

AI Security Gateway — Bedienungsanleitung

Das AISecurityGateway steht zwischen Ihren KI-Programmen und den KI-Anbietern. Jede Anfrage läuft hindurch: Persönliche Angaben werden vor dem Versand durch Platzhalter ersetzt und in der Antwort wieder eingesetzt. Der KI-Anbieter sieht die echten Werte nie; für den Anwender bleibt alles, wie es war.

Diese Anleitung hat zwei Teile:

- **Teil A** für die Mitarbeiter, die mit KI arbeiten.
- **Teil B** für die Verwaltung des Gateways im Admin-Panel.

Die Installation übernehmen wir; sie ist nicht Teil dieser Anleitung.

Teil A — Für Anwender

A1. Was sich für Sie ändert

Fast nichts. Sie arbeiten in Ihrem gewohnten Chat-Programm — etwa OpenWebUI oder AnythingLLM — mit den Modellen, die Ihr Betrieb freigegeben hat. Der Unterschied liegt im Hintergrund:

- Aus `max.mustermann@beispiel.de` wird auf dem Weg zur KI `[EMAIL_1]`, aus einer IBAN `[IBAN_1]`. In der Antwort steht wieder der echte Wert.
- Derselbe Wert bekommt im ganzen Gespräch denselben Platzhalter. Die KI kann sich also ganz normal auf „`[PERSON_1]`“ beziehen.
- Die Zuordnung zwischen Platzhalter und echtem Wert besteht nur für die Dauer der Anfrage und wird nirgends gespeichert.

A2. Zwei Stufen: Standard und @mask

Stufe	Wie wählen	Was ersetzt wird	Wofür
Standard	das Modell wie gewohnt	strukturierte Angaben: E-Mail-Adressen, Telefonnummern, IBAN, Kreditkarten, Steuer-ID, Zugangsschlüssel — dazu die Begriffe Ihres Betriebs (Projektcodes, Kundennamen)	die tägliche Arbeit
Vertraulich	die Modellvariante mit @mask im Modell-Menü, z. B. <code>gpt-4o@mask</code>	zusätzlich Namen von Personen und Firmen und Anschriften im Fließtext	Verträge, Bewerbungen, Schriftverkehr, Personalakten

Faustregel: Standard für den Alltag. **@mask** immer dann, wenn im Text Namen, Adressen oder andere persönliche Angaben in ganzen Sätzen stehen.

In der Stufe **@mask** prüft das Gateway auch die Sprache der Anfrage. Ist sie nicht sicher erkennbar oder nicht abgedeckt, wird die Anfrage **angehalten statt ungeprüft weitergeschickt**. Abgedeckt sind Deutsch, Englisch, Französisch, Spanisch, Italienisch und Russisch.

A3. Wenn eine Anfrage angehalten wird

Das Gateway bricht nichts ab. Statt einer Fehlermeldung bekommen Sie eine normale Antwort im Chat, die erklärt, was passiert ist. An die KI ist in diesem Fall nichts gegangen. Die häufigsten Gründe:

Hinweis im Chat	Was dahintersteckt	Was Sie tun
Ausweis oder Karte erkannt	Ein angehängtes Bild zeigt einen Ausweis, Pass, eine Gesundheits- oder Bankkarte. Diese werden immer gesperrt.	Das Dokument nicht an eine KI geben.
Anhang entfernt	Bilder und Dateien werden in der Grundeinstellung nicht mitgeschickt; der Text Ihrer Anfrage lief trotzdem weiter.	Den Inhalt als Text einfügen, wenn er dafür geeignet ist.
Audio oder Video gesperrt	Ton und Video lassen sich nicht auf persönliche Angaben prüfen.	Mit der Verwaltung klären, ob es dafür einen Weg gibt.
Sprache nicht abgedeckt	In der Stufe @mask ließ sich die Sprache nicht sicher bestimmen.	Die Anfrage in einer der abgedeckten Sprachen stellen.
Entwicklerwerkzeug gesperrt	Programmierungswerkzeuge wie Claude Code oder Cursor sind in der Grundeinstellung gesperrt.	Freigabe bei der Verwaltung anfragen.

