SEARCHINFORM
INFORMATION SECURITY

Как защититься от утечек через ИИ?

СОДЕРЖАНИЕ

Зачем контролировать коммуникацию сотрудников с ИИ-сервисами? ... 3
Как помогает «СёрчИнформ КИБ» ... 4
Как настроить контроль ИИ-сервисов: практический разбор ... 5
Задача 1. Зафиксировать общение ... 5
Задача 2. Проконтролировать запросы ... 8
Задача 3. Заблокировать запросы ... 10
Полный запрет доступа ... 11
Запрет отправки файлов ... 13
Запрет запросов по контенту ... 14
Что в итоге? ... 17

SEARCHINFORM
INFORMATION SECURITY

Зачем контролировать коммуникацию сотрудников с ИИ-сервисами?

Каждый третий россиянин использует ИИ в работе и повседневной жизни. При этом данные, загружаемые в публичные нейросети, могут быть скомпрометированы.

Во-первых, данные чатов остаются в памяти ИИ-сервисов и могут быть использованы для обучения моделей. Например, в 2023 году Amazon рекомендовала сотрудникам не общаться с ИИ по рабочим вопросам, после того как обнаружила в ответах чат-бота данные, похожие на закрытые внутренние отчеты.

Во-вторых, существует риск публичной компрометации чатов. Например, пользователи ChatGPT могут делиться личными диалогами, создавая публичные ссылки. Если при этом активирована опция «Сделать чат доступным для поиска», конфиденциальная информация может быть проиндексирована поисковыми системами.

По данным опроса «СёрчИнформ», в 2025 году в 49% российских компаний ИБ-специалисты видели риски утечки конфиденциальной информации через ИИ-сервисы, и ограничивали персоналу доступ к ним. 6% компаний уже столкнулись с фактическими утечками информации через ИИ-сервисы.

«СёрчИнформ КИБ» сможет проконтролировать, используют ли сотрудники ИИ в обход корпоративной политики и не сливают ли туда конфиденциальные данные.

Как помогает «СёрчИнформ КИБ»

1. **Фиксирует использование ИИ.**

Сотрудники могут использовать ИИ, чтобы облегчить рабочие задачи. Важно контролировать эту активность с точки зрения легитимности, продуктивности и безопасности – ведь общение с нейросетями может привести к компрометации корпоративных данных и инфраструктуры. Кроме того, в поиске удобного ИИ-сервиса пользователи могут наткнуться на фишинговые подделки: «бесплатные» сервисы, которые на деле заражают ПК или крадут данные. Поэтому важно контролировать в том числе факты поиска ИИ-инструментов.

2. **Контролирует, какие запросы сотрудники отправляют в ИИ-сервисы.**

Пользователи могут отдавать в чаты с нейросетями критичную информацию: персональные данные, части кода, конфиденциальные документы. Необходимо контролировать контент, который сотрудники отдают на обработку в ИИ-системы, в любом виде: внутри текстовых промптов или файлов.

3. **Блокирует ненужные варианты коммуникации с ИИ.**

Если организация оперирует крайне чувствительной информацией, нужно исключить риски ее компрометации с помощью ИИ. Можно полностью запретить доступ к ИИ-ресурсам или ограничить определенные действия сотрудников на них: например, любую загрузку файлов.

