конфликт версий и предусматривать обоснованное ограничение рискованных выводов, а не автоматически переписывать архивный документ.

Аудиторский показатель эффективности прав не сводится к количеству закрытых заявок. Следует измерять полноту поиска релевантных копий, правомерность отказов, обоснованность выбранных мер, проверяемость технического результата и доступность апелляции к человеку. Субъект не должен быть вынужден доказывать «точный URL, попавший в датасет», если контролёр сам по своей архитектуре не располагает такой прослеживаемостью.

При исполнении запросов субъекта данных первым шагом является определение объекта и контролёра. Не всякое обращение «удалите меня из ИИ» требует одинакового технического решения. Организация выясняет, имеется ли запись в исходном архиве, включалась ли она в очищенный набор, используется ли в retrieval-индексе, была ли задействована при fine-tuning, сохраняется ли в журнале запросов и способна ли модель воспроизводить сведения без внешнего контекста. Ответ должен различать подтверждённое наличие данных, проведённые действия и пределы технической проверки. Если часть сведений не может быть найдена надёжно, недопустимо выдавать неопределённость за доказательство отсутствия обработки.

Для практической защиты прав важно исключить побочное раскрытие сведений через сам канал обращения. Запрос на доступ может поступить от лица с совпадающим именем либо от постороннего, который пытается узнать, есть ли в наборе сведения о конкретном человеке. Метод подтверждения личности должен быть соразмерен риску и не превращаться в избыточный сбор документов. На стороне контролёра сохраняются сроки рассмотрения и порядок ответов, предусмотренные GDPR, а сложность работы с модельными весами не должна автоматически использоваться как основание для бессрочного молчания. Для споров по уголовным и санкционным совпадениям следует обеспечить путь пересмотра человеческим экспертом и инструкцию о том, какие первичные доказательства могут помочь восстановить точный контекст.

Процедурный срок по статье 12(3) GDPR: сообщение о действиях по запросам субъекта предоставляется без необоснованной задержки и, как правило, не позднее одного месяца после получения обращения. При необходимости срок может быть продлён ещё максимум на два месяца из-за сложности или числа запросов, но о продлении и его причинах контролёр сообщает в течение первоначального месяца. При решении не предпринимать действий применяются требования статьи 12(4) о мотивированном ответе и средствах защиты; дополнительная проверка личности возможна в установленных статьёй 12(6) пределах. Эти сроки следует закрепить в реестре обращений и договорных SLA, не обещая технически неподтверждённого удаления данных из весов модели [1].

22. Возражения против обработки и учёт ограничений сайтов

Возражение субъекта против обработки по статье 21 GDPR и ограничения автоматического извлечения, обозначенные владельцем сайта, юридически не совпадают. Первое является правом физического лица, второе может выступать сигналом об ожиданиях, договорных правилах или средствах контроля контента. Оба типа информации необходимо документировать, но исключение домена из crawler не гарантирует выполнения индивидуального запроса о конкретном лице [1][2].

EDPB и CNIL уделяют внимание техническим механизмам, которыми сайты выражают нежелание автоматического сбора, включая robots.txt и CAPTCHA. Их игнорирование увеличивает комплаенс-риск; однако наличие файла robots.txt само по себе не решает вопрос об исключении текста из ранее полученного набора. Поставщик должен отдельно контролировать upstream- и downstream-версии корпуса [2][6].

Предлагаемая архитектура ведёт deny registry со scope: URL, домен, тип контента, конкретный субъект, идентификатор документа и версия. Каждая запись содержит основание запрета, дату вступления, статус исполнения и связанные операции: не собирать повторно, исключить из следующего обучения, исследовать существующие копии, отследить жалобу. При отсутствии возможности соотнести запрос с техническим объектом решение не маркируется как полностью исполненное.

Для обоснованного несогласия с возражением требуется отдельная правовая оценка и мотивированный ответ. Нельзя автоматически отказывать заявителю по причине широких общественных интересов разработчика или на основании того, что у поставщика ещё нет механизма стирания влияния данных из весов. Невозможность технологии и отсутствие юридической обязанности в конкретной форме должны быть установлены отдельно, а не смешиваться.

23. Модельные веса, memorisation и вопрос «забывания»

Обучение переводит исходные наблюдения в параметры модели; при этом отдельные тренировочные фрагменты могут оставаться воспроизводимыми при определённых запросах. Технический риск зависит от размера, повторяемости, редкости и характера данных, алгоритма обучения и метода атаки. Тест на verbatim extraction, проверка чувствительных последовательностей и оценка генеративных выводов могут выявлять нарушения конфиденциальности, но не исчерпывают возможные способы выявления информации [3] [10].

В правовом отношении удаление объекта из raw dataset, из curated training set, из индекса retrieval и из резервных копий являются разными операциями. Изменение весов через machine unlearning, retraining или обновление модели требует отдельного подтверждения эффекта. Утверждение «запрос на удаление выполнен» должно указывать, какие уровни действительно обработаны, а какие остаются предметом риска и почему [1].

Часто менее рискованным промежуточным решением бывает исключение спорных фрагментов из последующего дообучения и отдельная защита от выдачи чувствительных данных. Но защитный фильтр выдачи не равнозначен стиранию персональных данных из всех обработок, и его можно обойти или он может ухудшить качество законных ответов. Доклад рекомендует тестировать остаточный риск, фиксировать ограниченность мер и по необходимости требовать архитектурного изменения.

Эксперимент с канареечными последовательностями или иными тестовыми идентификаторами допустим только в законно сформированной тестовой среде, чтобы проверка сама не становилась новым источником неправомерной обработки. При исследовании чужой модели предпочтительны разрешённые интерфейсы, собственные искусственные данные и соблюдение условий доступа. Отчёт должен отличать гипотезу о запоминании от доказанного факта разглашения.

Техническое требование «забыть» сведения не сводится к единственной операции. В архитектуре с retrieval-augmented generation достаточно быстро может измениться индекс, если проблема действительно возникла из него. В системах с fine-tuning возможны возврат к предыдущему checkpoint, повторная настройка на очищенном наборе либо меры ограничения выдачи. При предобучении крупной модели влияние отдельных образцов на веса трудно изолировать, а методы machine unlearning требуют самостоятельной проверки эффективности и нежелательных побочных эффектов. Удаление из исходного корпуса влияет прежде всего на дальнейшие циклы и не подтверждает автоматически изменение ранее выпущенной модели. Нельзя без квалификации обещать универсальное полное удаление из параметров.

Для независимого заключения полезна градация доказательств. Уровень A: подтверждено физическое удаление известной записи из контролируемого хранения. Уровень B: показано исключение её из будущих обучающих сборок. Уровень C: выполнены тесты поведения конкретной версии модели по заранее установленному протоколу. Уровень D: есть исследовательские свидетельства снижения влияния данных, но нет технической возможности доказать нулевой остаток во всех условиях. Формулировка результата должна соответствовать уровню, включая неуспешные тесты и пределы поиска. Особенно важно не

распространять оценку одной доступной версии модели на все копии, зеркала и варианты, находящиеся вне контроля проверяемой организации.

24. Анонимизация, псевдонимизация и повторная идентификация

Псевдонимизация снижает непосредственную доступность идентификаторов, но при наличии дополнительной информации и возможности соотнести запись с лицом остаётся обработкой персональных данных. Анонимизация требует оценки идентифицируемости с учётом объективного контекста и разумно вероятных средств. Простая замена имени на хэш, случайный номер или инициал не гарантирует анонимности [1][15].

EDPB в июле 2026 года опубликовал проект Guidelines 02/2026 on Anonymisation, рассматривающий contextual и simplified approaches и три теста: выделение записи, связывание и вывод информации о лице. Их результат не заменяет общий юридический анализ; статус проекта на дату доклада остаётся консультационным. При оценке большого тренировочного корпуса необходимо различать состояние данных для непосредственного контролёра, получателя и потенциального атакующего [15].

Практический тест анонимизации включает поиск редких сочетаний атрибутов, связь с внешними источниками, оценку участия отдельных лиц и возможность восстановления личности по выводу модели. Итоги фиксируются с описанием сценариев атак, версии внешних данных, допущений и ограничений. Сравнение с закрытой базой специалистов без права доступа не может быть описано как реально проведённый тест.

В отличие от декларации «данные анонимизированы», доказательная оценка требует раскрыть модель угроз, права доступа к ключам, режим удаления исходных данных и результаты повторной проверки при появлении новых связующих источников. При неопределённом статусе данных более безопасно продолжать применять защитные меры GDPR до мотивированного вывода, не выдавая этот подход за юридическое определение анонимности.

25. Каталог происхождения: проверяемые свидетельства для каждого источника

Происхождение данных представляет собой цепь от исходного URL или версии файла до конкретного набора обучения и модели. Реестр должен сохранять идентификатор ресурса, доменную категорию, период доступа, дату загрузки, статус публичности, применённое основание обработки, ограничения и последовательность преобразований. При этом для публикации можно использовать агрегированные или ограниченные сведения, если раскрытие прямых ссылок создаёт дополнительный вред [1][2].

Криптографический хэш позволяет установить, что представленные копии идентичны между проверками, но не подтверждает законность извлечения или подлинность первичного источника. Электронный журнал с подписанными временными метками помогает последовательности операций, но при отсутствии независимого подтверждения может отражать лишь утверждения самого оператора. Поэтому необходимы выборочные сопоставления с исходными системами, журналами сетевых запросов и договорами поставщиков.

Каждый преобразованный фрагмент должен наследовать источниковые ограничения, включая запрет повторного использования. При агрегации тысяч документов внутренние идентификаторы позволяют исключать отдельные линии происхождения без раскрытия лишних персональных подробностей. Архитектура данных должна предусматривать возможность корректировать ошибочную атрибуцию и фиксировать конфликт источников вместо принудительного выбора одной версии.

Независимая проверка определяет минимально достаточную трассировку: от выбранного результата модели или training example назад к manifest, операциям очистки и источнику. Невозможность трассировки является ограничением аудита и предметом рекомендаций. Она не даёт оснований автоматически утверждать, что весь материал получен незаконно.

Надёжная прослеживаемость предполагает связку минимум четырёх идентификаторов: источника, конкретной записи, версии датасета и версии модели. При изменении ссылок сайта необязательно хранить бессрочную полную копию спорного текста; в зависимости от цели и требований безопасности могут применяться хэши, ограниченные метаданные и отдельный архив с контролируемым доступом. Каждый метод имеет пределы: криптографический хэш подтверждает совпадение конкретной последовательности байтов, но не доказывает её юридическую достоверность и не объясняет происхождение. Подпись поставщика подтверждает подписанта, но не превращает недоказанные гарантийные заявления в установленный факт. Поэтому реестр содержит сведения о процедуре получения, доверенном источнике и подтверждённых проверках.

Особые требования предъявляются к цепочке последующих изменений. Допустим, первоначальный документ включён в пакет D1, после удаления идентификаторов получен D2, после дедупликации D3, а после объединения с другим корпусом сформирован D4. Если сохранены только итоговый номер D4 и общая фраза «очищено», невозможно восстановить, какой фильтр затронул спорную запись и какие модели успели её использовать. Предлагается вести журнал преобразований с кодом процедуры, временем, версией конфигурации, количеством исключённых объектов и связывающим идентификатором пакета. Для независимого аудита возможен ограниченный доступ только к необходимым фрагментам, при котором доказуемость не требует передачи всего корпуса эксперту. Правомерность хранения артефактов также подлежит отдельной оценке.

26. Международные передачи и доступ третьих стран

При доступе к персональным данным из третьей страны применимость главы V GDPR определяется фактической структурой передачи, ролями сторон и правовым контекстом. Размещение сервера в ЕС не исключает международную передачу, если уполномоченный иностранный подрядчик получает доступ к данным; одновременно не всякое использование зарубежного облачного бренда само по себе означает незаконную передачу [1].

Контролёр документирует страны получателей, правовой механизм передачи, субобработчиков, применимые договорные условия, меры безопасности и возможность доступа по законодательству иностранных государств. Необходимо отличать правовое основание самой обработки по статье 6 от механизма международной передачи по главе V: выполнение одного не устраняет другого.

Для глобальных корпусов важны репликация при машинном обучении, временные рабочие копии в облачных средах, логирование запросов операторов и shared debugging. Инвентаризация должна охватывать не только основной training bucket, но и pipeline оркестрации, мониторинговые сообщения и изображения из внешних инструментов. Каждый технический маршрут проверяется по фактам.

Независимый аудитор вправе запрашивать доказательства оценки трансграничных рисков и договорной ответственности, но не должен подменять компетентное судебное установление правомерности международной передачи общей шкалой «низкий/высокий». Когда доступны лишь заявления поставщика, это ограничение данных об исследовании должно быть явно раскрыто.

27. Безопасность набора данных и обязанности подрядчиков

Статья 32 GDPR требует надлежащих технических и организационных мер с учётом рисков обработки. Для web scraping это предполагает сегментацию raw, cleaned и training сред, разграничение ролей, журналирование доступа, управление ключами, контроль экспорта, шифрование там, где оно уместно, и ограничение сроков хранения резервных копий. Наличие одного «защищённого» облака не обеспечивает безопасность полного набора [1].

Отдельную угрозу представляют токены доступа к сайтам, cookies, случайно собранные конфиденциальные документы, устаревшие резервные экземпляры и подсистемы аннотирования с доступом подрядчиков. Нельзя опираться на предположение, что обучение на публичных данных делает все полученные фрагменты пригодными к неограниченному внутреннему распространению. Описанные риски требуют тестирования доступа на каждой критической стадии.

В соглашении с обработчиком по статье 28 GDPR должны быть реально обеспечены помощь при реализации прав субъектов, ограничения субподряда, порядок удаления и доступ к доказательствам. Практически полезен отдельный реестр субобработчиков с изменениями регионов хранения и категорией доступа. В случае выявления нарушения безопасности применимость обязательств статей 33–34 определяется отдельно от процедуры расследования незаконного scraping.

Независимая оценка безопасности должна включать проверку действительности контрольных записей, а не только наличие сертификатов. Допустимо использовать внешние удостоверения как элемент уверенности, однако область их охвата проверяется: сертификат управления информационной безопасностью не подтверждает сам по себе законность обработки каждого URL.

28. Интеллектуальная собственность: параллельный, а не альтернативный правовой режим

Скрейпинг может затрагивать произведения, охраняемые авторским правом, права изготовителей баз данных и договорные режимы доступа. В ЕС статья 4 Директивы (EU) 2019/790 предусматривает исключение или ограничение для text and data mining законно доступных произведений при условии, что правообладатель надлежащим образом не зарезервировал соответствующее использование. Указанное исключение применяется в пределах его условий и национальной имплементации [18].

Это не является освобождением от GDPR, поскольку персональные данные и авторское право охраняют разные объекты и интересы. Произведение может не содержать персональных данных, но быть защищённым авторским правом; личное сообщение может не иметь значимой творческой оригинальности, но подпадать под GDPR. В сборнике могут одновременно присутствовать оба режима, а также нормы о коммерческой тайне и технической защите доступа.

Для независимого аудита рекомендуется разделять два листа доказательств: законность обработки персональных данных и допустимость использования контента для TDM/обучения. Отсутствие результата по одной линии не восполняется положительным заключением по другой. При обнаружении rights reservation следует документировать технологию выявления, тип сигнала, дату и влияние на фактически используемые версии корпуса.

Нельзя трактовать запись в публичной сводке обучающих данных по AI Act как опубликование всех защищённых документов либо как правовую лицензию на них. Требование достаточной детализации сводки направлено на прозрачность, а не на передачу права воспроизводить весь исходный корпус. Конкретные условия применения исключений или заключённых лицензий подлежат самостоятельной оценке [11][12][18].

29. AI Act: документация GPAI, авторское право и публичная сводка

Статья 53 AI Act предписывает поставщикам general-purpose AI моделей создавать и обновлять техническую документацию, предоставлять определённые сведения downstream-поставщикам, иметь политику соблюдения права ЕС об авторских и смежных правах и публиковать достаточно подробную сводку контента, использованного для обучения, в соответствии с шаблоном AI Office. Правила применяются согласно действующему календарю и переходным положениям; необходимо учитывать специальные исключения для некоторых открытых моделей и отдельные обязанности моделей с системным риском [11] [12].

Для целей web scraping юридически значимо, что техническая документация по AI Act и record of processing по GDPR имеют разные назначение, адресатов и состав. Обязательство опубликовать summary of training content не означает обязанность раскрыть общедоступно каждый личный профиль или защищённый договор; но и не подменяет проверяемого внутреннего реестра происхождения и оснований обработки персональных данных.

В докладе предлагается карта соответствия: категория источников в AI Act summary ↔ внутренний dataset manifest ↔ запись legal basis/retention ↔ лицензия или rights reservation ↔ протокол фильтрации ↔ отчёт по тесту leakage. Такая связка позволяет избежать ситуации, когда публичный список категорий выглядит убедительно, но не имеет связи с реальными моделями и историями загрузки.

Органы, уполномоченные применять AI Act, и надзорные органы по защите данных выполняют различные функции. Проверка compliance по AI Act не заменяет установленной GDPR надзорной процедуры. Именно поэтому рекомендации по единому «сертификату законного scraping» не должны выдаваться за существующую государственную аккредитацию или универсальное разрешение.

Календарь и исключения требуют отдельной проверки. Обязанности главы V для соответствующих поставщиков GPAI-моделей начали применяться с 2 августа 2025 года. Для моделей, выведенных на рынок до этой даты, статья 111(3) устанавливает переход до 2 августа 2027 года. Статья 53(2) освобождает некоторые модели с действительно свободной открытой лицензией и публично доступными параметрами от обязанностей по пунктам 53(1) (a) и (b), но не от политики соблюдения авторского права и публичной сводки обучающего контента по пунктам (c) и (d); для моделей с системным риском данное освобождение не действует. Обязанности GPAI не следует смешивать со сроками применения отдельных правил для высокорисковых AI-систем. Установление соответствия AI Act само по себе не доказывает законность обработки каждой исходной записи в рамках GDPR [11][12][19].

