



loopthink

THE SOVEREIGN AI CONTROL PLANE

TECHNISCHES WHITEPAPER

MCP-Zugriffe auf Unternehmenssysteme steuern

Gateway-Architektur, Maskierung auf Feldebene und Deployment-
Topologien — für Architekten, Security-Engineers und CISOs, die
diese Produktklasse bewerten.

ZUSAMMENFASSUNG

Das Model Context Protocol hat es trivial gemacht, einen universellen KI-Assistenten mit einem internen System zu verbinden. Eine funktionierende Anbindung an ein ERP oder eine Produktionsdatenbank ist heute ein Nachmittag Arbeit statt eines Projekts. Diese Bequemlichkeit hat die Kontrollen darum herum überholt: In den meisten Organisationen kann niemand beantworten, wie viele solcher Verbindungen aktuell existieren, unter welchen Identitäten sie laufen und welche Daten die Grenze bereits überschritten haben.

Dieses Papier beschreibt, was eine steuernde Schicht zwischen KI-Clients und internen Systemen leisten muss, wie Maskierung auf Feldebene funktioniert und wo ihre Grenzen liegen, und wie sich drei Deployment-Topologien in der einen konkreten Frage unterscheiden, die für einen Security-Architekten zählt: Was verlässt das Netzwerk tatsächlich, und wer hält die Zugangsdaten?

Es richtet sich an Architekten, Security-Engineers und CISOs, die diese Produktklasse bewerten — und ist bewusst deutlich darin, was ein Gateway *nicht* löst.

1 Warum sich der Kontrollpunkt verschoben hat

Vor MCP bedeutete die Integration eines KI-Assistenten mit einem internen System, einen eigenen Dienst zu bauen: Authentifizierung, API-Wrapper, Fehlerbehandlung, Deployment. Die Kosten dieser Arbeit waren selbst eine Kontrolle. Projekte durchliefen das Architektur-Review, weil sie durch die Entwicklung mussten.

MCP beseitigt diese Kosten. Ein Entwickler kann eine interne Datenbank an einem Nachmittag als Satz aufrufbarer Tools verfügbar machen, und jeder MCP-fähige Client — ChatGPT, Claude, Copilot, Cursor, eine selbst gehostete Oberfläche — kann sie konsumieren. Das Protokoll wird inzwischen unter der Agentic AI Foundation weiterentwickelt, mit den großen Modellanbietern und mehreren Enterprise-Software-Herstellern als Beteiligten; es ist also keine Wette auf die Roadmap eines einzelnen Anbieters. Es entwickelt sich zur Standard-Infrastruktur.

Die Konsequenz für die Sicherheitsarchitektur ist konkret. **Die Autorisierung ist von der Integration in die Sitzung gewandert.** Im alten Modell hatte ein Service-Account einen festen Umfang — zur Bauzeit definiert, einmal geprüft. Im MCP-Modell verbindet eine menschliche Identität einen Client mit einem Server und stellt dann offene, natürlichsprachliche Anfragen gegen alle Tools, die dieser Server anbietet. Was der Nutzer tun kann, wird zum Verbindungszeitpunkt bestimmt, pro Nutzer, und ändert sich mit seinen Berechtigungen.

Steht nichts zwischen Client und System, sind drei Dinge gleichzeitig wahr:

- Das interne System sieht gültige Zugangsdaten und kann einen freigegebenen Agenten nicht vom Laptop eines Entwicklers oder einem privaten ChatGPT-Konto unterscheiden.
- Die Details der Tool-Aufrufe — welches Tool, welche Parameter, welche Identität, welches System, zu welcher Zeit — werden nirgendwo in der Form erfasst, die ein Audit oder eine Incident-Untersuchung verlangt. Die Konversationsprotokolle des Modellanbieters enthalten sie nicht.
- Nutzdaten, einschließlich personenbezogener Daten, wandern zum Modellanbieter — in genau der Form, in der das Tool sie zurückgegeben hat.

Ein Gateway existiert, um jede dieser drei Aussagen falsch zu machen.

2 Was ein steuerndes Gateway leisten muss

Fünf Funktionen, in der Reihenfolge, in der eine Anfrage auf sie trifft.

Identitätsbindung. Die MCP-Sitzung muss an einen echten, benannten Nutzer aus dem bestehenden Verzeichnis der Organisation gebunden sein, nicht an einen geteilten API-Key. Praktisch heißt das: Das Gateway agiert als OAuth-Resource-Server und authentifiziert die Sitzung gegen den Identity Provider des Kunden. Jede nachgelagerte Aktion erbt diese Identität.

Tool-Kuratierung und Scoping. Nicht jeder Nutzer sollte jedes Tool sehen. Das Gateway löst die authentifizierte Identität gegen ein Rollenmodell auf und liefert nur die Tools zurück, die diese Identität aufrufen darf. Ein Finanzcontroller und ein Support-Mitarbeiter, die sich mit demselben Endpunkt verbinden, erhalten unterschiedliche Tool-Listen.