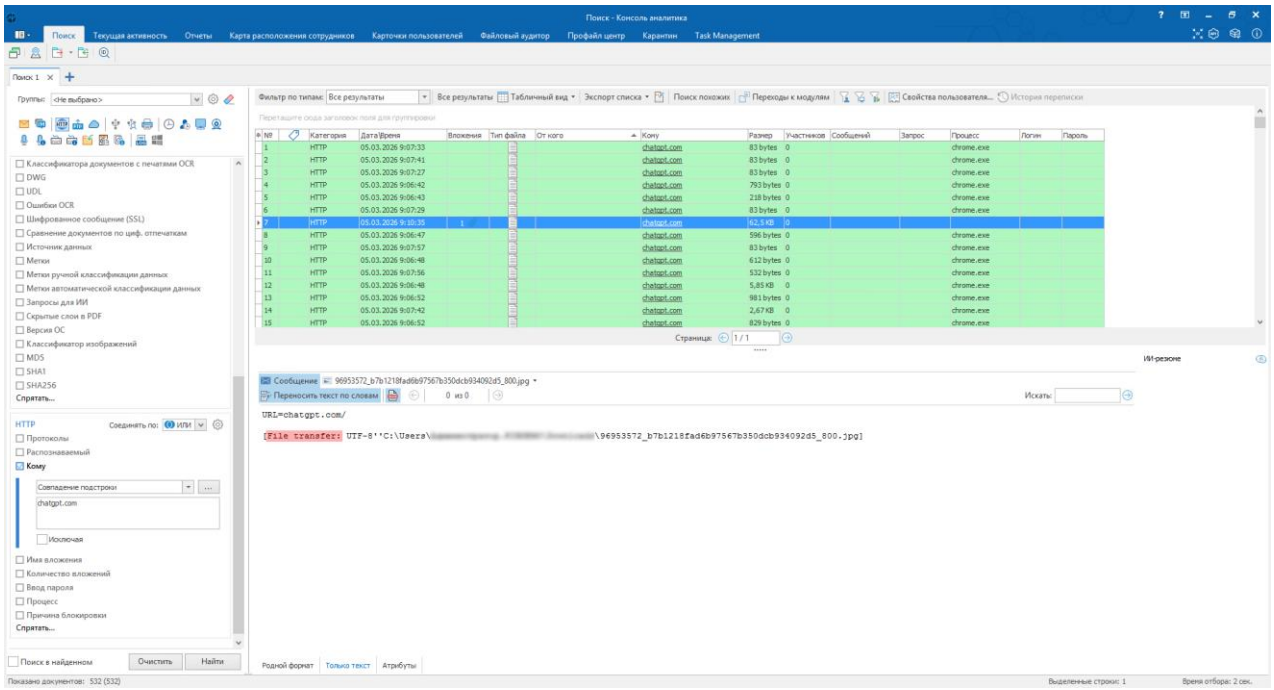
Как настроить контроль ИИ-сервисов: практический разбор

Задача 1. Зафиксировать общение

На первом этапе можно воспользоваться ручным поиском. Посмотреть, какие сайты посещали пользователи в течение дня, можно на вкладке Поиск в Консоли аналитика. Есть несколько способов это сделать.

- **источник данных** HTTPController отражает все обращения пользователей к любым веб-ресурсам. Чтобы найти посещения конкретных сайтов, задайте параметры поиска: например, в атрибуте «Кому» укажите названия известных ИИ-сервисов. Дополнительно уточнить запрос помогут фильтры и условия: по пользователю, дате, времени и т.д.
  Программа выводит результаты в виде таблицы, где указаны название ресурса и время отправки информации. Детали каждого визита представлены как пакеты трафика и фиксируются столько раз, сколько пользовать отправлял какие-либо данные на веб-ресурс.