Для отчёта Observatoire ARGA полезна «двойная матрица доказательств»: отдельно анализируются обязанности AI Act и отдельно GDPR. В столбце AI Act оценивают наличие документации о происхождении и содержании обучающих данных, политику авторского права и публичное описание. В столбце GDPR устанавливают цели, роли контролёров, основание статьи 6, дополнительные условия статей 9/10, информирование, минимизацию и права субъектов. Пересечение может давать полезные доказательства, но не подменяет раздельные тесты. Например, описание источника «публичные новостные сайты» помогает понимать класс контента, но не показывает, исключались ли неправомерные архивные обвинения и как исполнитель реагирует на исправление судебной записи.

30. Французский подход: CNIL и защитные меры при сборе в сети

CNIL опубликовала рекомендации по разработке систем ИИ и отдельно рассмотрела сбор публично доступных данных через web scraping, включая законный интерес и меры уменьшения вмешательства. Французский регулятор подчёркивает, что сам scraping не запрещён GDPR как технологическая операция, но законность зависит от конкретной цели, условий обработки и защитных мер [6][8].

Для практического проекта во Франции важно документировать факторы уязвимости источников, использование исключающих списков, публичное информирование и фильтрацию информации, не требуемой для цели обучения. Сигналы сопротивления сайта автоматическому сбору, в том числе технические, нельзя бездумно игнорировать. При этом юридическое содержание рекомендации CNIL должно отделяться от прямого текста GDPR, особенно когда национальная практика развивает общие положения [6].

Доказательная документация должна быть на понятном языке, описывать ответственного контролёра и реальные способы реализации прав. Политика на французском сайте, которая перечисляет лишь корпоративные интересы и не описывает источники персональных данных, не гарантирует выполнение принципа прозрачности. Аудитор проверяет доступность и актуальность информации, а не только наличие веб-страницы.

Методология ARGA предлагает использовать подход CNIL как проверочный минимум для операций, затрагивающих Францию, но не переносить его автоматически как обязательное право всех третьих стран. При международных трансграничных проектах необходимо установить применимые юрисдикции и материальную компетенцию каждого регулятора.

31. Великобритания: позиция ICO и сравнительные ограничения

Великобритания имеет собственный режим UK GDPR и регулирование защиты данных; позиция ICO полезна как сравнительный материал, но не является нормативным толкованием EU GDPR для всех государств Союза. В ответе на консультационную серию о генеративном ИИ ICO указывает, что legitimate interests в типичной нынешней практике mass web-scraping является реалистичным основанием лишь при прохождении трёхчастного теста, особенно необходимости и справедливого баланса [7].

ICO подчёркивает риск непрозрачной, масштабной и часто невидимой обработки. Если менее вмешивающийся способ достижения той же цели доступен, контролёр должен объяснить, почему его нельзя использовать. Аналогично существенные недостатки прозрачности, из-за которых человек не может реализовать свои права, делают баланс интересов особенно проблемным [7].

Для сравнительного исследования важно не подменять индивидуальную оценку национального закона универсальной таблицей «UK разрешает — ЕС запрещает». Реальные различия могут заключаться в отраслевых методиках надзорных органов, процессуальных инструментах, судебной практике и деталях регулируемого процесса. Рекомендуется применять к каждому кейсу карту норм, дат действия и компетентных органов.

Международная группа, использующая один и тот же датасет для моделей в ЕС и Великобритании, должна проводить оценку на уровне конкретных операций, имея в виду разные географические границы и договорные механизмы. Единый глобальный документ может содержать совместимые общие меры, но обязанность доказать применимость и соблюдение соответствующего режима остаётся.

32. Непрозрачные поставщики наборов: due diligence до закупки

Закупка «готового очищенного корпуса» может переносить техническую работу, но не даёт покупателю автоматического подтверждения законности исходного сбора. Проверка должна выяснять кто собирал данные, для какой цели, по какому основанию, какие ограничения источников учитывались, содержатся ли специальные категории или уголовные данные, как обеспечиваются права субъектов и допускается ли независимая проверка [1][2][4].

Особое внимание требуется при покупке материалов, собранных цепочкой брокеров. Если первичный поставщик не сохранил дату и ограничения источника, последующие гарантии легко становятся чисто договорными. Рекомендуется требовать provenance manifest на уровне категорий и репрезентативных выборок с возможностью защищённой верификации, а также подтверждения исключения материалов по жалобам.

Невозможность раскрыть весь корпус по коммерческим причинам не отменяет потребность в достаточном доказательстве. Могут применяться конфиденциальные комнаты данных, независимый эксперт, проверка выборки без копирования чувствительных материалов и аттестация конкретных процедур. Однако аудитор должен честно различать неполный доступ и полное подтверждение отсутствия нарушений.

Контракт содержит право на целевой аудит, определение допустимых источников, уведомление о существенных изменениях и инцидентах, сроки исполнения законных

запросов, порядок уничтожения копий и распределение расходов на исправления. Все договорные гарантии сопоставляются с реальной способностью их исполнить технически.

Процедура закупки обучающего корпуса должна начинаться до подписания лицензионного соглашения. Покупатель формирует риск-ориентированный список вопросов о происхождении, времени, частоте обновления, прежних целях сбора, типах сайтов, возможных специальных категориях, исключениях и правах субъектов. Если брокер отвечает на всё заверением «полная анонимизация», требуется узнать, какие идентификаторы удалены, как оценивалась возможность повторной идентификации, проводились ли связывающие тесты и кому доступны исходные версии. Отдельно исследуют условия, в которых переданные записи могут фактически вновь стать персональными после интеграции с другими системами. Невозможность раскрыть коммерчески чувствительную методику не всегда исключает более ограниченную независимую проверку в защищённой среде.

Контрактные средства имеют практическое значение, когда они привязаны к действиям: право на выборочный аудит, обязанность сообщать о новых обнаруженных категориях, запрет использования отдельных сегментов, уведомление об изменении источников, механизм срочного отзыва dataset version и содействие при обращениях субъектов. Без этого соглашение порождает лишь возможность взыскать убытки после обнаружения нарушения. Важен и вопрос о передаче третьим странам, если брокер фактически предоставляет доступ из внешней инфраструктуры. Решение о закупке следует фиксировать как отдельный управленческий акт с оставшимися сомнениями; оно не должно замещать юридическую оценку первичной и дальнейшей обработки.

33. Публичная отчётность: достаточно подробная, но не опасная

Прозрачность тренировочных источников нужна людям, регуляторам, правообладателям и исследователям, однако безоглядная публикация перечня всех URL способна повторно распространить чувствительные сведения. Следовательно, необходим дифференцированный подход к раскрытию: публичное описание категорий, периодов и механизмов прав; более детальный защищённый комплект для компетентного регулятора; выборочные доказательства для независимого аудитора при законных гарантиях доступа [1][2][11].

«Сводка источников» и «перечень всех персональных записей» не совпадают. В отчёте следует объяснять метод категоризации сайтов, охват и ограничения списка, период покрытия и изменения версий. Доверие повышают не декоративные проценты, а воспроизводимость того, что включено в знаменатель, чем подтверждена классификация и сколько объектов не имеют полноценного provenance.

Предлагается публиковать отдельно число документированных категорий, долю записей со временем сбора, долю источников с решением о допустимости, число обработанных заявлений по правам субъектов и методологию выборочной проверки. Если исходные данные не позволяют оценить знаменатель, показатель не публикуется с фиктивной точностью; прямо указывается отсутствие статистической основы.

В случаях реального вреда указывается, какой тип последствий затронут: неправомерное раскрытие, устаревшее представление о лице, скрытая повторная идентификация, невозможность реализации возражения. Данные заявителей, характер которых позволяет

вновь их идентифицировать, не включаются в публичные примеры без самостоятельного правового анализа.

34. Независимый аудит: предмет проверки и пределы заключения

Независимый аудит web scraping должен проверять четыре взаимосвязанных блока: юридическую мотивацию обработки, происхождение и разрешённость источников, реальную работу технических фильтров и действенность средств защиты субъекта данных. Ни один блок нельзя заменить заявлением разработчика о наличии внутренней политики. Однако заключение не имеет силы судебного решения или обязательного предписания надзорного органа, если закон не наделяет аудитора соответствующими полномочиями.

Объектом контроля выступает определённая версия коллектора, датасета, training run и интерфейса рассмотрения запросов. Для каждой версии аудитор формулирует период и объём покрытия, методы выборки, число просмотренных источников и классы ограничений доступа. Запись «компания соблюдает GDPR» без определённого предмета исследования не соответствует надлежащей методологии.

В условиях конфликта интересов, когда аудитор одновременно консультирует разработчика или представляет заинтересованную сторону, требуется раскрытие финансовых и организационных связей, независимая оценка риска самооценки и при необходимости отвод. Отсутствие формального статуса аккредитованного органа нельзя маскировать официально звучащими печатями и названиями сертификатов.

Выводы оформляются градуированно: «подтверждено проверенными материалами», «частично подтверждено с ограничениями», «не установлено ввиду отсутствия доказательств», «обнаружено конкретное несоответствие». Такой подход снижает риск необоснованных обвинений и защищает возможность заинтересованных сторон представить возражения.

Независимый аудит должен фиксировать цель, объём, доступные источники доказательств и степень проверки до начала сбора материалов. Полный доступ к конфиденциальному корпусу не всегда нужен и может создавать новый риск незаконного раскрытия. Возможны проверка выборочных манифестов, воспроизводимость фильтров на специально подготовленном стенде, анализ журналов и контрольные запросы к модели. Но ограниченный доступ требует честного названия объёма работ: внешняя оценка выходов модели не позволяет подтвердить весь состав тренировочного набора, а анализ процедур заказчика не доказывает, что они фактически применялись к каждому файлу. В заключении различаются документальное подтверждение, воспроизведённый эксперимент, косвенный индикатор и сведения со слов стороны.

Структура итогового аудиторского акта включает выявленные критерии, перечень подтверждающих артефактов, классификацию существенности, возможные альтернативные объяснения, пределы проверки и действия по устранению. Отдельный раздел посвящается независимости эксперта: полученные вознаграждения, связи с разработчиком, работа над спорной архитектурой, право клиента влиять на публикацию и доступ к конфликтующим данным. Организация, которая одновременно разрабатывает модель, консультирует её поставщика и публично оценивает законность, должна применять ясные институциональные

барьеры. Требование независимости здесь не означает отсутствие любого финансирования, но предполагает раскрытие интересов и механизм предотвращения влияния на выводы.

35. Статистическая выборка: почему несколько удачных тестов не доказывают безопасность

В огромном корпусе сплошная ручная проверка может быть невозможна. Для независимой оценки применяются вероятностные выборки с определённой генеральной совокупностью и известной процедурой включения элементов. Необходимо отдельно учитывать редкие, но тяжёлые события: документы с данными ребёнка или непубличной медицинской историей могут составлять ничтожную долю общего корпуса и при этом определять риск обработки [1][2].

Оптимальная схема может сочетать случайную выборку по всем записям со стратифицированной проверкой высокорисковых доменов и целенаправленными негативными тестами. Эти три компонента служат разным целям. Результаты targeted stress tests нельзя статистически распространять на весь корпус; результаты случайной выборки нельзя трактовать как полное отсутствие редких нарушений.

При расчёте долей необходимо сообщать объём выборки и неопределённость. Например, если в случайной выборке из 200 документов не обнаружено прямых идентификаторов, это не доказывает нулевую вероятность их присутствия во всей совокупности: приближённое верхнее значение двустороннего 95-процентного интервала Уилсона составляет около 1,9%. Расчёт зависит от выбранной статистической модели и не заменяет юридического анализа тяжести последствий.

Отчёт должен отдельно описывать недоступные к проверке хранилища, исключённые страны, неопределённость classification model и возможный selection bias. Рекомендации к ремедиации привязываются к конкретным наблюдениям. Не следует рисовать «процент законности» как механическую сумму категорий compliance, которым законодательство не присваивает числовых весов.

Статистическая оценка качества фильтра начинается с определения генеральной совокупности. Если выбирались только записи из одного языка, с коротких страниц и без изображений, полученный результат нельзя переносить на весь международный корпус. Необходимо различать precision, recall и prevalence, поскольку низкая доля обнаруженных чувствительных сведений может отражать как реальную редкость, так и недостаточную чувствительность обнаружения. Для аудита требуется независимая контрольная разметка, а разногласия экспертов должны документироваться. Если записи стратифицированы по типу источника, языку и временной давности, показатели общей совокупности вычисляют с учётом соответствующих весов; простое среднее по намеренно равным тестовым группам может искажать итог.

При нуле обнаруженных ошибок в выборке нельзя утверждать абсолютное отсутствие ошибок. Даже для простой независимой случайной выборки существует верхняя доверительная граница частоты, зависящая от размера и принятой вероятности. При кластерной зависимости, повторяющихся публикациях и целевой выборке стандартная биномиальная модель может быть неприменима без корректировок. Исследователь должен объяснить дизайн до просмотра результатов, не корректировать критерии успеха задним

числом и отдельно публиковать редкие классы, где данных недостаточно для надёжного вывода. Кроме того, результат о точности фильтра не является автоматическим правовым разрешением собирать рискованные категории: качество контроля и наличие основания обработки остаются разными вопросами.

36. Метрики эффективности фильтров и обработка ошибок классификации

Технический фильтр может обнаруживать имена, контакты, числа документов, медицинские термины и признаки несовершеннолетних. Его качество нельзя описать только долей совпадений. Для двухклассовой проверки нужны чувствительность (recall), точность положительных предсказаний (precision), число ложных отрицаний, частота пропусков по группам источников и описание ground truth разметки. При изменении языков и жанров показатели могут резко измениться.

Ложный отрицательный результат может сохранить чувствительные данные в corpus, ложный положительный — необоснованно исключить допустимые материалы. Предпочтительный баланс порогов зависит от тяжести потенциального вреда и цели, однако сам по себе threshold не доказывает законность. Решения об исключениях должны пересматриваться человеком с учётом рисков и при необходимости с минимальным доступом к самому содержимому.

Для аудита полезны три набора: frozen контрольный корпус с документированными разрешениями, adversarial набор с редкими и необычными форматами идентификаторов и репрезентативная выборка обычных страниц. Желательно фиксировать версию алгоритма и дату теста, иначе результаты нельзя сравнивать между релизами. Не допускается создание опасных тестовых материалов, содержащих реальные лишние персональные сведения без надлежащего основания.

Дополнительная проверка касается обхода исключающих правил после преобразования документа, например после OCR, перевода или сегментации. Сохранение исходного запрета в metadata позволяет не терять ограничение при технической переработке. Метрика «исключение успешно применено» должна проверять downstream путь до тренировочной версии, а не только первый фильтр.

37. Санкционный комплаенс: риск вторичного распространения неверной идентичности

Условный сценарий: в открытой статье человек с распространённой фамилией упомянут рядом с санкционным субъектом. Автоматический сбор сохраняет абзац, а при повторном использовании модель начинает ассоциировать законопослушного человека с ограничительными мерами. В этой ситуации предметом исследования являются источник исходной ошибки, способ её воспроизведения, юридический режим информации и меры восстановления точности. Никакие реальные лица в данном примере не описываются.

Санкционный список должен проверяться по официальному акту и компетентному источнику, с учётом идентификаторов, дат включения и исключения, возможных однофамильцев и изменения статуса. Генеративная выдача, граф связи или совпадение имени не являются самостоятельным доказательством нахождения под санкциями.

Использование данных о конкретном лице при финансовом решении требует отдельной оценки применимых правил защиты данных и процессуальных гарантий.

Независимый аудит связывает «подозрительный вывод» с provenance конкретных фрагментов, если доступен такой след, и проверяет, обеспечена ли возможность исправления. Отсутствие доступа к весам модели ограничивает вывод: можно установить демонстрируемое поведение при выбранных запросах, но не обязательно доказать причинный вклад одного конкретного документа.

Практическая рекомендация — запретить использовать модель как первичный источник юридического санкционного статуса, фиксировать время проверки официального списка и создавать процедуру приоритетной эскалации жалоб о репутационном вреде. Такие меры касаются как защиты человека, так и качества решений финансовых организаций.

Для санкционного скрининга источник данных критически важен не только из-за возможных персональных идентификаторов, но и из-за юридической неоднородности событий. Коммерческие базы могут объединять официальную санкционную запись, упоминание в новостной статье, судебный спор, связь с организацией и совпадение имени. Генеративная модель при обучении на таком материале способна превратить неоднозначные признаки в гладкое, но юридически ложное утверждение о статусе человека. Поэтому система должна учитывать дату, орган принятия решения, идентификаторы лица, действительность ограничения, отмену записи и уровень подтверждения источника. Особенно опасно обучение на агрегаторах, которые не сохраняют историю исправлений и смешивают однофамильцев.

Условный кейс: модель сообщает, что предприниматель включён в санкционный список ЕС, основываясь на старой новостной статье о другом лице. Независимая проверка не должна ограничиться демонстрацией неверного ответа. Необходимо отдельно установить, есть ли соответствующая запись в официальном списке на конкретную дату, возможно ли установить первоисточник ошибки, повторяется ли ошибка в различных версиях модели и предусмотрен ли механизм исправления. Если внутренний тренировочный набор недоступен, причинный вывод о происхождении ошибки остаётся гипотезой. Правомерность дальнейшего решения банка оценивается по отраслевым требованиям отдельно от защиты данных. Это разграничение предупреждает как необоснованные обвинения разработчика, так и опасную практику принятия санкционного решения только на основании AI-генерации.

38. Международный розыск, экстрадиция и данные уголовного производства

Условный сценарий: удалённая из публичного доступа публикация о национальном розыске остаётся в учебном корпусе, а модель воспроизводит её без указания даты или последующего отказа иностранного суда в выдаче. Опасность состоит в смешении разных процедур: национальная запись, возможное международное уведомление, решение об экстрадиции и статус в информационных системах INTERPOL нельзя приравнивать друг к другу.

При подготовке правового заключения необходимо установить первичный процессуальный документ, дату, субъект, юрисдикцию, актуальный статус и основания использования персональных данных. Article 10 GDPR, если применим, требует отдельной проверки правового режима. Подлинность международного розыска не подтверждается сгенерированным ответом модели, поисковой выдачей или старой газетной публикацией [1].