A4. KI-Webseiten sind gesperrt

Seiten wie chatgpt.com oder claude.ai lassen sich im Betrieb nicht aufrufen. Das ist gewollt: Über den Browser liefe Ihre Eingabe am Filter vorbei. Statt der Seite erscheint ein Hinweis. Für KI nutzen Sie das Chat-Programm, das Ihr Betrieb eingerichtet hat.

A5. Was protokolliert wird — und was nicht

Das Gateway hält fest, **wann** der Filter lief, **für welchen Rechner** und **wie viele Angaben welcher Art** ersetzt wurden. **Niemals** die Werte selbst und niemals den Inhalt Ihrer Anfragen. Fügt jemand eine ungewöhnlich große Zahl persönlicher Angaben auf einmal ein — etwa eine ganze Kundenliste —, wird der Vorgang markiert. Blockiert wird dabei nichts.

A6. Was der Filter nicht kann

Der Filter findet, was er erkennt. Er ist sorgfältig, aber nicht unfehlbar — er ersetzt nicht Ihr Urteil darüber, was in eine KI gehört. Und er wirkt nur auf dem betrieblichen Weg: Was jemand abtippt und privat eingibt, erreicht ihn nicht.

Teil B — Für die Verwaltung: das Admin-Panel

B1. Anmelden

Das Panel öffnen Sie im Browser unter

```
http://<adresse-des-gateways>:8000/admin/login
```

mit dem Benutzer **admin** und dem Kennwort aus der Einrichtung. Das Kennwort ändern Sie in der Karte **Konto** (mindestens 8 Zeichen).

Links steht eine Abschnittsliste, über die Sie zu jeder Karte springen. Ganz oben die Karte **Schutz-Ebenen**: ein Schaubild, welche Stufen eine Anfrage gerade durchläuft — grün aktiv, gelb abgeschaltet oder ungefiltert.

Grundprinzip: Jede Karte ist ein Regler zwischen Schutz und Bequemlichkeit. Die Grundeinstellungen sind die vorsichtigen. Wer lockert, trifft eine bewusste Entscheidung über das Risiko seines Betriebs — und sollte sie festhalten.

B2. Status & Lizenz

Zeigt, ob der Filter läuft, welche Erkennung arbeitet und den Lizenzstand. **Ohne gültige Lizenz leitet das Gateway keine Anfragen weiter.**

- **Speichern & aktivieren** — Lizenzschlüssel eintragen und aktivieren.
- **Verbindung prüfen** — testet die Verbindung zum Lizenzserver.
- **Deaktivieren** — gibt den Platz frei, etwa beim Umzug auf einen anderen Rechner.

B3. Blacklist — Ihre eigenen Begriffe

Begriffe in dieser Karte werden **immer** ersetzt, in beiden Stufen und in jeder Sprache, ohne Rücksicht auf Groß- und Kleinschreibung. Gedacht für Projektcodes, Kunden- und Produktnamen, interne Kennungen. Ein Begriff pro Zeile, dann **Blacklist speichern**.

Pflegen Sie die Liste laufend. Ein neuer Großkunde, ein neues Projekt — was nicht darin steht, erkennt die Grundstufe nur, wenn es wie eine E-Mail-Adresse, Nummer oder Kennung aussieht.