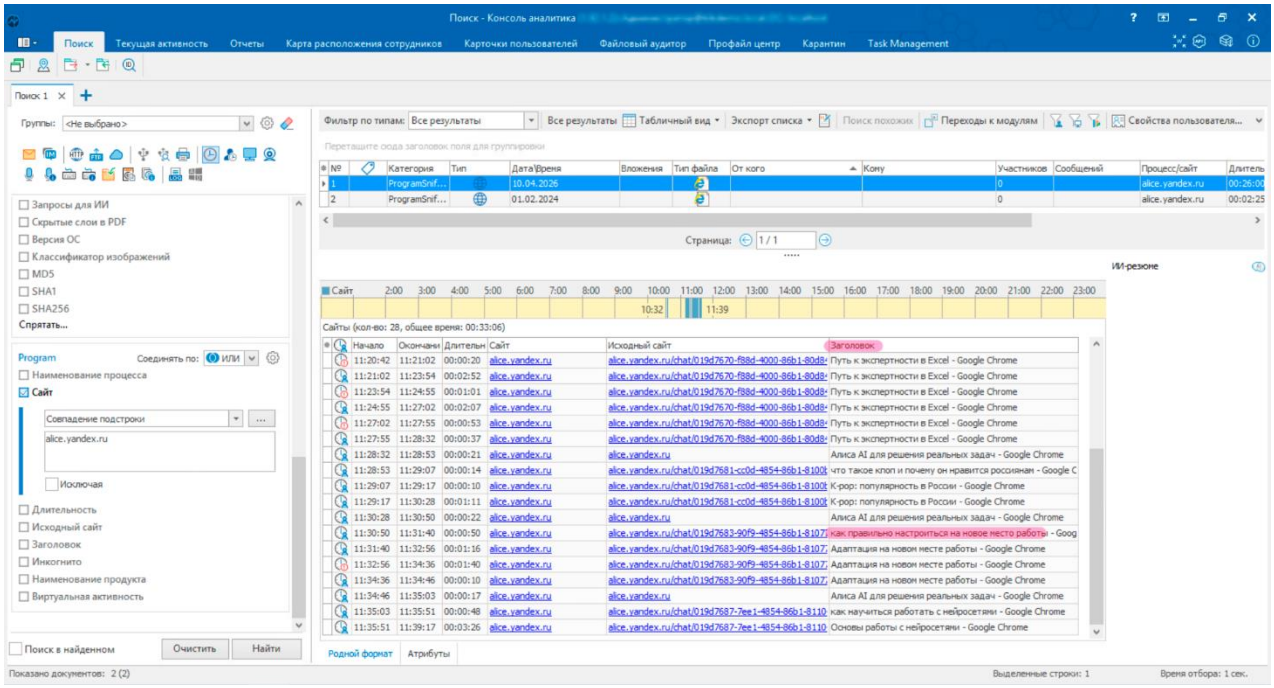
В большинстве случаев пакеты трафика содержат *текст пользовательского запроса*, факт *отправки файла* (документа, картинки или голосового сообщения), иногда – *содержимое отправленного файла*. В перехвате HTTPController с помощью фильтров можно найти нужные детали внутри пакетов и работать сразу с отобранными данными.



Результаты поиска по источнику HTTP: посещение ChatGPT

- **источник данных** ProgramController показывает все процессы и сайты, в которых был активен сотрудник. Чтобы сузить выборку, в панели фильтров выберите атрибут «Сайты» и задайте перечень известных доменов ИИ-сервисов, посещение которых хотите контролировать. Используйте метод «Совпадение подстроки»: тогда КИБ покажет все визиты на выбранные сайты, даже если во время визита URL у пользователя меняется – например, дополняется адресом поддомена.
  Программа покажет все случаи, когда пользователь посещал сайт ИИ-сервиса: с датой и временем, длительностью визита и т.д.

Важно, что ProgramController фиксирует *заголовок окна* при посещении сайта. Часто в нем можно увидеть *текст – или суть – пользовательского запроса* к ИИ. Это поможет разобраться, о чем именно он общался с нейросетью.



Результаты поиска по источнику Program: посещение Яндекс.Алиса

Чтобы каждый раз не искать обращения к нейросетям вручную, создайте политику безопасности в AlertCenter – тогда КИБ будет запускать нужный поиск автоматически по расписанию. Для этого:

1. При создании критерия поиска выберите **«Поиск по базам данных»**.
2. В перечне атрибутов выберите ProgramController.
3. В атрибуте **«Сайт»** укажите домены ИИ-сервисов. Например:
   - chatgpt.com
   - claude.ai
   - x.ai
   - perplexity.ai
   - chat.deepseek.com
   - giga.chat
   - alice.yandex.ru
   - gemini.google.com