Для независимой оценки training corpus важно определить, не превращаются ли опровергнутые обвинения в длительно сохраняющийся модельный признак. При невозможности технически устранить весь остаточный эффект разработчик должен оценить ограничения использования модели в правоприменительных или миграционных сценариях, особенно где вывод может повлечь лишение свободы или отказ в правах.

Рекомендуется протокол ускоренного исправления: сохранить доказательство выдачи, зафиксировать первичный источник и возможные судебные акты, квалифицировать чувствительность, ограничить рискованную downstream-выдачу и дать лицу понятный маршрут обжалования. Observatoire ARGA может разработать методику или провести независимую проверку в рамках доступных полномочий, но не заменяет официальный механизм пересмотра решений компетентных органов.

Информация о международном розыске особенно чувствительна к изменению правового статуса. Удаление сведений международной организацией, решение национального суда, отказ в выдаче, прекращение уголовного производства и журналистское сообщение об обвинении не являются взаимозаменяемыми документами. В публичном обучающем массиве они могут присутствовать в разной хронологии, а модель склонна суммировать наиболее повторяемую версию. Поэтому для независимой проверки сначала строится документальная временная линия, отделяющая установленные решения от предположений, а затем оценивается качество конкретного AI-ответа на ограниченном наборе тестов. Доступ к закрытым базам международного полицейского сотрудничества нельзя подразумевать по результату открытого поиска.

Этическая и юридическая дилемма возникает при публикации самой проверки. Чтобы показать, что модель неправомерно связала человека с преступлением, аудитор может вновь распространить исходное обвинение. Целесообразны ограничение идентифицирующих деталей, конфиденциальная полная версия для компетентного адресата и публичный статистический или обезличенный вывод, если цели можно достичь без повторного вреда. В частности, следует отказаться от механического включения в приложения полных скриншотов поисковой выдачи и паспортных данных. Это не универсальный запрет на использование доказательств, а требование соразмерной, защищённой процедуры с учётом статей 5, 6, 10 и 32 GDPR и применимого национального права.

39. Поставщики и разработчики: ответственность за решение при неполной информации

Разработчик, закупающий набор, не всегда способен провести самостоятельную сплошную проверку миллионов исходных страниц. Это объективное ограничение не делает достаточным единственное заявление поставщика «мы всё соблюдаем». Пропорциональная должная осмотрительность означает риск-ориентированные вопросы, договорные гарантии, выборочную верификацию, правила отказа от источника и возможность приостановки дальнейшего использования.

Необходимо разделять три ошибки: известное нарушение, неустранённая высокая вероятность нарушения и отсутствие достаточных доказательств для вывода. В отчёте для руководства они должны приводить к разным управленческим решениям. Назвать любую недостачу документа «установленным нарушением» некорректно; объявить её «соответствием» из-за отсутствия доказанного штрафа столь же необоснованно.

В договоре должны быть указаны юридически и технически исполнимые механизмы: уведомление о возникшей претензии, своевременная идентификация затронутых версий, порядок прекращения использования, сотрудничество в доказательной проверке, распределение информации о субъектах, защита конфиденциальных материалов и возможность независимого доступа в пределах закона.

С точки зрения корпоративного управления решение о принятии остаточного риска должен принимать уполномоченный орган после получения юридической и технической информации. Нельзя перекладывать такое решение на младшего инженера только потому, что он запускает scraper, либо на DPO как единоличного владельца бизнес-риска. Роли, сроки пересмотра и блокирующие условия фиксируются в управленческой матрице.

40. Предложения Observatoire ARGA к практике EDPB и международному диалогу

Первое предложение — установить проверяемый минимальный стандарт provenance, не предполагающий публикацию неограниченного списка персональных URL. Такой стандарт должен охватывать версии источников, цели сбора, категориальную фильтрацию, историю правовых оснований, сроки, связанные версии моделей и исполнение запросов. Способ раскрытия зависит от адресата: субъект, надзорный орган, независимый аудитор, публичная отчётность.

Второе предложение — подчеркнуть двустороннюю структуру эффективности opt-out: предотвращение нового сбора и оценка последствий уже завершённого обучения. Формальное исключение домена из crawler не должно автоматически описываться как стирание информации из весов модели. Поскольку возможности machine unlearning различаются, юридический стандарт должен требовать честного раскрытия доступных средств и доказательств их эффекта, а не технически необоснованного обещания.

Третье предложение — разработать специальные предохранители для данных детей, здоровья, уголовных правонарушений и пограничных/миграционных историй. Рекомендуется risk-first определение запретных категорий источников, карантин спорных фрагментов, ограничение downstream-использования в решениях с высокой тяжестью последствий и проверяемая эскалация жалоб. Это предложение институционального характера; оно не расширяет компетенцию EDPB или внешней организации.

Четвёртое предложение — публиковать воспроизводимые показатели аудита: величину обследованной совокупности, качество выборки, долю неизвестных источников, тесты чувствительности фильтра, фактические сроки разрешения жалоб и перечень ограничений независимого исследователя. Хорошая методика предпочтительнее «комплаенс-рейтинга» с придуманными весами.

Пятое предложение — укреплять процессуальные гарантии: мотивированный ответ по обращению, механизм представления дополнительных доказательств, защита частных сведений заявителя, отделение аудита от коммерческих консультаций и честное указание пределов заключения. Обсуждаемый текст Guidelines 03/2026 служит отправной точкой, а не уже действующим новым законодательством. Предложения следует рассматривать в рамках компетенции регуляторов и применимого права [2][13].

Предложения к EDPB следует формулировать как конкретные уточнения проекта Guidelines 03/2026, а не как просьбу «усилить защиту» вообще. Первое направление — проверяемые критерии достаточной прозрачности по статье 14 при сборе множества внешних источников, включая язык уведомления, описания категории источников и канал обращения. Второе — ориентиры по подтверждению мер минимизации для untargeted scraping, где быстро меняется состав посещённых сайтов. Третье — чёткие указания, какие доказательства практической эффективности фильтров могут учитываться в LIA, если известны значимые ложноотрицательные результаты на редких языках и специальных категориях. При этом предложения не должны противоречить законодательному тексту либо превращать добровольные методические формы в обязательные требования без надлежащего правового основания.

Четвёртое направление — совместимость прав субъектов с версионным устройством AI-моделей. Руководящие документы могли бы поощрять прозрачное разделение операций удаления из raw data, training corpora, индексов, checkpoint и вывода модели. Пятое — требования к проверяемости цепочки поставщиков без раскрытия персональных данных аудитору сверх необходимого. Шестое — регулярный пересмотр исключённых классов источников с учётом уязвимых групп, правомерных возражений сайтов и изменяющихся технических методов. Эти предложения направлены на снижение практической неопределённости; их допустимость и возможное включение в окончательный текст EDPB ещё не определены. Доклад фиксирует их как авторскую методологическую позицию Observatoire ARGA.

41. Выводы, ограничения и программа последующих исследований

Правомерность web scraping для генеративного ИИ не определяется одним фактом доступности страницы. Она требует отдельной оценки целей, основания по GDPR, категорий данных, ожиданий людей, альтернатив, защиты прав и фактического движения информации по всей технической цепочке. Публичность и инновационный характер проекта могут иметь значение в анализе интересов, но не заменяют обязательных правовых предпосылок [1][2][3].

На 22 октября 2026 года Guidelines 03/2026 EDPB являются проектом, представленным на публичное обсуждение до 30 октября 2026 года. Поэтому доклад не приписывает проекту силу вступившего в действие самостоятельного регламента и не предполагает неизменности его конечной редакции. Действующий GDPR и относимые нормы AI Act и авторского права оцениваются отдельно [2][11][18].

Предложенная методология независимой проверки требует доказательной последовательности от конкретного источника до изменения модели и исполнения прав субъекта. Из-за отсутствия доступа к внутренним корпусам большинства разработчиков такой подход не гарантирует полного восстановления training set во внешнем исследовании. Именно наличие документированных границ выводов отличает качественную экспертную оценку от недоказуемого публичного обвинения.

Дальнейшая исследовательская программа включает эмпирическую оценку качества фильтров на нескольких языках, сравнительный анализ применения статьи 14(5)(b) в государствах ЕС, методики подтверждения невозможности обратной идентификации и судебную практику по последствиям использования неправомерно собранных данных при

разработке моделей. Каждое направление требует самостоятельных проверенных источников, а не механического расширения этого документа.

Приложение 1. Терминологическая карта для независимого заключения

В рабочем заключении рекомендуется давать определения до оценки законности: crawler — программное обнаружение страниц; scraping — автоматизированное извлечение и сохранение; training set — конечная коллекция учебных записей; fine-tuning set — данные последующей настройки; provenance — подтверждаемая история происхождения; data broker — самостоятельный поставщик набора; rights reservation — указание правообладателя об ограничении использования для text and data mining в применимых условиях. Эти слова описывают разные юридические и технические отношения.

Карточка понятия включает английский термин, русскую форму, стадийность, пример доказательства и возможный риск ошибочного смешения. Так, public availability означает условия видимости информации, а legal basis — юридическое основание обработки персональных данных. Публикация на веб-сайте сама по себе не подтверждает наличие основания статьи 6. «Анонимизация» и «псевдонимизация» также не взаимозаменяемы: вторая обычно сохраняет возможность реидентификации при наличии дополнительной информации.

При исследовании отчётов внешних компаний следует включать таблицу использованных терминов и сведения об определениях, взятых из закона, регуляторной рекомендации или внутреннего стандарта. Термин, созданный авторами экспертной методики, помечается как авторский, чтобы не приписывать ему официальную юридическую силу.

Термин	Что надо подтвердить	Типичная ошибка
Публичный источник	Фактический доступ и ограничения	Считать свободным от GDPR
Законный интерес	Три условия статьи 6(1)(f)	Приравнять к коммерческой выгоде
Анонимный набор	Риски выделения, связи и вывода	Приравнять к хэшу имени
Удаление	Конкретные слои данных и проверка	Обещать стирание влияния без теста
Независимый аудит	Область, метод и ограничения	Выдавать за решение регулятора

Приложение 2. Матрица юридических оснований и дополнительных ограничений

Эта форма применяется к каждому отдельному этапу обработки, а не к «проекту ИИ» как неразделимому целому. В первой графе указывается операция, например извлечение, хранение, лицензирование набора, обучение или публикация вывода. Во второй — контролёр или совместные контролёры, определяющие цели и средства. В третьей — статья 6 GDPR с обоснованием её применимости. В четвёртой — отдельные ограничения статьи 9/10, защиты детей, передачи в третью страну и авторских прав.

Если исходные данные состоят из различных классов, карта составляется по рисковым группам. Положительная оценка для технической документации без персональных сведений не распространяется на сообщения о здоровье или индивидуальные обвинения. При

неизвестной категории источника запись переводится в статус «дополнительная проверка»; отсутствие доказательств не считается доказательством соблюдения.

Операция	Основание и дополнительное правило	Решение/доказательство
Сбор общедоступных контактов	Проверка статьи 6; баланса ожиданий	LIA + запреты категорий
Сведения о здоровье	Статья 6 плюс исключение статьи 9(2)	Legal review; при отсутствии — блок
Судимость/правонарушение	Статья 10 и применимое право	Правовая санкция или блок
Скачивание произведений	Применимость DSM/TDM и резервирования	Реестр rights reservation
Передача вне ЕЭЗ	Глава V GDPR дополнительно	Transfer assessment

Приложение 3. Форма LIA: конкретизация законного интереса

Форма открывается до сбора и получает уникальный номер и версию. Поле «интерес» должно содержать конкретную законную цель, измеримый желаемый результат, предполагаемых пользователей и предполагаемое воздействие на лиц. Недопустимый пример: «собрать весь интернет ради прогресса ИИ». Рабочий пример: «создать многоязычный словарь юридической терминологии для публичного информационного сервиса без построения индивидуальных профилей».

Затем указываются категории исходных сайтов, период сбора, географический охват, ожидаемый тип персональных сведений, необходимость каждой категории и возможные получатели. Поле «недостающие факты» обязательно: если источниковая структура неизвестна, нельзя оформлять решение как окончательно положительное. Следующий блок фиксирует связь данного интереса с заявленной целью по статье 5(1)(b) GDPR и отличает первоначальную публикацию от будущего использования [1][5].

Обязательное поле	Ответственный	Проверка
Конкретный законный интерес	Владелец проекта	Проверяемая цель, не лозунг
Категории лиц и материалов	Data lead	Фактический охват
Масштаб и сроки	Engineering	Реальные лимиты crawler
Альтернативы	Product + privacy	Документированы и сравнены
Решение по допуску	Уполномоченное лицо	Дата, версия, причины

При практическом заполнении LIA важно не оставлять описательные поля в форме деклараций. Для каждого класса источников оператор должен указать конкретную пользу предполагаемой обработки, альтернативные методы, аргументы против них и возможные негативные последствия для людей. Допустим, желаемая цель — улучшить качество распознавания редкого юридического языка. В записи можно указать недостаточность известных лицензионных терминологических корпусов, но вывод следует сопроводить сравнительными тестами или убедительным планом их получения. Ссылка на отсутствие готовых данных не доказывает, что разрешён массовый сбор публичных обсуждений с биографическими сведениями участников.

Независимая проверка LIA должна включать отрицательные решения. Если хотя бы одна категория источников не была отклонена, полезно проверить, рассматривались ли вообще аргументы противоположной стороны. Форма предлагаемой карточки содержит поле «контрдоказательство»: какой факт мог бы изменить принятое решение. Например, выявление высокой доли архивных сведений об уголовных преследованиях потребовало бы пересмотра допусков до запуска. Дата LIA, ответственное лицо и версия обоснования связываются с фактической конфигурацией crawler, чтобы при проверке можно было установить, действовали ли ограничения именно в момент извлечения данных. Это позволяет проследить организационное решение и не выдаёт внутреннее одобрение за признание законности надзорным органом.

Приложение 4. Форма LIA: необходимость и менее вмешивающиеся альтернативы

Сравнение проводится до выбора массового scraping. Включаются: лицензированный набор без персональных данных; агрегированный статистический корпус; синтетические примеры; ограниченный сбор по доменам; исключение личных профилей; минимизация метаданных; сокращение глубины архива. Для каждого варианта оценивают достижимость конкретной задачи, предполагаемое качество результата, влияние на права и достоверность исходных оценок.

Обоснование не может состоять только из фразы «альтернативы невозможны технически». Например, выборка из ограниченного класса отраслевых текстов может удовлетворять задаче специализированного ассистента и снижать риск извлечения чужих биографий. Если разработчик отвергает её из-за ухудшения качества, нужно указать контрольный эксперимент и величину изменения, а также размер и характеристики выборки.

Альтернатива	Проверка	Вывод
Лицензированные обезличенные тексты	Сравнение качества на hold-out наборе	Приемлема / недостаточна и почему
Сбор только тематических доменов	Потеря покрытия целевого языка	Контролируемое ограничение
Синтетические примеры	Leakage, качество, репрезентативность	Применение с защитными условиями
Предварительное исключение личных профилей	Доля пропусков при тесте	Утвердить / переработать фильтр

Приложение 5. Баланс интересов: модель причиняемого и предотвращаемого вреда

Матрица не вводит искусственный юридический балл. Она предназначена для структурированной аргументации. Описываются риски каждого вида: утрата контроля над публикацией, нежелательная выдача информации из модели, повторная идентификация, стигматизация, ошибочная криминальная или санкционная характеристика, непрозрачность и препятствия осуществлению права на возражение. Для каждого приводятся типичные затронутые лица, вероятные механизмы ущерба, смягчающие меры и остаточная неопределённость.

Отдельный пункт — reasonable expectations субъекта. Индивидуальный профессиональный профиль и личное сообщение на форуме могут быть одинаково индексируемыми, но ожидания дальнейшего использования отличаются. В мотивировке приводится не абстрактное мнение юриста, а факты контекста, масштаба и дальнейших целей. Приоритет защиты интересов ребёнка должен учитываться особо [1][2].

Заключение указывает, какие меры подтверждены тестом, какие только запланированы, и кто принимает окончательное решение. Если выраженное несоответствие нельзя устранить, форма предусматривает блокировку работы с затронутой категорией. Недостаток документальных доказательств фиксируется как нерешённый вопрос, а не маскируется словом «низкий риск».

Приложение 6. Допуск и запрет интернет-источников

Список доменов не должен сводиться к техническому allowlist. Каждому источнику присваиваются признаки: категория сайта, авторизация доступа, публичность, специальные типы персональных данных, наличие данных несовершеннолетних, robots.txt/CAPTCHA и других выраженных ограничений, известные договорные условия, период публикации и возможные права на контент. Источник может быть разрешён для технического мониторинга метаданных, но запрещён для сохранения полного текста.

Предлагаются четыре статуса: одобрен с условиями; временно ограничен; заблокирован; ожидает проверки. Статус имеет владельца решения и срок повторной оценки. Изменение архитектуры сайта, доступности или условий не наследует автоматически прежний допуск. При массовом сборе необходимо проверять, действительно ли crawler использует последнюю утверждённую версию источникового списка.

Признак	Пример управленческого решения
Форум с вероятными личными историями болезни	Предварительная блокировка; legal review
Официальный открытый набор технической терминологии	Допуск после проверки лицензии и данных
Ресурс с техническим возражением против сбора	Исключение из автоматического обхода
Домен с неизвестным статусом и детскими данными	Карантин до мотивированного решения
Старый судебный архив с данными о преступлениях	Отдельная проверка ст. 10 и права доступа

Приложение 7. Source Intake Sheet: карточка первичного источника

Карточка фиксирует дату принятия решения, исходный URL или ограниченный идентификатор, владельца ресурса, условия доступа, язык, юрисдикцию, предмет, сведения об источнике лицензии, класс персональных данных и обоснованность необходимости. При обработке особо чувствительных материалов прямой URL допускается хранить в закрытом реестре, а в публичной версии указывать безопасный псевдоним источника.

Пример условной записи: SOURCE-LEGAL-2026-001, «официальные публичные документы с юридической терминологией», доступ без персональной регистрации, robots.txt разрешает

только часть разделов, исключены страницы с персональными жалобами и поисковая выдача по физическим лицам. Решение: «одобрено для ограниченного корпуса технических текстов», срок пересмотра через три месяца, владелец — руководитель data governance. Запись не имитирует сведения реального сайта.