Policy-Durchsetzung zum Aufrufzeitpunkt. Die Liste einzuschränken genügt nicht, denn Tool-Argumente tragen ihr eigenes Risiko. Ein Read-only-Tool mit einem unbegrenzten Query-Parameter ist in keinem relevanten Sinn read-only. Die Policy muss den tatsächlichen Aufruf bewerten: welches Tool, welche Argumente, welches Ziel, unter welcher Rolle.

Transformation in der Datenebene. Ergebnisse durchlaufen Maskierung und DLP, bevor sie Richtung Client zurückgegeben werden. Das ist Gegenstand von Abschnitt 4 — und die Funktion mit den meisten Feinheiten.

Audit. Jede Sitzung, jeder Tool-Aufruf, jede Ergebnisgröße, jede Policy-Entscheidung — aufgezeichnet in einer Form, die ein Audit übersteht und in ein SIEM exportiert werden kann.

ANMERKUNG ZUR PERSONALISIERUNG DER TOOL-LISTE

MCP unterstützt sowohl die initiale `tools/list`-Anfrage als auch eine `notifications/tools/list_changed`-Nachricht, um einen Client während der Sitzung zu aktualisieren. In der Praxis ist die Client-Unterstützung für diese Notification uneinheitlich: Manche Clients respektieren sie, manche cachen die Tool-Liste für die Lebensdauer der Verbindung, manche aktualisieren erst beim Reconnect.

Die architektonische Implikation verdient eine klare Formulierung, denn sie betrifft die Frage, wie sich Berechtigungsänderungen fortpflanzen. **Behandeln Sie die Discovery zum Verbindungszeitpunkt als den verlässlichen Mechanismus und Aktualisierungen während der Sitzung als Optimierung.** Entwerfen Sie kein Autorisierungsmodell, das darauf angewiesen ist, dass ein Entzug innerhalb einer aktiven Sitzung wirksam wird. Der Entzug muss zum Aufrufzeitpunkt von der Policy-Engine durchgesetzt werden — nicht dadurch, dass ein Tool aus einer Liste entfernt wird, die der Client womöglich noch hält.

3 Der Outbound-only-Connector

Der übliche Weg, einem externen Dienst Zugriff auf ein internes System zu geben, ist ein exponierter Endpunkt: ein Reverse Proxy, ein VPN-Tunnel, ein freigeschalteter eingehender Port. Jede dieser Varianten ver-

ändert den Netzwerkperimeter, jede braucht ein Firewall-Ticket, jede schafft eine lauschende Angriffsfläche.

Die Alternative ist, die Richtung umzukehren. Eine Connector-Komponente läuft im Netzwerk des Kunden und baut eine ausgehende, authentifizierte Verbindung zur Control Plane auf. Sie fragt genehmigte Aufträge ab, führt sie lokal gegen das interne System aus und liefert die Ergebnisse über denselben verschlüsselten Kanal zurück.

Drei Konsequenzen folgen daraus, und sie sind die Substanz des Designs:

- 1 **Keine eingehenden Ports, kein VPN, keine Firewall-Änderung.** Der Connector nutzt denselben Egress-Pfad wie jeder andere ausgehende HTTPS-Client. Das ist in der Regel der Unterschied zwischen einem zweiwöchigen Network-Change-Request und einem Deployment am selben Tag.
- 2 **Zugangsdaten verlassen das Netzwerk nie.** Der SAP-Service-User, das Datenbankpasswort, der API-Key eines internen Dienstes — sie werden im Connector konfiguriert und in der Umgebung des Kunden gespeichert. Die Control Plane hält eine Referenz auf ein Credential, nie das Credential selbst.
- 3 **Die Ausführung erfolgt lokal.** Die Abfrage läuft dort, wo die Daten liegen. Über die Grenze geht das Ergebnis — nach der Transformation —, nicht der Rohdatensatz.

Der Preis dafür sind Latenz und Betriebsaufwand: Der Connector ist eine Komponente, die der Kunde nun betreibt, überwacht und aktualisiert. Diese Kosten sind real und gehören Käufern gegenüber offen benannt statt beschönigt.

4 Maskierung auf Feldebene

4.1 Wo maskiert wird

Die wichtigste Eigenschaft überhaupt ist die Position im Datenfluss. Maskierung ist nur dann von Bedeutung, wenn sie **innerhalb der Datenebene angewendet wird, bevor das Ergebnis das Kundennetzwerk verlässt** — also im Connector, nicht in der Cloud.

Maskierung, die erst greift, nachdem die Daten eine Cloud-Komponente erreicht haben, schützt nichts, worauf es ankam. Die unmaskierten Daten haben die Grenze bereits überschritten; die Transformation ist Kosmetik. Jede Architektur, die Maskierung als Souveränitätskontrolle beansprucht, muss auf die exakte Komponente zeigen können, in der die Transformation läuft — und belegen, dass sie auf der Kundenseite der Grenze sitzt.