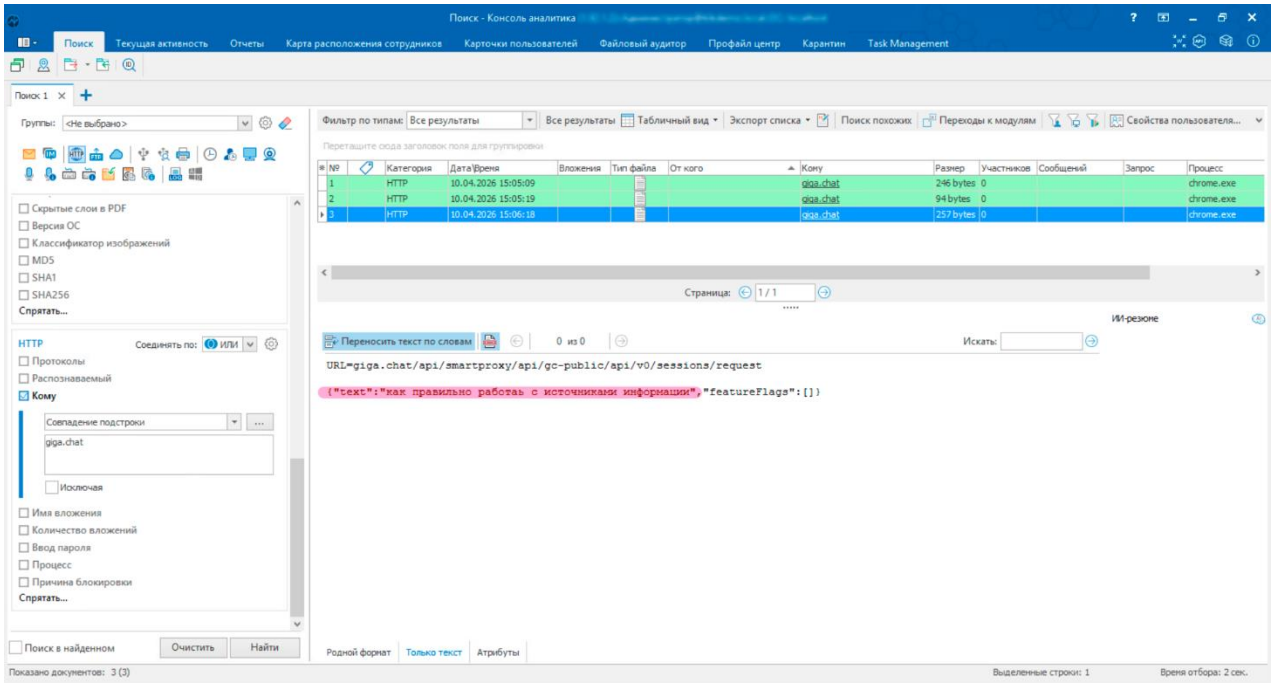
Можно одновременно указать несколько доменов. Лучше выбрать опцию «Совпадение подстроки». Не забудьте настроить расписание, чтобы сделать проверку по критерию регулярной.

Такая политика будет сообщать обо всех фактах обращений сотрудников к заданным нейросетям. Однако детали общения с ИИ не покажет –

чтобы контролировать, какие запросы шлют нейросетям пользователи, понадобится детализировать поиск.

Задача 2. Проконтролировать запросы

При работе с поиском по **HTTP** в **Консоли аналитика** обратите внимание на детали выдачи: каждый ИИ-сервис обрабатывает данные по-своему, поэтому результаты будут разными. Как правило, внутри зафиксированного пакета трафика окажется текст пользовательского запроса. Изучите, что его окружает: когда вопрос в чат с ИИ задается текстом, скорее всего, в пакете будет **специальное обозначение**. **Важно помнить: для каждой нейросети это обозначение будет уникальным.**



Пример обозначения передачи текстового запроса в пакете трафика

Поэтому при автоматизации контроля с помощью политик безопасности потребуется создавать отдельный критерий поиска для каждой нейросети.

Логика работы критериев при этом будет одинаковая.

1. При создании критерия выберите **«Сложный запрос»**.
2. Задайте условие 1 (A): **найти обращения к конкретным нейросетям.** Для этого:
   - выберите **«Поиск по атрибутам»**.
   - в перечне атрибутов выберите **HTTPController**.
   - в атрибуте **«Кому»** укажите домен ИИ-сервиса, запросы к которому хотите контролировать.
3. Задайте условие 2 (B): **найти, о чем у ИИ спрашивают пользователи.** Для этого:
   - выберите **«Фразовый поиск»,** вкладку **«Поиск по последовательности символов»**.
   - пропишите в блоке поиска необходимую строку – то самое «специальное обозначение», которое помечает в пакете трафика текст пользовательского запроса.