Поле карточки	Требуемая запись
Source ID и версия	Стабильный идентификатор
Дата и маршрут доступа	Фактическая проверка, не предположение
Правовой статус	GDPR, контент, специальные ограничения
Исключённые разделы	Конкретные шаблоны и пути
Уровень доказательности	Первичная проверка / тест / внешний аудит

На практике карточка первичного источника начинается с точной фиксации домена и вида материала, но этого недостаточно для установления происхождения конкретной записи. Следует различать сайт-хост, первоначального автора, агрегатор и последующего перепубликатора. Например, журналистская публикация может цитировать судебный акт, а другой портал может воспроизвести её с сокращениями. Если модель обучается на последнем варианте, сообщение об ошибке должно иметь путь к первоначальному документу, однако отсутствие такого пути не следует восполнять догадкой по текстовому сходству. В карточке фиксируются URL извлечения, дата, категория, известные исходные источники, степень подтверждения и ограничения на распространение копии.

Для контроля массовых коллекций полезна двухступенчатая проверка: автоматизированная первичная категоризация доменов и выборочная человеческая рецензия категорий, где ошибка может быть дорогостоящей. Важно сохранить не только одобренные сайты, но и причины отказа. Если сайт был исключён из-за явного возражения против scraping, системе требуется механизм обновлять запрет при изменении URL или зеркал. Однако технический сигнал robots.txt не заменяет полноценного правового анализа; его отсутствие не означает разрешение всех видов обработки. При противоречиях карточку переводят в статус «приостановлено» до выяснения обстоятельств.

Приложение 8. Dataset Manifest: минимальная структура доказательства происхождения

Унифицированный manifest включает Dataset ID, version, created_at, purpose ID, legal basis reference, source category, crawl job ID, source window, content type, PII-filter version, retention rule, transfer policy, downstream model runs и статус заявлений субъектов. Изменения версий должны быть append-only в журнале, а не переписываться без следа. Должен существовать способ доказать, какой файл использовался в конкретном запуске обучения.

Пример записи, предназначенный только для иллюстрации: DATASET-DEMO-07; версия 2.1; технические публикации за определённый период; общий корпус 18 000 условных документов; исключены личные профили, спорные сведения в карантине; контрольный журнал содержит 12 отклонённых URL; дальнейшее обучение допускается только после обновления фильтра. Все цифры здесь учебные, не описывают операционную статистику какой-либо компании.

Контроль полноты manifest не ограничивается заполняемостью. Аудитор сравнивает выборку строк с логами extraction, каталогами raw/clean, хэшами и записями использования в model registry. В случаях отсутствия связи вместо восстановления «по памяти» присваивается статус traceability gap с оценкой влияния на выводы.

Идентификатор	Назначение	Обязательное доказательство
Crawl Job ID	Процесс извлечения	Конфигурация, логи, дата
Dataset Version	Состояние набора	Хэш списка объектов
Model Run ID	Использование в обучении	Манифест загрузки и обучения
Rights Ticket ID	Исполнение запроса	Решение, действие, результат
Exception ID	Ограничение в источнике	Причина, владелец, срок

Приложение 9. Проверка поставщика уже собранного корпуса

Покупателю рекомендуется запросить цепочку происхождения по категориям и представительную выборку данных, не получая лишние персональные сведения. Запрашиваются география и периоды сбора, допустимые цели повторного использования, legal basis, статья 9/10, правовой статус контента, история возражений, удаление данных по жалобам, список субподрядчиков и содержание security controls. Отдельно проверяется, способен ли поставщик определить объём ранее полученных и уже удалённых данных.

Внутренняя шкала проверки не должна позволять компенсировать отсутствие правового основания большим количеством технических сертификатов. Юридически блокирующие недостатки отмечаются отдельно от обычных эксплуатационных несоответствий. Если поставщик не может предоставить сведения из-за коммерческой тайны, стороны могут организовать контролируемый просмотр доказательств независимым экспертом; это не должно вести к раскрытию данных без правового основания.

Вопрос поставщику	Доказательство	Статус ответа
Кто являлся контролёром при сборе?	Карта ролей и договоры	Подтверждён/оспаривается/неизвестно
Как исключаются запрещённые сайты?	Правила crawler и журнал тестов	Подтверждён/неподтверждён
Как исполняются возражения после продажи?	Процедура уведомления покупателей	Реализовано/не реализовано
Есть ли данные статьи 9 или 10?	Протокол классификации	Наличие/отсутствие/неизвестно
Можно ли проверить конкретную версию?	Manifest, хэши, dataset lineage	Полный/частичный след

Поставщику датасета стоит адресовать не общий вопрос «соответствует ли набор GDPR», а серию проверяемых запросов. Покупатель запрашивает категории сайтов, временные периоды, происхождение заявленных лицензий, основания дальнейшей обработки и описание процедур, которыми исключены специальные категории, детские данные и сведения об уголовном преследовании. Для части вопросов может быть достаточно агрегированного ответа, если подробное раскрытие само угрожает правам третьих лиц. Но

любой ответ должен иметь дату и относимость к конкретной версии массива; старый audit certificate на другую коллекцию не подтверждает свойства новой поставки.

Если поставщик заявляет «все записи анонимны», разумно запросить модель угроз повторной идентификации и ограниченный тест связывания. Иногда идентификация возможна не из одного поля, а из редкого сочетания содержимого и внешних данных. При отказе в независимом тестировании нужно отдельно оценить риск, а не превращать коммерческий спор о конфиденциальности в презумпцию нарушения. Решение о закупке фиксирует три исхода: достаточные подтверждения, исправимые пробелы и существенная неопределённость. При последнем исходе методически оправдано отказаться от допуска определённых категорий до устранения препятствий.

Приложение 10. RACI и разграничение полномочий участников

Одна команда не должна одновременно определять экономическую цель, признавать правовое основание доказанным, проверять собственные фильтры и утверждать внешний отчёт. Для проекта web scraping существенны роли владельца обработки, data lead, DPO/privacy counsel, security, product, специалиста по правам субъектов и независимого аудитора. DPO консультирует и контролирует, но не автоматически становится владельцем коммерческого решения.

Модель RACI: руководитель проекта формулирует цели и отвечает за ресурсы; privacy counsel проверяет правовое основание и риски; engineering отвечает за лимиты и сохранность версий; security — за доступ и инциденты; rights team — за обработку обращений; внутренний комитет принимает решение о принятии остаточного риска. Независимый аудитор оценивает ограниченную совокупность доказательств, не подменяя управление организацией.

Действие	Responsible	Accountable	Проверяющий
Цель и scope сбора	Product	Уполномоченный руководитель	Privacy counsel
LIA и DPIA	Privacy + data lead	Контролёр	DPO (функция контроля)
Фильтры данных	Engineering	Data governance owner	Независимый тестировщик
Права субъектов	Rights team	Контролёр	Privacy counsel
Публикация оценки	Observatory audit lead	Редакционный орган	Независимая рецензия

Приложение 11. Публичное информационное уведомление по статье 14 GDPR: проект структуры

Организация, которая получает персональные данные из общедоступных интернет-ресурсов для обучения генеративной модели, должна подготовить информацию понятным языком. Эта примерная структура не является универсальным официальным шаблоном: конкретные обязательные сведения определяются статьёй 14 GDPR, фактическим источником и характеристиками обработки. Уведомление включает наименование и контакты контролёра, цели и основания обработки, категории сведений, типы источников, получателей, сроки, возможные передачи, права и порядок обращения [1][2].

Пример содержания для адаптации: «Мы используем ограниченные категории общедоступных материалов для разработки системы обработки текстов. В некоторых случаях материалы могут содержать сведения, относящиеся к идентифицируемым лицам. Мы применяем правила отбора, исключения источников и очистки данных. Вы можете обратиться по указанному каналу, чтобы запросить информацию об обработке, исправление, удаление, ограничение или заявить возражение в предусмотренных законом случаях. Мы рассматриваем каждый запрос с учётом фактической обработки и применимого права».

Необходимо указать реальные категории доменов и периоды сбора, а не оставлять абстрактную фразу «все общедоступные источники». При применении исключения из индивидуального уведомления по статье 14(5)(b) публичное сообщение должно оставаться легко доступным и содержать предусмотренную законом информацию. Перед публикацией нужно проверить, не раскрывает ли перечень источников новые частные сведения. Наличие такого текста не освобождает от иных обязанностей.

Блок уведомления	Что проверяет рецензент
Контролёр и контакты	Реальная возможность обратиться
Цель и основание	Определённая, а не общая фраза
Источники	Категории, период, ограничения
Права	Рабочие процедуры и сроки ответа
Получатели и передачи	Соответствие реальной архитектуре

Приложение 12. Обоснование статьи 14(5)(b): документированная процедура

Контролёр описывает количество и категории потенциальных субъектов, временную глубину материала, наличие контактных данных, реалистичность прямого уведомления, предполагаемые расходы и дополнительные риски установления личности ради отправки сообщения. Затем сравнивает отсутствие индивидуального уведомления с возможными последствиями для прав, указывает альтернативные способы информирования и дополнительные меры безопасности [1][2][14].

Пример условной оценки: в библиотеке исторических форумных обсуждений за 15 лет возможно присутствуют упоминания нескольких сотен тысяч лиц без актуальных контактных данных. Создание перечня адресов может потребовать дополнительного массового извлечения, повышающего риск. Это обстоятельство не освобождает от общей защиты: требуется опубликованная политика, ограничение персональных сведений, доступный канал обращений и периодический пересмотр обоснования. Окончательный вывод зависит от конкретного набора.

Форма содержит подпись ответственного за цели и документированную правовую оценку, но юридическое основание определяется нормой, а не количеством подписей. Пересмотр проводится при возникновении новых средств связи, выявлении чувствительных материалов или изменении масштаба сборки. Решение с неизвестными параметрами маркируется как предварительное.

Форма для статьи 14(5)(b) должна отражать не только число потенциальных субъектов, но и затраты на индивидуальное извещение, доступность контактной информации, реальные препятствия установления лиц и тяжесть последствий для их прав. Простая оценка «данных слишком много» не раскрывает критерий несоразмерности. Полезен сценарный расчёт по отдельным категориям: например, для профессиональных регистраций адресов электронной почты может быть достаточно, тогда как для архивных постов без надёжных контактов ситуация отличается. Нельзя автоматически переносить вывод по одной категории на весь универсальный корпус. При этом сам сбор адресов исключительно ради уведомления способен создавать новый самостоятельный риск, который также оценивается.

Если исключение применимо, проект требует мер, защищающих законные интересы субъектов. Среди них может быть доступное публичное уведомление, описание источников, механизм обращений, эффективные средства предотвращения новой обработки после обоснованного возражения и отдельные правила чувствительных категорий. В акте должны быть представлены не только намерения, но и результат проверки: открывается ли уведомление без регистрации, понятен ли язык, поступают ли заявки, исполняются ли они в установленный срок. Исключение из индивидуального информирования не означает освобождения от остальных принципов GDPR и не аннулирует права субъекта.

Приложение 13. DPIA: карта угроз правам и свободам

Первый лист описывает операции, цели и необходимость. Второй — категории лиц, объём данных, правовой режим и сроки. Третий — угрозы и имеющиеся меры. Четвёртый — остаточный риск, обоснование решения и необходимость предварительной консультации по статье 36 GDPR в соответствующих обстоятельствах. Оценка должна осуществляться по фактическим признакам, а не по универсальной формуле «масштаб большой — нарушение доказано» [1][16].

В учебном примере генеративная система собирает научные публикации, но дополнительно сохраняет неанонимизированные обращения пациентов на медицинском форуме. Потенциальный вред включает раскрытие здоровья, ошибочные медицинские выводы и невозможность контроля дальнейшего распространения. Предлагаемые меры: запрет домена и аналогичных источников, уничтожение уже собранных копий, аудит восстановления сведений, перепроверка связанных training runs и информирование о доступных правах в установленном порядке.

Риск	Механизм	Предлагаемая защита
Ненужные сведения о здоровье	Автоматический обход форума	Доменные исключения; фильтры; карантин
Недостоверное обвинение	Старый новостной корпус	Временные атрибуты; проверка статуса
Повторная идентификация	Объединение таблиц	Ограничить linking; модель угроз
Невозможность возражения	Нет источниковой связи	Rights ticket + dataset lineage
Международный доступ	Удалённые подрядчики	Глава V и контроль доступа

Приложение 14. Стоп-правило для данных статьи 9 GDPR

Правило «остановить — изолировать — проверить — документировать» применяется при вероятном появлении специальных категорий. Сначала фиксируется категория и минимально необходимое доказательство обнаружения. Затем затронутый batch переводится в закрытое хранилище без допуска к обучению. Юрист проверяет, существует ли конкретное исключение статьи 9(2) при наличии законного основания статьи 6; при отсутствии надлежащей конструкции данные исключаются и удаляются согласно утверждённой процедуре.

Не рекомендуется показывать чувствительные документы широкому кругу сотрудников для «уточнения», поскольку сама проверка может усилить риск. Внутренний отчёт указывает долю срабатываний и ограничение теста. Если фильтр обнаружил лишь часть подобных сведений, невозможность оценить полноту должна сохраняться в выводах и планах дообследования.

Этап	Кто действует	Как фиксировать
Обнаружение вероятных данных статьи 9	Data governance	Risk event ID, тип признака
Карантин	Security / engineering	Ограничение доступа и журнал
Правовая квалификация	Privacy counsel	Ст. 6 и конкретная ст. 9(2)
Ремедиация	Data engineering	Удалённые версии, retraining risk
Повторная проверка	Независимый reviewer	Выборка и ограничения заключения

Приложение 15. Стоп-правило для сведений об уголовных правонарушениях

Обработка данных статьи 10 GDPR требует отдельного правового анализа и не должна автоматически попадать в общую процедуру для обычных публичных документов. Для подозрительной записи сохраняют только необходимые признаки: официальный или неофициальный источник, тип предполагаемого правового события, дату, статус, возможный идентификационный конфликт и юрисдикцию. Отдельно устанавливается, вправе ли организация обрабатывать такой класс данных под применимым правом [1].

Условная запись «персона связана с расследованием» не должна преобразовываться в метку «судимость» без компетентного первичного документа. Полицейские диффузии, розыск, запрос на выдачу и судебный приговор имеют различные правовые условия и доказательное значение. При недостатке подтверждений для использования данных в обучении разумно приостановить включение такого фрагмента.

Карта ремедиации содержит сравнение ошибочной и корректной версии, доказательства распространения, решение о запрете использования и способы ограничить модельные выводы о конкретном лице. Внешнее сообщение не должно публиковать дополнительные данные заявителя без самостоятельного обоснования.

Приложение 16. Защита детей: правила доменных и содержательных исключений

Команда определяет источники, где ожидаемы детские сведения, включая школьные страницы, молодёжные обсуждения, семейные фотогалереи и помощь несовершеннолетним. Действие по умолчанию в высокорисковых категориях — отказ от сбора до отдельной проверки, а не автоматическое включение с последующим анонимизированием. При работе с возрастными индикаторами нужно избегать необоснованного дополнительного сбора документов личности [1][2].

Протокол проверки включает набор безопасных синтетических примеров, оценку пропусков фильтра на разных языках, тест обхода через OCR и краткие мультимодальные подписи. В отчёте показываются ошибки, ограниченные набором тестов. Не следует считать, что большая языковая модель способна достоверно определить возраст лица по тексту каждого интернет-поста.

Показатель проверки	Вопрос
Предварительное исключение источников	Есть ли классы сайтов, не посещаемые crawler?
Ошибочные пропуски	Какие детские упоминания не замечены?
Последующее использование	Исключаются ли сведения из следующих training runs?
Права родителей/представителей	Есть ли действующий канал обращения?
Остаточная неопределённость	Документированы ли границы классификатора?

Приложение 17. Процедура обработки запроса субъекта данных

Каждому обращению присваивается номер, категория требований, дата получения, контакт для безопасного ответа, проверка личности только в необходимом объёме, связанный dataset ID и ответственная команда. Далее осуществляется поиск по допустимым идентификаторам, анализ правовой обязанности, определение мер и документирование ответа. В карточке автоматически фиксируются контрольная дата одного месяца по статье 12(3) GDPR, отметка о своевременном уведомлении заявителя в случае мотивированного продления и дата окончательного ответа. Продление возможно ещё максимум на два месяца при соблюдении установленных условий [1].

Интерфейс заявителя не должен требовать точного знания внутренней технической структуры. Запрос, содержащий имя и ссылку на проблемную публикацию, может служить отправной точкой поиска, но риск ошибочного совпадения требует подтверждения прежде, чем передавать сведения постороннему. Если контролёр не может безопасно идентифицировать человека, он объясняет препятствие и законные способы уточнения запроса.

Шаг	Документируемое действие
Приём	Номер, тип запроса, дата
Идентификация	Минимально достаточная проверка
Поиск	Источники и связанные версии

Шаг	Документируемое действие
Правовая оценка	Применимая статья и исключения
Ремедиация	Raw/clean/index/model, степень подтверждения
Ответ	Мотивировка, дата, способ оспорить

Рабочая процедура запроса строится вокруг журнала обязательств с ограничением доступа. После обращения регистрируются дата поступления, объект запроса, тип предполагаемого права, необходимая проверка личности и перечень затрагиваемых контролируемых систем. Далее определяется юридическая применимость соответствующего права: запросы об удалении, ограничении, исправлении и возражении не имеют полностью одинаковых условий. Не следует отказывать в целом только потому, что технически невозможно доказать удаление влияния из всех весов модели; при наличии контролируемых исходных или индексированных записей допустимые меры должны оцениваться отдельно.

При взаимодействии с брокерами важно предусмотреть маршрутизацию: кому передаётся запрос, кто принимает решение, какие подтверждения возвращаются и может ли покупатель заблокировать повторную загрузку из старой версии. Запросы от лиц с одинаковыми именами не должны приводить к раскрытию данных другого человека. Протокол ответа указывает только проверенные действия, оставшиеся ограничения, контакт для последующего оспаривания и информацию о возможных средствах защиты согласно применимому праву. Подобная процедура полезна и для оценки качества внутреннего комплаенса: она позволяет измерить время реакции без публикации чувствительных деталей жалоб.