B4. Medien — Bilder, Dokumente, Audio, Video

Das Gateway filtert Text. Anhänge werden nicht maskiert, sondern nach festen Regeln behandelt:

Einstellung	Grundeinstellung	Bedeutung
Ausweis-/Karten-Scan aktiv	an	Anhänge werden im Haus per Texterkennung gelesen; Ausweise, Pässe, Gesundheits- und Bankkarten werden immer gesperrt. Hat Vorrang vor allem anderen.
Bild/Dokument-Policy	strip	<i>strip</i> entfernt den Anhang, der Text läuft weiter. <i>block</i> lehnt die ganze Anfrage ab. <i>allow</i> reicht durch — ohne Schutz .
Bild-Uploads verbieten	an	Fotos erreichen die KI nicht. Abwählen erlaubt Bildauswertung durch die KI; das Ausweis-Veto bleibt aktiv.
Erweiterter Dokument-Scan	an	erkennt zusätzlich Rechnungen, Kontoauszüge, Lohnabrechnungen, Steuerbescheide, ärztliche Unterlagen und Verträge — wo Bilder durchgereicht werden
Audio/Video durchlassen	aus	eingeschaltet geht Ton und Video ungeprüft an die KI
Bildbearbeitung erlauben	aus	Bild hinein, verändertes Bild heraus; jedes hochgeladene Bild durchläuft das Ausweis-Veto

Unter **Aktuelle Wirkung** steht in Klartext, was die gewählten Schalter gerade bewirken. Unter **Heute** die Zahl der gesperrten und entfernten Anhänge.

B5. Entwickler — Programmierwerkzeuge

Werkzeuge wie Claude Code, Codex, Cursor oder Copilot übertragen ganze Quelldateien an die KI. Sie sind in der Grundeinstellung **gesperrt**. **Agentic-Coding-Tools zulassen** gibt sie frei; ihre Anfragen laufen dann weiter durch den Filter.

Die Anmeldung mit einem **persönlichen Abo** (etwa Claude Pro statt eines Betriebsschlüssels) wird **immer** abgelehnt — unabhängig von diesem Schalter. Ein solcher Zugang hat keinen Vertragspartner, keine Kostenzuordnung und lässt sich vom Betrieb nicht entziehen.

Ein freigegebenes Werkzeug muss auf das Gateway zeigen, sonst greift der Filter nicht:

Werkzeug	Einstellung
Claude Code	ANTHROPIC_BASE_URL=http://<gateway>:8000/anthropic und ein Betriebsschlüssel
Codex und OpenAI-kompatible Werkzeuge	OPENAI_BASE_URL=http://<gateway>:8000/v1
Cursor, Erweiterungen für Entwicklungsumgebungen	Feld „Base URL“ auf http://<gateway>:8000/v1

B6. Shadow-AI Shield — die Sperre der KI-Webseiten

Immer aktiv, nicht abschaltbar. Der Shield sperrt KI-Webseiten an der Namensauflösung, damit niemand am Filter vorbei im Browser arbeitet. Zusätzlich einschaltbar:

Schalter	Wirkung
KI-API-Hosts sperren	schließt den direkten Weg von Programmen zur KI-Schnittstelle. Programme, die über das Gateway laufen, bleiben nutzbar.
Agentic-Coding-Backends sperren	sperrt die eigenen Dienste von Copilot, Cursor-App, JetBrains AI und ähnlichen
KI-Übersetzer sperren	DeepL, Google Übersetzer und andere — erst einschalten, wenn es einen Ersatzweg gibt
DoH-Canary sperren	schaltet die verschlüsselte Namensauflösung der Browser zentral ab, damit die Sperre nicht umgangen wird

In **Gesperrte Domains & Ausnahmen** ergänzen Sie eigene Einträge — eine Domain pro Zeile, ohne https://. **Ausnahmen** überstimmen alle Listen. Diese Änderungen und **Eigene URL-Listen** wirken sofort; Änderungen an den mitgelieferten Listen und der Engine brauchen einen Neustart des Sperrdienstes.

B7. Audit — der Nachweis

Die Karte **Audit · heute** zeigt die Zahl der Anfragen und ersetzten Angaben, nach Art und nach Rechner. Gespeichert werden **nur Metadaten**, nie Werte oder Inhalte — das Protokoll eignet sich deshalb als Nachweis für Datenschutz-Dokumentation und Folgenabschätzung.