Примеры обозначений отправки текста:
- **для ChatGPT:** {"content_type":"text","parts":["
- **для GigaChat:** "text:

4. Задайте логику объединения условий: A **and** B
5. Задайте расписание проверки по критерию.

Получившийся критерий поиска будет отбирать только те данные перехвата, которые содержат непосредственно текст пользовательского промпта, а не служебную информацию. Его можно использовать как шаблон и на его основе уточнять запросы: например, искать общение с ИИ на конкретную тему (условием C по «Фразовому поиску» или «Поиску по словарю» на интересующую тематику) или загрузку туда документов (изменением условия B).

Отправка файлов в ИИ-сервисы как правило фиксируется на уровне пакета данных: например, в параметре "URL=" появится дополнение /upload/ или /attachment/, добавится команда [File transfer: ...]. Уточните обозначения в пакетах конкретных нейросетей при ручном поиске в Консоли аналитика.

HTTPController в КИБ перехватывает прикрепляемые файлы, их можно просмотреть в выдаче в Консоли Аналитика и при сработке политики в AlertCenter.

Это могут быть не только документы, но и картинки и даже голосовые запросы. Чтобы сразу работать с их содержимым, подключите в AAServer OCR для извлечения текста из изображений и ASR для распознавания речи. Если автоматическое распознавание не подходит, содержание файлов можно узнать вручную:

- для картинок – просмотреть снимки MonitorController,
- для аудио – проверить записи MicrophoneController.

Ищите по пользователю, процессу (браузерам) и времени, когда в HTTP зафиксировался запрос.

Задача 3. Заблокировать запросы

КИБ позволяет жестко запретить доступ к выбранным ИИ-сервисам или сделать это гибко: например, блокировать промпты с определенным содержанием или загрузку файлов в ИИ-сервисы.

Полный запрет доступа

Если в компании вообще запрещено использовать ИИ, настройте блокировку доступа к ИИ-сервисам по URL. Функция настраивается в EndpointController.

1. Зайдите в раздел «**Блокировки и перехват**» и выберите подраздел «**Сайты HTTP**».

2. Добавьте новое правило и выберите действие «**Запрещено**».