Приложение 18. Карта исполнения удаления на разных уровнях модели

Техническое действие «удалить данные» расшифровывается по уровням: исходный scrape, очищенный документ, training dataset, временный cache, retrieval-индекс, оценочный набор, журналы и модельные веса. Для каждой ступени фиксируется факт наличия, возможность технического удаления, подтверждение исполнения и остающееся ограничение. Важно не заявлять успешное «забывание» лишь из-за удаления одного файла.

В учебном сценарии спорная запись обнаружена в индексе поиска, но происхождение model weights не позволяет исключить более раннее использование. Команда удаляет запись из индексного набора, запрещает её повторное извлечение, исследует связь с тренировочными версиями и отдельно оценивает методы снижения риска выдачи. Отчёт сообщает, что подтверждено, а что нельзя установить.

Слой данных	Потенциальное действие	Доказательство
Raw objects	Удалить или изолировать	Журнал и проверка резервных копий
Curated set	Исключить запись/версию	Dataset manifest до и после
Retrieval index	Удалить вектор и фрагмент	Тест отсутствия результата
Model weights	Оценить retraining/unlearning	Документированный эксперимент
Выход модели	Фильтр + контрольный тест	Поведение на оговорённых запросах

Техническая карта удаления связывает каждый запрос с разными хранилищами и моделями. Для raw data подтверждается удаление или ограничение конкретной записи; для подготовленного корпуса — исключение из версии и будущих сборок; для retrieval — переиндексация; для fine-tuning — пересмотр контрольной точки; для параметрической модели — тестирование извлечения и оценка доступных мер. При этом хранение доказательств исполнения само должно иметь основание, срок и соответствующие меры безопасности. Нельзя ради подтверждения удаления бессрочно сохранять полную копию спорной записи в открытом тикете.

Акт должен различать результат и прогноз. Запись «планируется исключить при очередном обучении» не равна подтверждённому удалению из модели, доступной сегодня. Напротив, неуспех конкретного теста извлечения данных не доказывает, что модель никогда не сможет сгенерировать их иным способом. Рекомендуются ограниченные формулы: «удалено из указанного индекса на дату», «исключено из набора версии X», «на протоколе тестов Y не выявлено воспроизведения» и «влияние на веса не установлено». Для каждой формулы сохраняются идентификаторы проверенной системы и условия испытания. Особенно важно не заявлять о действиях независимых третьих сторон, когда организация не контролирует их копии и не получила подтверждений.

Приложение 19. Протокол аудита фактической работы crawler

Аудитор запрашивает configuration snapshot, версии parser, правила обхода ссылок, используемый user-agent, schedule, лимиты скорости и список разрешённых и запрещённых доменов. Из сетевых логов берётся случайная выборка посещений, по которым устанавливают, совпадают ли реальные операции с утверждённым scope и как отрабатывают исключающие правила. Если полнота журналов недостаточна, это указывается как ограничение проверки.

Тест проводится в контролируемой среде либо на законно доступных страницах. Недопустимо ради аудита обходить CAPTCHA, аутентификацию или техническую защиту чужого ресурса без правомерного основания. Точность процедур может быть проверена на собственных имитационных сайтах, содержащих искусственно созданные данные и различные запреты.

Тест	Ожидаемый результат
robots.txt disallow для контрольного пути	Crawler не извлекает страницу
Страница вне allowlist	URL отклонён
Страница с запрещённым классом содержимого	Batch карантинизирован
Изменение правил в процессе	Применяется версия на дату задания
Сбой сети/логирования	Ошибка записана, не скрыта

Приложение 20. Модель защиты данных в цепочке подрядчиков

Пример архитектурной карты: владелец проекта утверждает цель и набор источников; подрядчик-скрейпер исполняет инструкции и ведёт extraction logs; облачный оператор

хранит ограниченную коллекцию; компания разметки получает отобранное подмножество; разработчик обучает модель; права субъектов обрабатывает назначенный контролёром контакт. Для каждой стороны устанавливаются фактические роли, доступы, регион, операции и срок хранения [1][4].

Аудитор рассматривает не только формальное наличие DPA, но и исполнение договорных условий: есть ли журнал субподрядчиков, регламент удаления копий после окончания договора, возможность вернуть/удалить объекты и помочь при запросе субъекта. Непрозрачный путь передачи особенно опасен при данных особых категорий или сведениях о судимостях.

Рекомендуется проводить тест «отзыв внешнего доступа»: прекратить доступ бывшего подрядчика, подтвердить удаление его рабочих копий и проверить, что обучение не зависит от непроверяемых сторонних архивов. Положительный результат одного теста не подтверждает отсутствие иных копий, поэтому ограничение теста указывается прямо.

Приложение 21. Авторское право и право на text and data mining: отдельная проверка

Юридическая карточка контента должна отличать существование охраняемого произведения, применимое исключение или лицензию для соответствующей операции, законность доступа и наличие надлежаще выраженного reservation of rights. По статье 4 Директивы 2019/790 исключение для TDM при законном доступе может быть обусловлено отсутствием надлежащего резервирования использования правообладателем. Конкретная имплементация и действие иных норм проверяются отдельно [18].

При аудите фиксируются тип ресурса, способ выражения ограничений, период актуальности, правило отбора и связь с training content summary GPAI. Юридическая оценка персональных данных ведётся параллельно и не заменяется выводом по TDM. В формулировке заключения целесообразно раздельно сообщать: «условия авторско-правового использования подтверждены / не подтверждены» и «основание персональной обработки подтверждено / не подтверждено».

Источник	Что документировать	Возможный вывод
Лицензированный корпус	Объём лицензии и downstream-права	Использование в пределах условий
Открытый сайт	Правовой доступ, TDM reservation	Требуется отдельная оценка
Чужая закрытая база	Договор, технические ограничения	Без допуска не собирать
Публикация с персональными сведениями	GDPR плюс права на контент	Два параллельных анализа

Приложение 22. Связка AI Act, GDPR и раскрытия содержания обучения

Матрица связывает публичную сводку по статье 53 AI Act с внутренними документами поставщика и защитой персональных данных. Общедоступная сводка по контенту не обязана раскрывать полные идентифицируемые документы, но должна быть достаточно информативна в соответствии с применимым шаблоном и правовыми требованиями.

Техническая документация в полном объёме имеет иной круг получателей, чем публикация для широкого круга [11][12].

В рабочем примере компания указывает в публичной сводке «публикации открытых научных сайтов, специализированные отраслевые статьи и лицензированные материалы». Внутри организации каждой категории соответствуют record ID, временной период, правила сбора, отсутствие или наличие личных данных, политика исключений и связь с моделями. При отсутствии внутренней связки публичная формула не является доказательством фактического provenance.

Публичный объект	Внутреннее доказательство
Категория тренировочного контента	Dataset ID и источник категории
Период сбора	Crawl logs и timestamp
Политика соблюдения авторских прав	Rights reservation register
Правомерность персональной обработки	LIA, DPIA и related records
Доступность прав субъекта	Tickets и документированные ответы

Приложение 23. Карта трансграничного доступа к персональным данным

Контролёр составляет карту физических мест хранения, облачных регионов, административного доступа, субобработчиков и маршрутов резервного копирования. Для каждого потенциального международного доступа определяется, существует ли юридически значимая передача, и если да, какой механизм главы V GDPR используется и какие меры необходимы. Режим международной передачи дополняет, а не заменяет общую проверку статьи 6 [1].

Практический аудит включает проверку служебных учётных записей, внешних обработчиков аннотаций, диагностических логов и команд поддержки, имеющих удалённый доступ. Отдельный вопрос — могут ли персональные данные из prompts пользователя оказаться в системе мониторинга, не описанной как часть исходного training pipeline.

Элемент	Контрольный вопрос
Облачный регион	Где расположена фактическая копия?
Администрирование	Кто и из какой страны имеет доступ?
Субобработка	Есть ли разрешение и журнал изменений?
Правовой механизм передачи	Установлен ли и применим ли?
Передача спорных данных	Возможны ли временная блокировка и исправление?

Приложение 24. План статистической проверки источниковых запретов

Генеральную совокупность определяют как все элементы конкретной версии корпуса, к которым применяются правила сбора. Формируют случайную выборку с фиксированным

seed и отдельно целевую высокорисковую выборку. Для первой допустима оценка доли определённых ошибок с интервалом неопределённости; для второй вывод ограничен проверенными классами, и её нельзя без весов экстраполировать на всю совокупность.

Учебный пример: из 200 случайно отобранных документов два относятся к категориям, требующим ручной проверки. Наблюдаемая доля равна 1%. Двусторонний 95-процентный интервал Уилсона составляет приблизительно 0,27–3,57%, если применима соответствующая модель выборки. Эти цифры не являются оценкой реального корпуса и не означают, что 99% документов законны: классификатор выявляет только выбранный признак, а юридическая правомерность включает другие требования.

Переменная	Значение учебного примера
Размер случайной выборки	200
Найденные спорные категории	2
Наблюдаемая доля	1%
Интервал Уилсона (95%)	≈0,27–3,57%
Запрещённый вывод	«99% корпуса законны»

Статистический план должен предусматривать вероятность отбора каждой записи или честно объяснять, что выборка является целевой. Условно случайные «пятьдесят красивых примеров» пригодны для иллюстрации проблемы, но не для надёжного вывода о доле нарушений в многомиллионном наборе. При высоком риске слабопредставленных языков целесообразна стратификация: крупные языковые группы оцениваются отдельно от редких, а чувствительные типы сайтов имеют собственные контрольные наборы. Если результат впоследствии агрегируется, вес каждой группы должен учитывать дизайн исследования. Незнание долей групп ограничивает возможность вычисления общей частоты.

В протоколе фиксируются размер генеральной совокупности, единица отбора, способ случайного выбора, критерий ошибки, инструкция аннотаторам, процедура разрешения разногласий и план вычисления неопределённости. Важно измерять как ошибочное исключение безопасных данных, так и сохранение запрещённого контента. Для правового анализа последняя ошибка часто имеет более серьёзные последствия, но нельзя механически объявить любую обнаруженную ошибку незаконной обработкой: требуется знать фактическое основание, контекст и принятые меры. Аудитор должен разделять статистический результат, техническое объяснение и юридический вывод.

Приложение 25. Тестовая матрица обнаружения персональных данных

Оценка фильтра проводится на нескольких языках и типах источников: обычный текст, HTML с нестандартной разметкой, переведённые фрагменты, OCR, перепосты и структурированные формы. Обучающие и контрольные примеры должны быть разделены, а для внешнего аудита предпочтительны искусственные или специально разрешённые тестовые записи. Метрики вычисляются по известной разметке, с учётом баланса классов и доли случаев неопределённой квалификации.

Рабочая карточка включает версию фильтра, набор данных, объём теста, истинно положительные, ложно положительные, ложно отрицательные результаты, precision, recall и ограничения. Если в 100 подтверждённых чувствительных образцах обнаружены 92, recall равен 92%. Но этот результат сам по себе не показывает precision, пока не известны ложные положительные срабатывания среди прочих документов. Аналогично успешность на английском не доказывает успешность на французском или русском.

Тестовый класс	Основной риск	Контроль
Юридические имена и однофамильцы	Ошибочная атрибуция	Контекст и официальный статус
Свободный медицинский текст	Ложные отрицания	Семантическая проверка
Переведённые страницы	Потеря чувствительности	Многоязычный hold-out
OCR документов	Скрытые номера и даты	Тест ошибки распознавания
Спорные данные детей	Возрастная неоднозначность	Блокировка источников и review

Приложение 26. Программа независимого аудита на 20 рабочих процедур

Аудит начинается с согласования scope: какая версия модели, какой corpus и какие страны охвачены. Затем согласуют правовой доступ к материалам и ограничения конфиденциальности. Отдельными тестами проверяют: цель обработки; LIA; DPIA; категорию данных; список доменов; запреты сбора; источник и время; versioned manifests; контракты с брокерами; контролёрские роли; права субъектов; хранение и удаление; международный доступ; конфигурацию crawler; точность фильтров; downstream traceability; memorisation risk; авторско-правовые оговорки; публичную сводку GPAI; независимость подписанта отчёта.

Для каждой процедуры фиксируют запрос документов, полученные доказательства, выполненный тест, результат и неопределённость. Недоступность журнала не должна приводить к заполнению результата на основании интервью. Отчёт различает проверенное отсутствие явного нарушения на данной выборке и доказательство соответствия всех действий нормативному режиму. В случае правового спора используются осторожные формулы и возможность представить позицию проверяемого участника.

Процедуры	Тип доказательства
1–4: цель, LIA, DPIA, категории	Решения, оценки, карты данных
5–9: домены, запреты, provenance, брокеры	Логи, manifest, договоры
10–14: роли, права, хранение, переводы, crawler	Доступы, tickets, тестовые выгрузки
15–18: фильтры, модели, leakage, copyright	Отчёты тестов, правила и код
19–20: публичность и независимость	Сводки, декларации конфликта

Программа независимого аудита делится на проверку проектных решений и фактического исполнения. На проектном этапе собираются утверждённые цели, архитектура, LIA, DPIA,

карта контролёров, договоры поставщиков и политика исключения доменов. На операционном этапе анализируются логи, реальные конфигурации crawler, версия фильтров, цепочки преобразований, доказательства реакции на жалобы и периодические тесты утечки. Третий этап проверяет способность организации устранить обнаруженную ошибку, а не только показать существование письменной процедуры. Для каждого пункта рекомендуется заранее определить тип доказательства и критерий достаточности, чтобы не менять стандарт проверки ради удобного результата.

Проверка может выполняться в защищённой среде с независимыми наблюдателями и минимальным копированием данных. Отсутствие полномасштабного доступа должно отражаться в заключении, включая типы данных, которые эксперт не смог проверить. Если заказчик отказывается показывать ключевые журналы под предлогом конфиденциальности, это не является автоматическим доказательством нарушения GDPR, но означает ограничение объёма положительного заключения. Отчёт полезно разделять на подтверждённые результаты, недостаточные данные, существенные отклонения и предложения по улучшению. В случае конфликта интересов назначаются независимый рецензент и уполномоченный для решения вопроса о публикации.

Приложение 27. Evidence Chain: образец движения одного документа

Условный документ D-2026-001 найден на разрешённом источнике в определённый день. Лог crawler содержит время и URL, raw object получает хэш, процесс очистки исключает контактные строки, а manifest записывает идентификатор итогового фрагмента. Training run ссылается на версию каталога. Если приходит жалоба, rights ticket включает проверку связанного документа и всех последующих версий.

Доказательство должно позволять ответить: существовал ли объект в момент загрузки; как он менялся; на каком основании был извлечён; попал ли в обучение; какие данные из него сохранились; какие исправления проведены. Без такого следа нельзя установить источник отдельной генерации с надёжностью, которую иногда обещают маркетинговые модели «data lineage».

Нужно различать доказанную цифровую связь и причинность генеративной выдачи. Даже если спорный текст входил в training set, конкретная формулировка может возникнуть из других данных. И наоборот, отсутствие обнаруженного training record при неполном доступе не доказывает его исторического отсутствия.

Артефакт	Контроль целостности
Исходная страница	Время, URL, контекст доступа
Raw object	Хэш, ограниченное хранение
Cleaned fragment	Version ID, фильтрация
Training manifest	Связь с run ID
Rights ticket	Действия и итоговая проверка

Приложение 28. Учебный кейс: ошибочная санкционная атрибуция

Моделируемая ситуация: публичное сообщение связывает однофамильца с лицом, в отношении которого применены санкции. Генеративная система воспроизводит сомнительную связь и не указывает источник, дату и степень уверенности. Этот пример не относится к конкретному человеку или делу. Независимая проверка начинает с фиксации выдачи модели и точных условий получения результата, затем проверяет публичный первоисточник и возможные официальные документы.

Далее определяется вероятность присутствия ошибочного фрагмента в сборке, если доступны исходные сведения; отдельно оценивается использование модели в KYC/AML-процессе. Необходимо отличать технический вывод «при данном запросе модель сформулировала связь» от юридического вывода «лицо действительно имеет санкционный статус». Выбор меры зависит от происхождения ошибки, степени публичного распространения и дальнейших последствий.

План корректирующих действий: временное ограничение рискованной downstream-выдачи, проверка официального списка, корректировка набора, уведомление заинтересованных подразделений, мотивированный ответ лицу и последующий контроль тестов. Никакой этап не должен создавать ещё одну общедоступную копию спорных личных сведений без правового основания.

Для учебного случая ошибочной санкционной атрибуции требуется создать набор искусственных записей, не воспроизводящих реальные приватные обвинения. Условно лицо А и лицо Б имеют сходные фамилии, но разные даты рождения и юрисдикции. В новостной статье упоминается расследование в отношении Б, тогда как санкционная мера в официальном источнике относится к иному субъекту. Через несколько агрегаторов текст превращается в сокращённую запись с неопределённым местоимением. Модель отвечает, что А находится под санкциями, хотя первичный список не подтверждает этого. В протоколе отдельно оценивают качество ответа, возможное смешение записей и отсутствие доказательств, что ошибка возникла именно из конкретного обучающего документа.

Ремедиация в таком кейсе включает проверку официальных актуальных реестров, уточнение идентификаторов, изменение downstream-процедуры принятия решения и канал исправления записи в контролируемом корпусе. Если банковская система использует модель для проверки, необходима самостоятельная оценка требований к финансовому учреждению и человеческой проверке. Учебный пример демонстрирует разницу между допустимым исследованием механизма ошибки и недопустимым распространением обвинений о конкретном человеке без надлежащих доказательств. При публикации желательно использовать синтетическую фабулу и открыто обозначить её вымышленный характер.

Приложение 29. Учебный кейс: устаревшая информация об уголовном преследовании

Моделируемая ситуация: первоначальная онлайн-заметка упоминает расследование, но позже обвинение снято либо судебный акт изменён. В некоторых видах систем такой материал может продолжать влиять на ответы модели. Исследование фиксирует исходную и актуальную процессуальные ситуации отдельно и не приравнивает запись прессы к действующему розыску, международному уведомлению или судебному приговору.

Контрольные точки: применимость статьи 10 GDPR; источник и дата первоначальной публикации; наличие официального акта; качество идентификации лица; правовой статус обучения; возможность исправления и удаления; фактическое поведение модели; компетентный маршрут защиты. Эксперт не утверждает наличия записи в INTERPOL без подтверждения соответствующего компетентного источника.