- **CSV** und **PDF** — Ausfuhr für die Ablage.
- **Archivieren & zurücksetzen** — legt das laufende Protokoll ins Archiv; der Nachweis bleibt erhalten.
- Ein **Massendaten-Verdacht** heißt: Eine Anfrage enthielt ungewöhnlich viele Personenbezüge auf einmal. Es wurde nichts blockiert und alles Erkannte normal ersetzt. Der Hinweis ist dazu da, den Vorgang mit dem Mitarbeiter zu klären.

Grenze des Nachweises: Das Protokoll belegt, dass der Filter lief und was er ersetzt hat. Es kann nicht belegen, dass nichts Unerkanntes durchging — dafür müsste es Inhalte mitschreiben, und genau das tut es bewusst nicht.

Client-Namen

Rechner erscheinen im Audit als `key_xxx` — ein nicht umkehrbarer Abdruck des verwendeten Schlüssels. In der Karte **Client-Namen** geben Sie ihnen lesbare Namen, eine Zuordnung pro Zeile: `key_xxx = pc-empfang`. Die Liste ist mit den heute gesehenen Kennungen vorbefüllt.

Kosten

Eine Schätzung der Kosten des Tages, gesamt und je Rechner. Grundlage sind die hinterlegten Modellpreise; die Abrechnung des Anbieters kann abweichen.

B8. Weitere Karten

Karte	Wofür
Sprach-Routing	welche Erkennung für welche Sprache arbeitet. Nicht abgedeckte Sprachen werden in der Stufe @mask angehalten.
Bildmodell	Anbieter, Modell und zentraler Schlüssel für die Bilderzeugung. Jede Bildanfrage durchläuft den Filter und das Audit.
Konfiguration · sichern & übertragen	alle Panel-Einstellungen als Datei, etwa um eine zweite Instanz gleichzuziehen. Kennwort und Lizenz sind nicht enthalten. Schlüssel der Bild-Anbieter nur, wenn ausdrücklich angehakt — sie stünden sonst im Klartext in der Datei.
Gateway-Adresse, Ausfallschutz	nur bei einer zweiten Instanz als Rückfall; die Einrichtung übernehmen wir

B9. Einen neuen Arbeitsplatz anbinden

Im Chat-Programm des Arbeitsplatzes das Gateway als Adresse der KI eintragen:

<code>http://<gateway>:8000/v1</code>	OpenAI-kompatibel (die meisten Programme)
<code>http://<gateway>:8000/anthropic</code>	für Anthropic-Clients
<code>http://<gateway>:8000/openai</code>	für OpenAI-Clients

Der echte API-Schlüssel wird vom Programm mitgeschickt; das Gateway reicht ihn durch und speichert ihn nicht. Soll der Rechner im Audit gleich mit Namen erscheinen, setzt das Programm den Header `X-ASG-Client` — oder Sie tragen ihn unter **Client-Namen** ein.

Was Sie selbst im Blick behalten sollten

- **Die Blacklist pflegen.** Neue Kunden und Projekte gehören hinein.
- **Das Audit regelmäßig ansehen** — besonders Massendaten-Hinweise.
- **Lockerungen festhalten.** Wer Bild-Upload, Audio/Video oder Entwicklerwerkzeuge freigibt, entscheidet über ein Risiko. Halten Sie fest, wer das wann warum entschieden hat.
- **Die Lizenz** — ohne sie leitet das Gateway nichts weiter.

RB-COMPUTERSERVICE · Roman Brunetti · Elbestraße 8c, 90453 Nürnberg · 0911 6320968 · info@rb-computerservice.de

RB-COMPUTERSERVICE · Elbestraße 8c, 90453 Nürnberg · 0911 6320968 · info@rb-computerservice.de · rb-computerservice.de