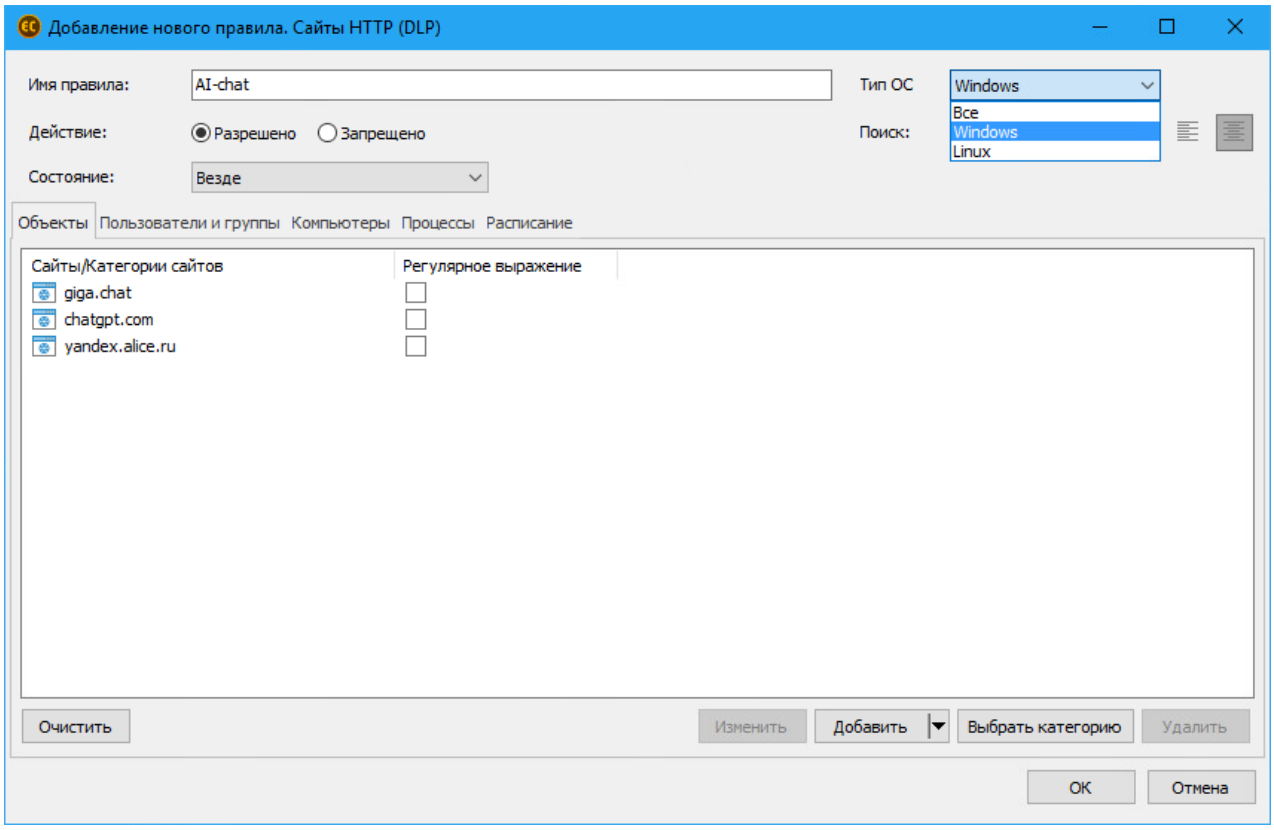
3. На вкладке «**Объекты**» пропишите список доменов ИИ-сервисов, к которым хотите заблокировать доступ.

4. Задайте дополнительные условия:
   - Перейдите на вкладку «**Пользователи**» внесите список сотрудников, для которых будет работать правило.
   - На вкладке «**Компьютеры**» можно указать список ПК, для которых будут действовать ограничения.

   Это удобно, если запрет должен работать выборочно.

5. Выберите **тип ОС**, для которых будет применяться блокировка. Правило может работать для пользовательских ПК на Windows, на Linux или везде сразу. Нюансы работы блокировки для разных ОС могут отличаться.

6. Уточните, в какое время будет работать блокировка на вкладке «**Расписание**». Например, можно разрешить использование ИИ в рамках корпоративной сети и запретить, если сотрудник работает удаленно или находится в командировке.



Настройка правила блокировки «Сайты HTTP»

Сотрудников можно оповещать о том, что использование ИИ-сервисов запрещено, или перенаправлять на разрешенные ресурсы при попытке доступа.

Чтобы настроить это в КИБ:

- **в подразделе блокировок «Сайты HTTP» перейдите в раздел «Оповещения»,**
- **выберите подходящий тип уведомления: *«не уведомлять»*, *«перенаправлять на сайт»* или *«выводить сообщение»*.**

Например, можно вывести сотруднику текстовое уведомление о запрете на использование ИИ-сервиса или отправку конфиденциальных данных, или перенаправить его на страницу с правилами корпоративной безопасности.

Полный запрет доступа подходит, если компания оперирует крайне чувствительной информацией. Однако стоит понимать, что тотальный отказ от ИИ-сервисов вызовет у сотрудников желание пользоваться ими тайно, чтобы, например, сэкономить время и быстрее выполнить работу.

Оптимальный вариант – прописать регламенты и настроить точечные блокировки по контенту или конкретным действиям. Так у сотрудников будет понимание, что использовать в работе искусственный интеллект можно (например, для генерации изображений в дизайне), но с оглядкой на правила безопасности.