В заключении сообщаются только подтверждённые факты, механизм риска, ограничение доступа к внутренним данным и меры по предотвращению дальнейшей ошибочной атрибуции. Для заявителя создаётся понятный процесс предоставления актуальных судебных документов, но нельзя перекладывать на него обязанность восстановить всю скрытую архитектуру чужой модели.

Приложение 30. Учебный кейс: брокер продаёт «обезличенный» набор

Моделируемая ситуация: брокер передаёт разработчику архив форумных сообщений с удалёнными именами, но с редкими датами, профессиями и географическими описаниями. На основе других открытых источников отдельные авторы могут быть снова установлены. Название «анонимный» в договоре не снимает юридического вопроса: требуется анализ средств идентификации, реально доступных сторонам и разумно ожидаемым получателям [1][15].

Аудитор запрашивает схему анонимизации, правовое основание первоначального сбора, описание возможных внешних связей и результаты попыток повторной идентификации с разрешёнными тестовыми данными. Если доказательства недостаточны, формулируется «статус анонимности не подтверждён», а не «подтверждён незаконный сбор». Разработчик принимает отдельное решение о допустимости дальнейшего использования.

Практическая мера — временно ограничить использование спорного корпуса, провести targeted тесты, пересмотреть договорную гарантию, при необходимости изменить набор и предусмотреть процедуру обработки будущих заявлений субъектов. Цель метода не только найти нарушение, но и предотвратить дальнейшее неоправданное распространение материала.

Приложение 31. Публичный отчёт о прозрачности: минимальные показатели

Публичный отчёт не должен заявлять процент юридического соответствия, вычисленный из произвольной суммы галочек. Предпочтительны ограниченные проверяемые показатели, по которым ясно определена генеральная совокупность, период и метод. Пример: доля URL в зарегистрированной версии корпуса, для которых известен источник; процент записей с временной меткой; число подтверждённых обращений по правам субъектов, разрешённых в определённый период; среднее время расследования сложного заявления.

Для каждого индикатора указываются числитель, знаменатель, исключённые элементы, сборочный процесс и ограничения. Если извлечение затрагивает множество связующих документов, один URL и один человек не являются взаимозаменяемыми единицами. Внешний читатель должен отличать количество обработанных запросов от количества уникальных субъектов. Методика защищает и регулятора, и общество от неверной статистической картины.

Индикатор	Числитель/знаменатель	Ограничение
Traceability coverage	Объекты с корректной связью / все объекты проверенной версии	Не подтверждает законность автоматически
Filter recall	Выявленные контрольные случаи / все размеченные контрольные случаи	Зависит от тестовой выборки
Rights resolution	Исполненные запросы / принятые к рассмотрению	Не измеряет юридическую правильность отказа
Source timestamp completeness	Источники с отметкой / все проверенные источники	Не доказывает актуальность факта

Приложение 32. Проект предложений к EDPB Guidelines 03/2026

В части accountability предлагается рекомендовать версионированную инвентаризацию источников, позволяющую проверить путь конкретной категории данных до training run. При этом не следует требовать открытой публикации всех идентифицирующих URL, когда это усиливает риск для людей. Дифференцированное раскрытие публичности, надзорным органам и конфиденциальным независимым рецензентам позволит совместить доказуемость и минимизацию.

В части legitimate interest целесообразно включить практические примеры альтернатив «массовый сбор vs ограниченный тематический набор», а также более явное различие между измерением производительности модели и юридическим тестом необходимости. Положительный эксперимент по качеству не заменяет баланс прав, но создаёт проверяемые данные для его проведения. При отсутствии возможности документировать выбор альтернатив решение не должно описываться как полностью мотивированное.

В части transparency важно подчеркнуть, что reliance on Article 14(5)(b) не означает освобождения от публичной информации и защитных мер. Следует рекомендовать структуру searchable source categories, указание периода сбора и прямой канал для субъектов. Для data brokers полезна схема передачи информации о ранее выраженных запретах конечному разработчику.

В части special categories и criminal offence data предлагается ясный workflow «исключить или передать на отдельную правовую проверку», без ложного впечатления, что общедоступность автоматически создаёт исключение из статей 9/10. Эти предложения отражают независимую исследовательскую позицию и не являются уже принятыми нормами.

Предложения к консультации EDPB должны включать ясную формулировку предлагаемого уточнения, правовую проблему, техническую причину возникновения проблемы и пределы применимости. Например, можно предложить методическую рекомендацию документировать категории источников и статистическую эффективность фильтров для широких untargeted-сборов. При этом важно не представить такое предложение как уже существующую обязательную норму. Точно так же просьба конкретизировать уведомления по статье 14 должна быть совместима с тем, что сама статья устанавливает условия и исключения; рекомендации могут объяснять применение, но не переписывать закон.

Второе направление состоит в разработке формулировок для реального ответа субъекту, когда оператор исключил его запись из дальнейшего обучения, но не может подтвердить полное устранение влияния из доступной модели. Рекомендуется различать результат по разным слоям технической системы и указывать ограничения проверки. Третье направление — независимое измерение качества фильтров по языкам и классам данных. Четвёртое — требования к проверяемости заявлений брокеров без передачи всего корпуса аудиторской организации. Такие предложения ориентированы на практическое исполнение принципов GDPR и не предполагают создания новых самостоятельных полномочий для неправительственного Observatoire.

Приложение 33. План внедрения организационной методики на 90 дней

Первые 30 дней посвящены установлению фактического состояния: список проектных целей, роли контролёров и обработчиков, inventory источников, карта трансграничных доступов и первичная оценка LIA/DPIA. Каждому неопределённому источнику присваивается видимый статус с решением о временном ограничении. Запуск новых объёмных сборов без завершённой проверки запрещённых классов рекомендуется приостановить по управленческому решению.

Во вторые 30 дней создают технический контроль: allow/deny registry, versioned manifest, фильтрацию высокорисковых данных, надёжный журнал решений, процесс жалоб и протокол удаления из всех релевантных копий. Тесты строятся на искусственных и правомерно полученных данных. До распространения на все системы обязательно оценивается ложная уверенность в покрытии фильтра.

На 61–90-й день независимая группа выбирает репрезентативные источники, тестирует LIA и dataset lineage, рассматривает права субъектов, сопоставляет публичное описание с фактами и готовит ограниченное по предмету заключение. Неустранённые блокирующие недостатки и неопределённость не закрываются фразой «план внедрения выполнен».

Период	Результат	Приёмочный критерий
День 1–30	Карта целей, ролей и источников	Неназванные риски видимы в реестре
День 31–60	Фильтры, логи, rights workflow	Проведены документированные испытания
День 61–90	Независимая выборка и план исправлений	Выводы ограничены проверенными фактами

Приложение 34. Критерии принятия отчёта независимой проверкой

Отчёт может быть представлен руководству при наличии: определённого объекта и периода проверки; карты норм и статуса руководящих документов; списка полученных доказательств; результатов тестов; перечня неохваченных источников; выявленных недостатков и мер устранения; независимого редакционного решения. По высокорисковым персональным данным должны быть отдельно отражены статьи 9 и 10, а по непрямому получению — основания и пределы статьи 14(5)(b) GDPR.

Блокирующие условия принятия: отсутствие любой определённой цели обработки при продолжающемся массовом сборе; невозможность установить собственную роль и контрагентов; доказанный захват особых категорий без необходимой правовой проверки; отсутствие механизма зафиксировать спорный источник или учесть обращение субъекта в пределах доступных данных; публичное заявление о полной законности без релевантных подтверждений. Это предлагаемые организационные критерии качества, а не закрытый перечень правонарушений.

Заключение должно содержать явное «не проверено», когда доступ к внутренней модели, части корпусов или историческим логам не был предоставлен. Такая прозрачность повышает ценность исследования: заинтересованные лица смогут отличить поддающийся установлению факт от гипотезы и поставить отдельные вопросы в компетентной процедуре.

Приложение 35. Реестр правовых и технических вопросов для последующей рецензии

Перед официальной публикацией полного исследования правовой рецензент проверяет редакцию GDPR и AI Act, статус рекомендаций EDPB 03/2026, применимость французских положений и национальных имплементаций DSM/TDM, а также корректность специальных режимов персональных данных. Технический рецензент сверяет архитектуру lineage, различие удаления файла и изменения весов, воспроизводимость выборочного аудита и валидность предлагаемой формулы неопределённости.

Отдельно требуется проверить, что ни один приведённый учебный пример не может быть воспринят как описание конкретного клиента, а методологическое предложение — как уже полученное официальное полномочие Observatoire ARGA. Опубликованные контакты, логотип и реквизиты организации должны соответствовать фактической утверждённой редакции. Если документ используется для подготовки позиции по консультации EDPB, он не подменяет самостоятельно оформленное заявление и подтверждение отправки.

Предмет рецензии	Кто проверяет	Что нужно получить
Юридическая актуальность	Профильный юрист	Датированное заключение
Точность ссылок	Редактор	Активные первоисточники
Цифровая методика	Технический специалист	Тест протоколов и ограничений
Статистические примеры	Аналитик	Пересчитанные показатели
Независимость и публикация	Редакционный орган	Решение, конфликты и scope

Приложение 36. Проектирование crawler как юридически контролируемой системы

На стадии проектирования автоматического сборщика целесообразно ввести разделение policy plane и execution plane. Первый содержит утверждённую цель, допустимые домены, категории контента, временные окна, ограничения региона, исключения для уязвимых групп и максимальные сроки хранения. Второй выполняет сеть запросов согласно конкретной версии политики. Каждый crawl job должен ссылаться на неизменяемый policy ID, иначе невозможно воспроизвести, каким правилам подчинялся извлекатель в момент спорной

операции. Наличие policy-файла без обязательной проверки перед запросом не является действительным контролем.

Особое значение имеет семантика правила «разрешён домен». Поддомены, зеркала, архивы, URL-параметры и URL-shorteners могут приводить к тому, что одобренная категория неожиданно расширяется. Проект требует нормализации URL, канонизации домена и обязательной проверки после redirect. Для динамических сайтов с браузерным рендером необходимо учитывать доступность дополнительных элементов, которые не видны в исходном HTML. Любое изменение способа открытия страницы может менять и юридически релевантные факты доступа, например прохождение аутентификации либо ранее скрытой части страницы.

Перед извлечением реализуется последовательность проверок: принадлежность ресурса утверждённой категории; отсутствие в deny registry; учёт выраженных технических ограничений; контроль допустимого типа документа; исключение путей с высокой вероятностью чувствительных сведений; лимит на частоту и глубину обхода. При негативном результате задание должно быть заблокировано, а причина сохранена как запись аудита. Инженер не должен иметь возможность незаметно отключить все предохранители ради повышения объёма собранного корпуса.

В спорном случае следует различать отсутствие запрета и наличие юридического разрешения. Если robots.txt не содержит запрета, это не означает, что обработка соответствующих персональных данных имеет основание по статье 6 GDPR. И наоборот, технический запрет сайта может касаться договорного или авторско-правового режима, который исследуется параллельно с GDPR. Регуляторный проект EDPB и практика CNIL относят выраженное сопротивление ресурса к значимым признакам оценки; индивидуальная применимость норм должна быть установлена по фактам [2][6].

Независимый тест проводится с моделируемыми ресурсами: разрешённый домен, запрещённый путь, перенаправление на запрещённый путь, вложение с персональными сведениями, отсутствие связи с реестром правил и изменение deny registry во время выполнения. Отчёт включает не только успешность блокировок, но и то, какие этапы были недоступны при аудите, были ли скрытые административные исключения, кто мог менять правила и сохранялась ли история их действия. Рекомендуемая архитектура обеспечивает не декларативную «этику crawler», а проверяемое исполнение определённых ограничений.

Контрольный объект	Тест отказа	Доказательство результата
Policy ID	Задание без утверждённой политики	Запуск отклонён и записан
Домен	Redirect на запрещённый ресурс	Повторная проверка конечного адреса
Sensitive path	Ссылка на детские/медицинские материалы	Нет извлечения, есть лог запрета
Изменение запрета	Новая deny-запись во время job	Определена дата и область действия
Административный обход	Попытка отключить фильтр без полномочий	Отказ / отдельное одобрение

Для crawler проектируется цепь двух независимых контуров: правила допуска и фактический исполнитель. Первый утверждает цель и классы источников, запрещённые домены, срок

хранения, предельный объём и обязательные действия при сомнениях. Второй исполняет задания, но не может самовольно расширять разрешённую географию или тип данных только потому, что нашёл новые ссылки. Между ними вводится журнал изменений: кто предложил правило, кто утвердил, какая версия вступила в силу и какие задания были запущены до её отмены. Это позволяет доказательно проверять исполнение и оперативно остановить сбор спорной категории.

Контрольные тесты должны охватывать не только штатную работу. Например, проверяется поведение при неожиданной авторизации, перенаправлении на сайт медицинской тематики, изменении robots.txt, появлении API с данными несовершеннолетних, аварии фильтра и повторном запуске после сбоя. При техническом отказе предпочтительно безопасное поведение: временная остановка или помещение задания в очередь на разбор, а не расширение сбора по умолчанию. События остановки фиксируются как доказательства реального контроля. Результат такого испытания не подтверждает автоматически законность самой выбранной цели, поэтому юридический и технический блоки аудитора сохраняются раздельными.

Приложение 37. Судебная и доказательная оценка происхождения спорного фрагмента

В условиях правового спора часто требуется установить, когда опубликован определённый материал, когда он был извлечён, кто имел возможность управлять содержимым и в каком виде оно передано в обучение. Простая ссылка на текущую веб-страницу недостаточна: страница могла измениться или исчезнуть. Должен существовать протокол фиксации версии с учётом правомерности самого сохранения, хэширования копии, сетевого журнала и описания контекста доступа. Стороне спора необходимо предоставить достаточные данные для проверки происхождения без безразборного распространения чужой частной информации.

Временные отметки должны относиться к разным событиям: создание первичного текста, публикация на сайте, первое наблюдение crawler, фактическая загрузка, первичная обработка, включение в dataset version и использование в training run. Они не взаимозаменяемы. Например, архивная страница с датой публикации в 2018 году, впервые скачанная в 2026 году, не доказывает, что разработчик обладал сведениями с 2018 года. При построении хронологии аудитор отдельно рассматривает возможное рассогласование часов и альтернативные источники отметок времени.

Причинная связь между обучающим документом и конкретным выходом генеративной модели особенно трудна. Даже полное совпадение фразы не исключает самостоятельного возникновения текста из нескольких источников. При отсутствии доступа к training runs внешнее тестирование позволяет зафиксировать воспроизводимость выхода в оговорённых условиях, но не всегда установить конкретный источник. В юридическом заключении следует использовать отдельные уровни формулировки: факт извлечения; факт включения в набор; факт использования при обучении; факт воспроизведения; степень доказанности связи.

Процедура учитывает защиту коммерческой тайны разработчика и персональных прав других субъектов. Возможно ограниченное предоставление технического manifest независимому эксперту, обезличенная статистика источников или судебный режим доступа к

закрытым доказательствам. Однако нельзя описывать не предоставленный журнал как проверенный. Участнику предоставляется возможность оспорить идентификацию записи, условия доступа и временные утверждения, а аудитор должен сохранять доказательства альтернативной версии.

При выявлении ошибки рекомендуется составить отдельную матрицу восстановления: что точно известно о страницах, что известно о копиях, какие изменения подтверждены и какое влияние нельзя установить. Для международного розыска, санкций или финансовых решений такая матрица особенно важна: вывод о правовом статусе лица должен опираться на компетентный официальный источник, а не на совпадение поисковой выдачи и фразы модели.

Доказательный вопрос	Возможный объект	Допустимый вывод
Существовал ли текст?	Версионированная копия и атрибуция	Историческое состояние источника
Был ли он загружен?	Crawl log, manifest, hash	Факт конкретного извлечения
Вошёл ли в обучение?	Training run manifest	Использование в оговорённой версии
Был ли воспроизведён?	Повторяемый test transcript	Выход модели при данном тесте
Какова причинная связь?	Совокупность доступных фактов	Уровень доказательности с оговорками

Приложение 38. Деперсонализация, linking и риск обратной атрибуции

Одна из слабостей упрощённой оценки данных заключается в том, что фильтр идентификаторов тестируется по отдельным строкам, тогда как реидентификация может происходить через связь между несколькими фрагментами. Профессиональная должность, приблизительное место проживания, необычная дата и описание редкого события в совокупности способны указать на человека без прямого имени. При оценке анонимизации нужно рассматривать не только уникальность отдельного атрибута, но и возможность singling out, linkability и inference в конкретной модели угроз [1][15].

Пример синтетического теста: один учебный фрагмент говорит о редкой должности, второй — о переезде в небольшой город, третий — о конкретной профессиональной награде. По отдельности сведения не устанавливают лицо, но при соединении с доступным справочником могут существенно повысить вероятность идентификации. Тест проводится на искусственных профилях, чтобы не создавать новый риск, и фиксирует набор доступных противнику внешних данных. Результат не сводится к «обнаружено совпадение имени»; важна вероятность связывания и качество альтернативных объяснений.

В архитектуре больших генеративных моделей linking возникает не только внутри тренировочного корпуса. Retrieval-слой, plug-ins внешнего поиска, пользовательские запросы и другие базы могут предоставить недостающие атрибуты. Поэтому аудит рассматривает хотя бы типовые downstream-сценарии, в которых система соединяет воспроизводимые фрагменты с новыми источниками. Наличие безопасной исходной очистки не гарантирует отсутствие комбинационных рисков в готовом продукте.

Доказательная форма содержит список проверенных угроз, доступные вспомогательные наборы, права доступа, меры минимизации, степень устойчивости к linking и условия, при

изменении которых вывод требует пересмотра. В случае невозможности проверить реальные внешние базы устанавливается ограниченная область заключения. Для конфиденциальных форумных и медицинских историй рекомендуется более строгая стратегия исключения, а не бесконечное совершенствование вероятностного фильтра.

Юридически следует отдельно решить, для кого и при каких условиях сведения способны идентифицировать человека. Слово «псевдонимизировано» не исключает GDPR, а предварительная оценка анонимности должна быть пересмотрена при появлении новых средств объединения или изменении доступа. Контролёр документирует именно основание своего вывода, а не степень собственного доверия к названию технологии.

Проверка	Сценарий	Критерий отчёта
Singling out	Редкие сочетания полей	Можно ли выделить запись без имени?
Linking	Две разрешённые тестовые базы	Есть ли устойчивое совпадение?
Inference	Вывод чувствительного свойства	Можно ли обоснованно получить личный атрибут?
Модельный вывод	Связь retrieval и ответа	Сохраняется ли риск после фильтра?
Пересмотр	Новые источники или права доступа	Нужна ли повторная DPIA?

Приложение 39. Управление удалением данных из резервных копий и обучающих эпох

Удаление спорной записи представляет собой задачу управления жизненным циклом, а не одиночный файловый вызов. Большой корпус может иметь raw snapshot, deduplicated snapshot, токенизированный кеш, shards, промежуточные checkpoints, embeddings и backup-версии. Некоторые технические копии могут иметь непрямую связь с исходным URL. Поэтому восстановление затронутых объектов требует прослеживаемости до операции построения соответствующей версии; отсутствие такой способности увеличивает неопределённость отчёта [1][10].

Процедура начинается с ограничения дальнейшего доступа и заморозки нового обучения на спорном фрагменте. Затем определяются связанные идентификаторы, цель и основание удержания каждой копии. Применимые обязанности сохранения определённых доказательств или резервных элементов рассматриваются отдельно; удаление не должно уничтожать необходимые по закону сведения без правового анализа. Разработчик документирует, что именно удаляется, что сохраняется на законном основании и в течение какого времени.

Если модель уже обучалась на фрагменте, технические варианты включают retraining без данных, подтверждённый метод machine unlearning, ограничение retrieval и подходящие фильтры ответа. Ни один метод не объявляется универсально эквивалентным стиранию. Испытания могут включать извлечение канареечных искусственных последовательностей, меморизационные стресс-тесты и контроль изменения поведения на законных задачах. Внешний отчёт указывает конфигурацию и ограничение, не претендуя на доказательство отсутствия любого скрытого влияния.

Системный процесс предусматривает журнал requests, отдельные статусы raw removal, index removal, training future exclusion и model risk mitigation. Для каждого статуса существует проверочный артефакт и владелец исполнения. Если отсутствие traceability препятствует выполнению части мер, это должно быть явно сообщено по законным требованиям, с реальным планом устранения технического ограничения. Ложное подтверждение полного удаления способно ухудшить положение заявителя.

Для международных организаций следует учитывать копии у подрядчиков. Смена имени корпуса или перенос его в другую юрисдикцию не устраняет исходную задачу по правам субъекта. Договоры со связанными обработчиками должны обеспечивать координированную передачу запросов и доказательств исполнения. Аудитор проверяет выборку завершённых дел, сопоставляя обещанные действия с реально изменёнными техническими объектами.

Объект	Действие	Условие завершения
Raw и clean sets	Исключение/ограничение доступа	Согласованные manifest и логи
Temporary caches	Удаление после технического срока	Проверка lifecycle/retention
Retrieval index	Удаление и переиндексация	Тест поисковой выдачи
Future training	Внесение отрицательных ограничений	Контроль следующего run
Model weights	Ремедиация и тест остаточного риска	Ограниченный вывод без ложной гарантии

Приложение 40. Процедура оспаривания решения независимого аудитора

Независимость внешнего отчёта не освобождает его автора от процессуальной справедливости. Проверяемой организации должны быть представлены конкретные установленные факты, описание тестов и разумная возможность исправить неточности до публикации, если это не уничтожает доказательства и не создаёт угрозы заявителю. Вместе с тем право представить объяснения не означает право организации заблокировать добросовестную публикацию только из-за несогласия с неблагоприятным выводом.

Процедура предусматривает четыре стадии. Сначала формируется evidence list с указанием доступности материалов. Затем формулируются предварительные выводы, строго разделяющие наблюдение и оценку. После получения позиции проверяемой стороны аудитор фиксирует, какие факты приняты, опровергнуты, остались спорными или недоступными для независимой проверки. Наконец, редакционный орган принимает мотивированное решение, оценивая необходимость защитить персональные сведения и источники сообщения.

Когда спор касается конкретного лица, его права и безопасность могут потребовать ограниченного раскрытия материалов другой стороне. Например, опубликовать полное содержание жалобы о состоянии здоровья нельзя только ради иллюстрации ошибки модели. При необходимости допускается безопасная обобщённая версия, удаление идентификаторов и выделение документальных доказательств, необходимых для опровержения спорного утверждения. Целесообразно заранее определить правила защиты информаторов и конфиденциальных источников.

Роль аудитора должна оставаться отделённой от юридического представительства одной из сторон, если это создаёт существенный конфликт интересов. Observatoire ARGA может одновременно развивать методологию и вести другие виды деятельности в пределах своих полномочий, но конкретный вывод требует раскрытия связи с финансированием, коммерческим поручением или заинтересованным заявителем. При высоком конфликте предлагается внешняя рецензия либо отвод.

Итоговый отчёт содержит раздел dissent или response, если возражение существенно для понимания доказательной картины. После публикации действует процедура исправления фактических ошибок с сохранением истории версий, чтобы добросовестная коррекция не превращалась в незаметное переписывание уже распространённого текста. Это повышает воспроизводимость исследования и уменьшает риск нового вреда от неверной экспертной публикации.

Стадия	Документ	Процессуальная гарантия
Предварительные наблюдения	Evidence list	Проверка фактов, не обвинение
Ответ организации	Response record	Возможность представить контрдоказательства
Редакционное решение	Decision memo	Независимость и причины изменений
Публикация	Redacted report	Защита персональных сведений
Исправление	Version history	Прозрачное объяснение правки

Приложение 41. Проектирование понятного уведомления о scraping: тест с реальным пользователем

Принцип прозрачности не достигается одним юридически выверенным, но нечитаемым документом. При сборе общедоступных источников человек может вообще не знать, что его публикации использовались для разработки ИИ, тем более не знать названия конкретной модели, разработчика или посредника. Статья 14 GDPR требует информирования в предусмотренном объёме, а при наличии соответствующих условий исключения из индивидуального уведомления по статье 14(5)(b) предполагает надлежащие альтернативные меры, включая публичную доступность сведений [1][2].

Для независимой проверки предлагается использовать тест «пять вопросов». Может ли читатель за несколько минут установить: какая организация выступает контролёром; ради какой определённой цели собирается его информация; какие типы веб-источников затронуты; какие права в его ситуации доступны; каким способом обратиться и получить мотивированный ответ. Такой тест является методическим инструментом и не заменяет формальных требований регламента. На странице можно разместить поиск по категориям сайтов и временным периодам, но не обязательно раскрывать URL, содержащие чужие чувствительные сведения.

Система уведомлений должна быть многоуровневой. Краткая публичная страница объясняет общий смысл, документ второго уровня содержит обязательную информацию, а техническое приложение показывает классы источников и границы обработок. Для регулятора можно подготовить закрытый набор с более точными списками, если раскрытие допустимо. При

составлении языковых версий следует проверять одинаковость существенных правовых утверждений и корректность понятия «возражение», а не ограничиваться машинным переводом кнопок.

Полезен независимый пользовательский тест с разными типами затронутых лиц: автор публичной статьи, бывший пользователь форума, представитель организации и человек, ошибочно связанный с санкционным лицом. Участникам предлагают определить, куда направить возражение, как узнать о возможной обработке и где находится перечень источников. Интервью не должно требовать раскрывать свои реальные чувствительные данные; применяются учебные ситуации. Измеряются успех, время поиска, неправильные представления и необходимость дополнительной помощи.

В отчёте следует отдельно раскрывать известную полноту каталога источников. Если компания не знает точного содержания приобретённого датасета, заявление «вы можете проверить все ваши данные в каталоге» вводит в заблуждение. В таком случае уведомление должно честно описывать ограничения контроля и реальный канал расследования, а разработчик — иметь план восстановления происхождения и условий обработки.

Вопрос человека	Ожидаемый элемент интерфейса
«Как вы могли получить мой текст?»	Категории источников и период сборки
«Кто отвечает?»	Наименование контролёра и контакт
«Что именно происходит с данными?»	Описание сбора, обучения, выдачи
«Могу ли я возразить?»	Рабочая форма с правовой квалификацией
«Что меняется после обращения?»	Разделение удаления из набора и мер для модели

Приложение 42. Различие частного scraping и правоприменительного доступа к данным

Проект Guidelines 03/2026 EDPB ограничен случаями сбора из внешних интернет-источников в связи с генеративным ИИ, выполняемого частным сектором [2]. Это ограничение имеет практическое значение: исследования правоохранительных баз, государственный доступ к уголовно-процессуальным данным и автоматизированные решения компетентных органов могут подпадать под другие специальные правила, включая законодательство об обработке данных в правоохранительных целях. Нельзя переносить на такие режимы все частноправовые примеры EDPB без самостоятельной квалификации.

Для ARGA важен сценарий, когда публичная информация об уголовном деле поступает в коммерческий корпус, а downstream-организация использует модель при проверке клиента, заявителя на убежище или участника судебного процесса. Частный сбор и официальное правоприменение могут быть разными обработками со своим законным основанием, целью и компетентным надзором. Если модель производит ошибочную биографическую ассоциацию, исправление тренировочного корпуса не подменяет пересмотра административного либо судебного решения.

Информационная цепочка должна сохранять источник правового утверждения: официальный судебный акт, пресс-релиз, новость, агрегатор либо генерируемый вывод. Для каждой стадии обозначаются степень надёжности, срок действия и возможность

опровержения. Если правовой статус лица зависит от конкретного официального решения, текст СМИ не заменяет первичный акт. Вопрос, была ли запись в системах международного полицейского сотрудничества, нельзя определять по выводу генеративной модели.

Предлагаемая методика включает отдельный блок «межрежимной границы». Он задаёт вопрос: кто использует результат и по какому праву; какая юридическая сила приписывается выводу; предусмотрена ли человеческая проверка; доступно ли лицу исправление исходных данных и обжалование последующего решения. Этот подход позволяет не сводить все последствия к одной проверке GDPR, игнорируя отраслевые процессуальные гарантии.

Независимое исследование при отсутствии доступа к государственным базам не должно делать утверждений об их содержимом. Оно может оценить общедоступную выдачу модели, правовую архитектуру интерфейса и документы, добровольно предоставленные заинтересованной стороной, но должно ясно обозначить пределы компетенции. Правильная квалификация границ повышает доказательную ценность работы и защищает субъектов от нового распространения спорных обвинений.

Этап	Что нельзя подменять	Надлежащая проверка
Публичная новость	Официальный статус дела	Первичный компетентный акт
Коммерческая модель	Полицейская база	Ограниченный тест модели
Частный scraping	Правоприменительный сбор	Раздельная карта режимов
Техническое исправление	Судебное обжалование	Самостоятельный процессуальный путь

Приложение 43. Предобучение, дообучение, retrieval и обработка пользовательских запросов

Обозначение «модель обучена на открытых данных» скрывает несколько разных стадий. Предобучение может включать огромные разнородные корпуса; fine-tuning обычно использует целевые коллекции меньшего объёма; retrieval-augmented generation обращается к внешнему индексу без необходимости менять основные параметры модели; эксплуатация сохраняет или обрабатывает пользовательские запросы и журналы. Каждый процесс имеет собственную цель, контролёра, источник, период и возможное законное основание [1][9][10].

Проблема особенно заметна при исполнении запроса об удалении. Если имя субъекта воспроизводится из retrieval-индекса, удаление или ограничение индекса может изменить результат быстро и проверяемо. Если сведения были усвоены в параметрах модели при предобучении, требуются другие технические меры. Если они появились исключительно из текущего запроса пользователя, правомерность исходного scraping может вообще не быть главным вопросом. Процедура не должна путать все три случая.

Для анализа предлагаем разделить модельную карту на четыре уровня. Level 1 — исходные сохранённые данные и доступ к ним; Level 2 — подготовленные training/fine-tuning наборы; Level 3 — параметры и checkpoints; Level 4 — служебные индексы, плагины, prompts и журналы эксплуатации. Для каждого документируются категории персональных данных, техническая возможность поиска, срок хранения и ответственный контролёр. Наличие одного общего регистра «AI data» слишком грубо для воспроизводимой проверки.

Специально рассматривается continuous learning и последующее пользовательское обучение. Если коммерческая система сохраняет промпты для улучшения модели, цель такого использования должна быть отдельно объяснена, а права субъекта и действующие ограничения — оценены. Нельзя автоматически переносить основание для ответа на пользовательский запрос на непрерывное обучение на всех вводимых персональных сведениях. ICO справедливо выделяет раздельность целей разработки модели и downstream-применений как важный элемент purpose limitation [9].

Независимая оценка требует фактической архитектурной диаграммы и возможности теста на каждом уровне. При отсутствии такого разграничения организация рискует отвечать заявителю о «данных в модели», когда подтверждены только сведения в журнале API. Эксперт обязан обозначить эту подмену как ограничение, но не домысливать неизвестную внутреннюю архитектуру.

Стадия	Объект контроля	Типичное исправление
Предобучение	Large training set и weights	Исключение в следующих runs, проверка остатка
Fine-tuning	Специализированные записи	Пересборка/отмена версии при необходимости
Retrieval	Внешняя индексируемая база	Удаление записи и переиндексация
Эксплуатация	Prompts, logs, responses	Политика хранения и доступ к правам

Приложение 44. Метод ответа на распространённые недоказанные правовые утверждения

Утверждение «любая информация в сети находится в общественном достоянии» смешивает четыре разных понятия: открытый технический доступ, защиту персональных данных, авторско-правовой статус произведения и условия повторного коммерческого использования. Для юридического ответа необходимо последовательно установить, к какому правовому объекту относится фраза. Иначе спор превращается в обмен лозунгами и не позволяет оценить конкретную обработку [1][18].

Утверждение «мы действуем во имя технологического прогресса, поэтому имеется legitimate interest» не завершает тест статьи 6(1)(f). Да, интерес инновационной или коммерческой деятельности может быть законным, но отдельно доказываются необходимость и баланс прав. Существование позитивной общественной цели не устраняет обязанности учитывать масштаб невидимого сбора, старые сведения, уязвимые группы, информирование и правомочие возражения [2][3][5].

Утверждение «мы полностью анонимизировали данные, потому что удалили имена» требует технической проверки. Сохраняются редкие сочетания дат и обстоятельств, географические и социальные признаки, embeddings или возможность связи с другим набором. Анонимность должна оцениваться для рассматриваемого контекста, а не выводиться из одного преобразования. При невозможности проверить представленные средства допустимо констатировать, что заявленное свойство не подтверждено доступными доказательствами [15].

Утверждение «данные удалены из модели» требует определения объекта: raw data, clean dataset, retrieval index, fine-tuning checkpoint, weights или output filter. Для каждого требуется подтверждение. Нельзя обещать полное стирание влияния без методики тестирования; нельзя и утверждать, что всякое исправление невозможно. Разработчик должен предоставить информированный и технически честный ответ по конкретному запросу [10].

Наконец, утверждение «у нас есть внешний сертификат, значит сбор законен» подменяет область проверки. Сертификация системы управления, например безопасности, может подтвердить определённые организационные механизмы, но не законность каждого источника или основания обработки. Публичная проверка должна указывать предмет удостоверения, применяемый стандарт и ограничения. Такой стандарт ответов позволяет Observatoire ARGA не занимать заранее обвинительную или оправдательную позицию.

Заявление	Какая проверка обязательна
«Всё открыто и разрешено»	Право доступа, GDPR, контент, договоры
«Инновация — достаточный интерес»	Три этапа LIA
«Имена удалены — данные анонимны»	Linking/inference и контекст идентификации
«Удалили из ИИ»	Уровень модели и проверенный эффект
«Сертификат закрывает вопросы»	Область сертификации и предмет нормы

Приложение 45. Кейс переговоров с брокером: подробная процедура доказательной закупки

Предположим, разработчик намерен приобрести большой архив общедоступных сообщений, но брокер не раскрывает исходный список сайтов, ссылаясь на коммерческую тайну. Сначала покупатель запрашивает не необработанный весь корпус, а категориальный manifest: страны и типы доменов, периоды, заявленное основание сбора, исключения для личных историй, признаки статей 9/10, права на контент и механизм последующего удаления. Такой запрос сочетает необходимость проверки с принципом минимизации раскрытия данных [1][2].

Следующий этап — выборочная верификация. Стороны могут согласовать в закрытой среде проверку ограниченного числа записей, отобранных независимым способом, с фиксацией направления выборки, известной генеральной совокупности и способа обеспечения конфиденциальности. Коммерческая тайна не должна превращаться в пустое заявление «юрист брокера видел, значит всё законно», однако аудитор не вправе требовать раскрытия чувствительных сведений сверх допустимого объёма только ради удобства анализа.

После предварительных тестов договор должен определить запретные классы источников, обязанность сообщать об обнаруженных спорных данных, процесс совместного исполнения запросов субъектов и возможность оперативно исключать целые подмножества до завершения расследования. Устанавливаются технические идентификаторы версий, логика запрета на обучение старым batch и условия приостановки поставок при существенном отказе предоставлять доказательства.

Применимое право и юрисдикция рассматриваются отдельно от происхождения. Даже если поставщик базируется вне ЕС, могут возникнуть обязанности по GDPR при выполнении

соответствующих условий. В то же время сама покупка набора европейской организацией не означает, что все данные автоматически относятся к жителям ЕС. Необходимо анализировать территориальные факты, гражданские права и конфиденциальность по реальному составу обработки.

Финальное решение подразделяется на «принять при условиях», «запросить дополнительные данные», «исключить конкретные категории», «отложить закупку» и «отказаться». Каждое решение имеет письменные причины. Если существенный класс данных нельзя проверить, формальное принятие риска руководством не отменяет обязанность соблюдать право; оно лишь отражает организационную ответственность за выбранную позицию.