Запрет отправки файлов

Если стоит задача пресечь утечки любых рабочих документов в ИИ-сервисы, используйте комплексную блокировку загрузки файлов в веб. Для этого:

1. В EndpointController зайдите в раздел «**Блокировки и перехват**» и выберите подраздел «**Передача файлов по HTTP**».
2. Добавьте новое правило и выберите действие «**Запрещено**».
3. На вкладке «**Объекты**» пропишите список доменов ИИ-сервисов, к которым хотите заблокировать доступ.
4. На вкладке «**Файлы**» укажите типы файлов (по имени или расширению), которые хотите запретить к отправке. Например, все файлы офисных форматов, все PDF, изображения PNG и пр.
5. Укажите дополнительные условия: задайте **пользователей** или **компьютеры**, для которых будет действовать блокировка, на соответствующих вкладках.

Правила блокировки «Передача файлов по HTTP» работают в браузерах: *Chrome, Opera, Brave, Яндекс* и *Edge*. Список расширяется. Пока функция доступна только для ПК пользователей под управлением Windows.

В результате будут заблокированы все попытки отправить файлы выбранных форматов в ИИ-сервисы. **Блокировку можно использовать и для запрета голосового общения с ИИ**, если нейросети конвертируют аудио-запросы пользователей в файлы перед отправкой на свои сервера. Уточните формат этих файлов при поиске по HTTP в Консоли Аналитика и укажите в правиле блокировки.

Запрет запросов по контенту

Самый гибкий вариант запретов – ограничить отправку в ИИ-сервисы промптов и документов только с критичным содержимым. Для этого подойдет контентная блокировка веб-трафика.

1. В EndpointController зайдите в раздел «**Блокировки и перехват**» и выберите подраздел «**Сайты по содержимому**».

2. Добавьте новое правило и выберите действие «**Запрещено**».

3. На вкладке «**Объекты**» пропишите группу сайтов: список доменов ИИ-сервисов, к которым хотите заблокировать доступ. Установите флажок, к какому действию должна применяться блокировка:
   - «**Отправка данных**» будет срабатывать на отправку текстовых промптов;
   - «**Отправка файлов**» будет работать для загружаемых в ИИ-сервис документов и картинок.

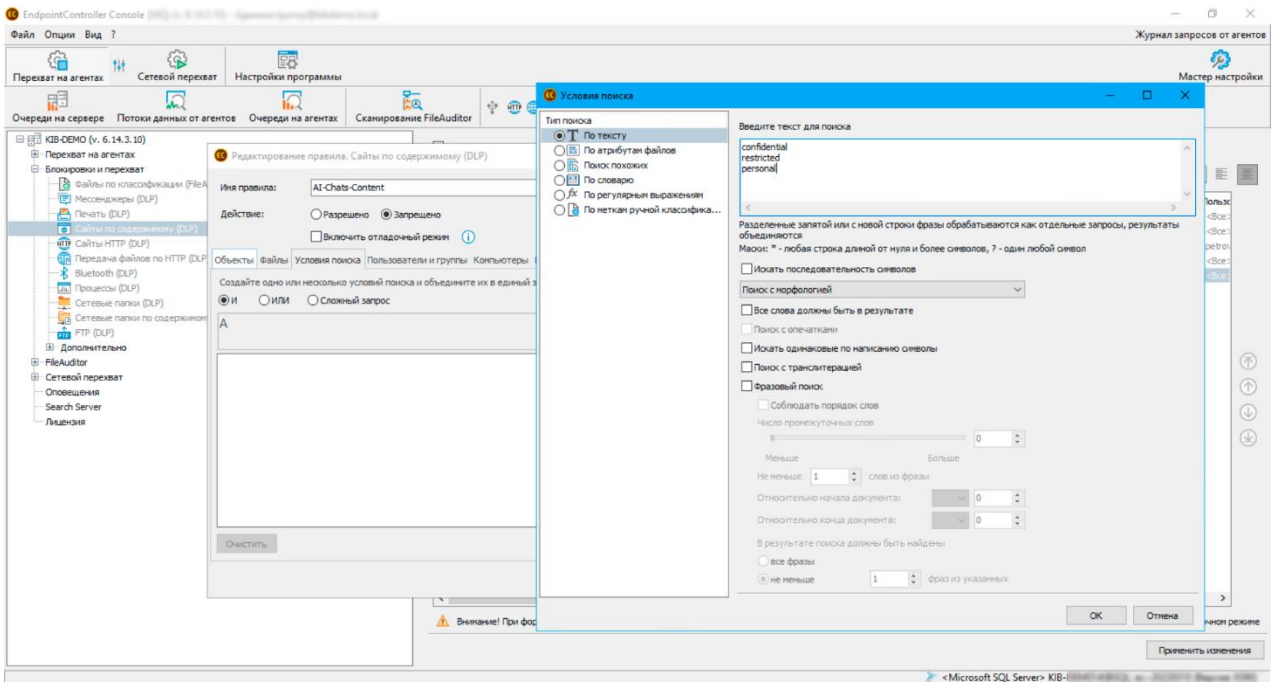
   Установите оба флажка, чтобы DLP проверяла контент и обычных запросов, и прикрепленных файлов.

4. На вкладке «**Условия поиска**» укажите, какой контент будет запрещен к отправке. Доступны поиски по тексту, словарям, регулярным выражениям, атрибутам и смысловой схожести, а также меткам классификации «СёрчИнформ FileAuditor» (для файлов).

5. При необходимости на вкладке **«Файлы»** укажите типы файлов, содержимое которых должно проверяться при отправке на ИИ-ресурс.

6. Включите **OCR** для распознавания контента в изображениях.

7. Задайте дополнительные условия:
    - **пользователей** или **компьютеры**, где будет действовать блокировка;
    - **процессы**, при работе которых будет действовать блокировка – например, конкретные браузеры.



Настройка условий поиска в правиле блокировки «Сайты по содержимому»

Блокировка будет действовать для пользовательских ПК на Windows и Linux. Нюансы работы блокировки для разных ОС могут отличаться. А в случае сработки запрета можно сообщать пользователям, что их действия нарушают политики безопасности: уведомления настраиваются в разделе «Оповещения».

В результате получится правило, которое позволит сотрудникам безопасно пользоваться ИИ-сервисами: доступ к ним останется открыт, но отправить туда запрос или документ с критичным для компании контентом не получится. Например, при попытке загрузить в нейросеть персональные данные или «запрещенный» документ, сотрудник увидит, что чат с ИИ «завис».

Важно помнить, что блокировка по содержимому не сработает, если какой-то ИИ-сервис использует нетипичные форматы передачи данных, фрагментирует передачу или шифрует данные вне спецификации HTTP(S).

Также блокировка не поможет остановить голосовые команды с конкретным содержимым – ASR-распознавание для «Сайтов по содержимому» недоступно.

Поэтому для оптимальной защиты стоит комбинировать разные варианты блокировок.

Что в итоге?

Часто сотрудники отправляют рабочие данные в ИИ, потому что не понимают: нейросеть – это «посторонний», и переданная информация останется у сервиса навсегда. Это ведет к недооценке рисков, ошибкам и непреднамеренным утечкам. С правовой точки зрения ИИ-сервисы однозначно признаются «третьей стороной», а значит, передача им информации может трактоваться как разглашение коммерческой тайны, компрометация персональных данных, трансграничная передача данных и т.д. Это чревато для организаций не только репутационными и финансовыми потерями, но иногда и наказанием от регуляторов.

Поэтому контроль использования ИИ-сервисов критически важен для обеспечения безопасности компании.

DLP-система **«СёрчИнформ КИБ»** помогает предотвратить утечки данных через ИИ-сервисы на всех уровнях. С ней ИБ-специалист сможет:

1. Узнать, кто в коллективе использует ИИ-сервисы.
2. Проверить, какими данными сотрудники делятся с нейросетями.
3. Исключить нежелательные действия сотрудников с ИИ.

Следите за нашей рассылкой и блогами: в следующих ИБ-инструкциях мы расскажем больше о возможностях DLP, DCAP и SIEM от «СёрчИнформ», а также поделимся лайфхаками, как решать практические задачи безопасности с помощью всего комплекса интегрированных инструментов.