Этап закупки	Документ	Блокирующее обстоятельство
Предварительная оценка	Purpose/legal basis map	Не определён контролёр
Проверка происхождения	Category manifest	Отсутствует traceability
Выборочный просмотр	Протокол ограниченного доступа	Неуправляемое раскрытие PII
Договор	Rights/remediation clauses	Невозможно исполнить запрос субъекта
Начало обучения	Допуск dataset version	Незакрытые high-risk категории

При предварительном согласовании закупки условного корпуса юрист, инженер по данным и ответственный за защиту данных должны работать над одной версионной карточкой. Юрист устанавливает применимые основания и ограничения, инженер проверяет фактическую структуру и воспроизводимость фильтров, а специалист по защите данных оценивает риски для людей. Если менеджер по закупке получает только три несвязанных одобрения, неизвестно, к одному ли набору относятся выводы. Поэтому необходим сквозной идентификатор поставки и блокирующее правило: значимое изменение категорий источников или способа обработки возвращает карточку на повторную оценку.

Договор может устанавливать санкции за недостоверные заверения, но санкции не заменяют обязанность предотвращать неправомерную обработку. Особое внимание уделяют условиям прекращения договора: сохранению резервных копий, возврату исходных данных, помощи при запросах субъектов, удалению рабочих копий у субподрядчиков и исключению материала из будущих обновлений. Если продавец не контролирует уже перепроданные версии, нельзя обещать глобальное удаление всем затронутым лицам. Учебный акт итоговой проверки должен отдельно содержать подтверждённые факты, нерешённые ограничения и конкретное решение: допуск, ограниченный допуск, отсрочка или отказ.

Приложение 46. Обоснование несоразмерного усилия при индивидуальном уведомлении

Исключение статьи 14(5)(b) GDPR не должно превращаться в автоматический выход из обязательств прозрачности. Для его применения контролёр обязан исследовать невозможность или несоразмерность индивидуального информирования с учётом характеристик массива и влияния отсутствия уведомления на субъектов. В проекте EDPB 03/2026 обсуждаются число лиц, давность информации и возможные защитные меры; подчёркивается необходимость сохранять публичную доступность предусмотренной информации [1][2].

Методика предлагает обосновывать несоразмерность через пять вопросов. Во-первых, существует ли у контролёра достоверный способ связаться с затронутыми лицами без дополнительной массовой обработки? Во-вторых, каков ориентировочный охват лиц и степень дублирования? В-третьих, какие действия придётся выполнить только ради уведомления и какой новый риск они создадут? В-четвёртых, насколько отсутствие сообщения помешает человеку воспользоваться правами? В-пятых, какие публичные, технические и организационные меры реально уменьшают эти последствия?

Два учебных примера показывают пределы обобщений. В первом проект обрабатывает несколько сотен добровольно предоставленных профессиональных резюме с действующими адресами, и само по себе массовое назначение задачи не делает индивидуальное уведомление несоразмерным. Во втором обрабатывается историческая библиотека форумных обсуждений, многие записи псевдонимны, контакты отсутствуют, и для индивидуальных сообщений потребовалось бы новое сопоставление с социальными сетями. Здесь обстоятельства могут поддерживать аргумент об избыточном вмешательстве ради уведомления. Однако окончательный юридический вывод требует полного анализа, а не простой арифметики стоимости.

Если контролёр ссылается на исключение, публичное уведомление должно содержать существенные сведения статьи 14 и доступный способ обращения. Недостаточно разместить на сайте универсальную фразу «мы можем использовать данные из интернета». Следует определять категории источников, цель, основания обработки, временные диапазоны и ответственных лиц. При наличии пробелов честно сообщаются границы реестра, а не придумывается точность.

Независимая проверка включает выборочную проверку доступных контактов, внутреннюю оценку затрат без доступа к лишним данным и тест пользовательской доступности публичного уведомления. В отчёте следует отдельно изложить факты, правовую оценку и степень зависимости результата от предположений. Если условия со временем изменились, решение подлежит пересмотру.

Фактор	Что подтвердить	Возможное возражение
Число субъектов	Реалистичный метод оценки	Дублирование сообщений
Доступность адресов	Возможность связи без нового сбора	Неполнота и устаревание
Давность данных	Распределение по периодам	Новые сведения могут быть доступнее
Вред от отсутствия уведомления	Последствия для реализации прав	Недостаточные альтернативные меры
Публичное информирование	Доступность и полнота	Формальная страница без смысла

Приложение 47. Национальные правовые ограничения публичных реестров: карта коллизий

Публичный реестр не представляет единого юридического режима. Существуют различия между торговыми, судебными, профессиональными, административными и архивными системами; доступ может преследовать определённую публичную цель и сопровождаться ограничениями массового выгружения, повторного использования, поиска по физическому лицу или сохранения исторических данных. Разработчик должен устанавливать условия

каждого класса ресурса и соответствующей юрисдикции, а не делать общий вывод из того, что веб-интерфейс доступен [1][2].

В исследовании различных государств необходимо отделить четыре слоя: правила доступа к первичному реестру; основания обработки персональных данных при самостоятельном извлечении; допустимость дальнейшей коммерческой или исследовательской переработки; правила использования результата в решениях о конкретном человеке. Законный просмотр публичной карточки может не давать достаточного основания для создания глобальной базы биографических рисков. В некоторых контекстах дополнительное значение имеют ограничения по сохранению сведений о правонарушениях и судебных решениях.

Типичный технический риск проявляется при интеграции нескольких реестров, когда одинаковые имена и похожие даты создают ошибочные персональные связи. Поэтому независимо от условий доступа нужны правила идентификации, атрибуции первичного правового статуса, фиксации времени и пересмотра результата. Ошибочные связи с уголовным преследованием, санкциями или несостоятельностью не должны становиться самостоятельным машинным юридическим фактом.

Юрисдикционная карта содержит название страны и органа, категорию источника, публичную цель, тип возможных персональных данных, правовой статус массового извлечения, наличие специальных правил повторного использования и возможный трансграничный доступ. Источники национального права должны быть проверены профильным юристом при конкретном проекте: данный международный доклад не претендует на исчерпывающую таблицу действующих запретов всех государств.

Для независимого аудитора полезно применять правило «не переносить допуск»: разрешение одной категории реестров в одной стране не означает допустимость аналогичного сбора из другой страны. Организационное решение сопровождается датой и ссылкой на применимый акт или официальную политику, а при отсутствии надёжного материала устанавливается статус ожидания правовой проверки. Такая модель снижает риск скрытой подмены международной сравнительной работы удобным общим лозунгом.

Слой анализа	Контрольный вопрос
Первичный доступ	Кто и на каких условиях публикует реестр?
Массовый сбор	Разрешён ли автоматический объём операций?
Персональные данные	Какие GDPR/национальные правила применимы?
Сведения статьи 10	Есть ли специальное законное разрешение?
Downstream-решение	Как проверяется актуальный правовой статус?

Приложение 48. Статистическое сравнение качества двух систем фильтрации

Сравнение версий фильтра должно опираться на фиксированный контрольный набор, ясную разметку и заранее определённые показатели. Если команда меняет тестовую выборку одновременно с алгоритмом, наблюдаемое улучшение может объясняться облегчением теста. Нужны независимая разметка, временная фиксация критериев и отдельный adversarial

набор для редких случаев. Результат по одному языку не переносится автоматически на другие.

Предположим, фильтр версии A проверен на 100 синтетических примерах со сведениями специальной категории и выявил 82, а фильтр B на той же выборке выявил 92. На этом наборе наблюдаемая чувствительность увеличилась с 82% до 92%, то есть на 10 процентных пунктов. Но нужно сохранить парные данные: какие именно десять примеров стали обнаруживаться и не появились ли новые пропуски при других типах. Простое сравнение двух процентов не устанавливает статистическую значимость изменения и не характеризует ложные положительные результаты.

Для юридического применения важна цена каждой ошибки. Если из-за ложного отрицания сведения о здоровье попадут в тренировочный набор, последствия могут быть серьёзнее обычного ухудшения качества датасета. С другой стороны, избыточный фильтр может удалять законные материалы определённого языка, культуры или научной дисциплины, искажающие результаты модели. Выбор порога требует анализа обеих сторон риска без юридически фиктивных «баллов законности».

При последующем использовании OCR, перевода или сегментации контекст меняется: фильтр должен тестироваться не только на исходных строках, но и на реальных преобразованиях. Отчёт описывает известные условия каждого теста, статистическую неопределённость и сценарии, не охваченные проверкой. Тестовая среда не должна содержать реальные избыточные персональные данные; если привлекаются правомерно полученные реальные материалы, их обработка отдельно обосновывается.

Независимый специалист может повторить вычисления на предоставленном confusion matrix и проверить соответствие выбранного контрольного набора заявленному использованию. Если предоставлены только итоговые маркетинговые показатели без исходных подсчётов, допустимый вывод ограничивается констатацией неполноты доказательств. Сопоставление качества фильтров полезно как инженерный контроль, но не заменяет проверки законного основания и цели данных операций.

Показатель	Формула	Что он не доказывает
Recall	TP/(TP+FN)	Законность источников
Precision	TP/(TP+FP)	Полноту выявления редких случаев
False negative rate	FN/(TP+FN)	Отсутствие ошибок вне теста
Разница recall	Recall B – recall A	Сами по себе причинность и значимость
Покрытие источников	Проверенные объекты/целевая совокупность	Полное соответствие GDPR

Приложение 49. Инцидент при раскрытии обучающего корпуса: соединение двух процедур

Отдельная проблема возникает, когда в обучающий корпус неправомерно попадают персональные сведения, а затем доступ к самому хранилищу получает неуполномоченная сторона. Эти события нельзя сливать в одно. Первое касается, в частности, законности, минимизации и прозрачности обработки; второе может составлять нарушение безопасности

персональных данных в смысле статьи 4(12) GDPR и в соответствующих условиях влечь оценку обязанности уведомления по статьям 33 и 34 [1].

Практический сценарий: подрядчик публикует диагностический файл с несколькими тысячами записей, в которых сохранились имена и данные о частной жизни. Организация должна немедленно защитить инфраструктуру, сохранить доказательства события, определить категории и число возможных субъектов, характер компрометации и возможные последствия. Параллельно нужно выяснить, почему эти данные вообще находились в training pipeline, было ли законное основание их извлечения и почему фильтры не удалили лишние поля.

Раздельная юридическая оценка важна для сроков. Уведомление о персональном data breach надзорному органу по статье 33 GDPR имеет собственный критерий риска и отсчёт, а публичный спор о законности scraping сам по себе не автоматически запускает тот же срок. При наличии риска для прав и свобод проводится анализ необходимости сообщения затронутым лицам по статье 34 и применимых исключений. Отчёт должен указывать дату осведомлённости, адресатов, решения и фактические доказательства, а не автоматически применять единый внутренний шаблон ко всем инцидентам.

В части исправления необходимо устранить возможность доступа, проверить связанные копии, оценить наличие фрагментов в действующей модели и изменить process design, из-за которого риск возник. Если организация лишь отзывает публичную ссылку на файл, но не пересматривает необоснованное хранение или training data lineage, коренная причина может сохраниться. Рекомендуется отдельная оценка системного изменения после устранения оперативной угрозы.

Независимый аудит должен соблюдать конфиденциальность доказательств, не публиковать сам скомпрометированный массив и не расширять число лиц, получивших доступ. При выводе формулируются отдельно: установленный технический инцидент, законность исходной обработки и качество ремедиации. Это предотвращает ситуацию, когда проблема государственной отчётности скрывает отдельный структурный дефект управления источниками.

Направление	Предмет	Доказательство
Инцидент безопасности	Несанкционированный доступ	Логи, временная шкала, scope
Исходный сбор	Законность попадания в корпус	LIA, source register
Хранение	Минимизация и сроки	Retention records
Уведомления	Риск, адресаты и сроки	Решения по ст. 33/34
Системное исправление	Предотвращение повторения	Проверка pipeline и доступа

Приложение 50. Учебное дело: запрос об удалении при совпадении имён

Генеративная модель выдаёт утверждение о якобы существующем уголовном деле против гражданина с распространённым именем и фамилией. Заявитель направляет требование удалить сведения. В интернете действительно существует старое сообщение о другом лице с такими же именем и фамилией. При рассмотрении нельзя принять простую логику «запись

найдена, значит она о заявителе», как нельзя и полностью отказаться от рассмотрения из-за сложности идентификации.

Первый этап — безопасно установить, что именно было сгенерировано, в какой конфигурации модели и при каких параметрах. Сохраняется минимальный фрагмент, необходимый для оценки, без распространения полного личного досье. Второй этап — отделить возможный первичный источник от правовой квалификации лица. Если сведения относятся к уголовным правонарушениям, анализируется статья 10 GDPR и применимые дополнительные условия. Третий этап — найти возможные внутренние training/retrieval связи, если доступ разрешён и технически возможен.

Для проверки личности организация не должна собирать лишние документы у заявителя или раскрывать ему сведения другого лица. Может потребоваться строго ограниченная проверка отдельных атрибутов, но она должна быть пропорциональна и документирована. Если совпадение подтверждено как ошибочное, решение о коррекции включает удаление соответствующей связи из доступных индексов, запрет на повторный сбор спорной записи в контексте заявителя и меры по снижению повторения неверного вывода.

Отдельный правовой вопрос — возможная обработка данных другого, действительно упомянутого лица. Исправление одной ложной атрибуции не делает автоматически законным обучение на исходном криминальном материале. Для первичного документа и каждой будущей обработки проводится независимая оценка. Эту проблему нельзя решить перестановкой фамилий в общедоступном ответе.

Завершение обращения предусматривает мотивированное письмо с указанием осуществлённых действий и их ограничений. Если установлено лишь изменение фильтра выдачи, не следует писать «мы полностью удалили данные из модели». Если техническое происхождение остаётся неизвестным, необходимо сообщить об ограничении в допустимой форме и продолжить соответствующую проверку. Сценарий подчёркивает, что защита данных человека — это юридическая процедура с цифровыми доказательствами, а не быстрый поиск похожей строки.

Сложность	Правильная мера
Однофамилец	Не отождествлять по имени
Уголовный контекст	Отдельная ст. 10 GDPR
Нет полного происхождения	Обозначить предел установления
Есть модельная выдача	Сохранить ограниченный воспроизводимый тест
Исправлена только выдача	Не обещать изменение всех весов

При совпадении имени необходимо различить подтверждение личности и идентификацию данных в информационной системе. Заявитель может подтвердить собственную личность, но это не доказывает, что спорная запись относится к нему. Наоборот, совпадение имени в векторном поиске может вернуть данные однофамильца, и необдуманный ответ на запрос о доступе раскроет третьему лицу чужую чувствительную информацию. Поэтому процедура использует достаточные дополнительные критерии только там, где это законно и необходимо, и не поощряет сбор полного комплекта удостоверений ради простого исправления публичного факта.

Учебное дело должно включать четыре исхода. Первый: данные действительно относятся к заявителю, их дальнейшая обработка признана недопустимой, и подтверждены конкретные меры исправления. Второй: запись принадлежит иному лицу, а вывод модели возник из ложного совпадения; корректируется процедура сопоставления, без раскрытия данных настоящего субъекта. Третий: сведения не обнаружены в доступных системах, но отсутствие в источниках не исключает внешних копий; это честно разъясняется. Четвёртый: доказательств недостаточно, и требуется ограниченная дополнительная проверка. Каждый итог сопровождается указанием, что было исследовано, какие версии систем охвачены и можно ли подать повторное обращение.

Источники и нормативно-методические материалы

По состоянию на 22 октября 2026 года. Проекты руководящих документов EDPB отмечены отдельно; приведённые источники не являются свидетельством проведения собственных полевых проверок.

[1] Европейский парламент и Совет ЕС. Регламент (EU) 2016/679 (GDPR), действующий текст; статьи 5, 6, 9, 10, 12–22, 25, 28, 32, 35, 36 и глава V. Официальный источник

[2] EDPB. Guidelines 03/2026 on web scraping in the context of generative AI, version 1.0, приняты для публичной консультации 7 июля 2026 года (проект). Официальный источник

[3] EDPB. Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, 17 декабря 2024 года. Официальный источник

[4] EDPB. Guidelines 07/2020 on the concepts of controller and processor in the GDPR, окончательная редакция 7 июля 2021 года. Официальный источник

[5] EDPB. Guidelines 1/2024 on processing of personal data based on Article 6(1)(f) GDPR, версия для консультации; статус не равен окончательной редакции. Официальный источник

[6] CNIL. La base légale de l’intérêt légitime: mesures à prendre en cas de collecte des données par web scraping, 19 июня 2025 года. Официальный источник

[7] ICO. The lawful basis for web scraping to train generative AI models, ответ на консультационную серию. Официальный источник

[8] CNIL. AI system development: recommendations to comply with the GDPR, включая сбор данных методом web scraping. Официальный источник

[9] ICO. Purpose limitation in the generative AI lifecycle, ответ на консультационную серию. Официальный источник

[10] ICO. Engineering individual rights into generative AI models, ответ на консультационную серию. Официальный источник

[11] Европейский парламент и Совет ЕС. Регламент (EU) 2024/1689 (AI Act), ст. 53 в консолидированной редакции от 27 июля 2026 года. Официальный источник

[12] Европейская комиссия. General-purpose AI obligations under the AI Act; краткое изложение обязанностей и публичной сводки обучающих данных. Официальный источник

[13] EDPB. Public consultation: Guidelines 03/2026 on web scraping in the context of generative AI, 8 июля – 30 октября 2026 года. Официальный источник

[14] EDPB/WP29. Guidelines on transparency under Regulation 2016/679 (WP260 rev.01), приняты WP29 и одобрены EDPB. Официальный источник

[15] EDPB. Guidelines 02/2026 on Anonymisation, version 1.0, для публичной консультации (проект). Официальный источник

[16] CNIL. Privacy Impact Assessment (PIA): методика и материалы для DPIA. Официальный источник

[17] Суд ЕС. Дело C-136/17, GC and Others v CNIL, решение Большой палаты от 24 сентября 2019 года, полный текст о правовых пределах деиндексации чувствительных данных. Официальный источник

[18] Европейский парламент и Совет ЕС. Директива (EU) 2019/790 о праве на авторское право в цифровом едином рынке, статьи 3 и 4. Официальный источник

[19] Европейский парламент и Совет ЕС. Регламент (EU) 2024/1689 (AI Act), статья 111(3): переходный режим для поставщиков GPAI-моделей, размещённых на рынке до 2 августа 2025 года. Официальный источник