4.2 Konfigurationsmodell

Maskierung wird pro Importer und pro Feld konfiguriert. Ein Importer ist eine konfigurierte Datenquelle — eine HTTP-API oder eine Datenbankverbindung (PostgreSQL, MySQL/MariaDB, MS SQL Server, Oracle, MongoDB). Für jeden Importer werden die zurückgegebenen Felder aufgelistet und pro Feld eine Maskierungsregel zugewiesen.

Regel	Verhalten	Einsatz
Redigieren	Ersetzen durch ein festes Token	Felder, die das Modell nie braucht (Passwörter, interne Schlüssel)
Pseudonymisieren	Ersetzen durch ein stabiles Surrogat, Mapping wird lokal vorgehalten	Identifikatoren, die das Modell über Datensätze hinweg korrelieren muss
Partiell	Präfix oder Suffix bleibt erhalten	Kontonummern, IBANs, wenn das Institut relevant ist
Generalisieren	Präzision reduzieren	Geburtsdaten auf das Jahr, Postleitzahlen auf die Region
Durchlassen	Keine Transformation	Nicht personenbezogene Geschäftsdaten

Der Standard für einen neu konfigurierten Importer sollte **Deny-by-Default für unklassifizierte Felder** sein, nicht Durchlassen. Ein Feld, das niemand klassifiziert hat, ist ein Feld, das niemand bewertet hat.

4.3 Der Rückweg

Pseudonymisierung ist nur nützlich, wenn die Antwort in die reale Welt zurückgeholt werden kann. Wer nach überfälligen Rechnungen fragt, braucht in der finalen Antwort den echten Kundennamen, nicht `CUSTOMER_7F3A`.

Die Mapping-Tabelle liegt deshalb im Connector, innerhalb des Netzwerks. Ausgehend werden echte Werte durch Surrogate ersetzt. Eingehend, wenn die Antwort des Modells ein Surrogat referenziert, setzt der Connector den echten Wert ein, bevor er den Bildschirm des Nutzers erreicht. Das Modell denkt in Surrogata; der Nutzer sieht die Realität; das Mapping überquert die Grenze nie.

Für strukturierte Felder funktioniert das sauber. Für Freitext funktioniert es schlecht — der nächste Punkt.

4.4 Was Maskierung nicht leistet

Dieser Abschnitt existiert, weil die Glaubwürdigkeit der gesamten Architektur davon abhängt, hier ehrlich zu sein.

Maskierte Daten sind in der Regel weiterhin personenbezogene Daten. Unter der DSGVO ist Pseudonymisierung eine Sicherheitsmaßnahme, keine Ausnahme — pseudonymisierte Daten bleiben im Anwendungsbereich der Verordnung. Nur echte Anonymisierung, bei der eine Re-Identifizierung für niemanden mit vernünftigem Aufwand möglich ist, fällt heraus. Weil das Mapping bewusst im Connector vorgehalten wird, erzeugt diese Architektur pseudonymisierte Daten, keine anonymisierten. Jede Behauptung, Maskierung allein mache eine Verarbeitung datenschutzkonform, ist falsch — und wird von jedem kompetenten Datenschutzbeauftragten als falsch erkannt.

Freitext hebt Feldregeln aus. Eine Regel auf einer `customer_name`-Spalte erfasst denselben Namen nicht, wenn er in einem `notes`- oder `description`-Feld auftaucht. Mustererkennung über Freitext findet gängige Formate — E-Mail-Adressen, IBANs, Telefonnummern —, aber sie ist probabilistisch, und die Trefferquote

ist nie vollständig. Wo Freitextfelder personenbezogene Daten tragen, sind die ehrlichen Optionen: das Feld ausschließen oder ein Restrisiko explizit akzeptieren.

Kombination re-identifiziert. Einzelne harmlose Felder können eine Person gemeinsam identifizieren. Den Namen zu maskieren und zugleich Abteilung, Rolle und Eintrittsdatum zurückzugeben, schützt in einem Sechs-Personen-Team womöglich niemanden.

Nutzen und Schutz stehen in direktem Zielkonflikt. Ein Feld zu maskieren, das das Modell zur Beantwortung der Frage braucht, verschlechtert die Antwort. Das ist kein Tuning-Problem, das sich lösen lässt; es ist eine Grenze, die bewusst gezogen werden muss — pro Anwendungsfall, mit dem Fachverantwortlichen am Tisch.

Maskierung ist keine Zugriffskontrolle. Sie reduziert die Sensibilität der Daten, die ein erlaubter Aufruf zurückliefert. Gegen einen Aufruf, der nie hätte erlaubt sein dürfen, richtet sie nichts aus. Das ist die Aufgabe der Policy-Engine, und die beiden dürfen in einem Design-Review nicht verwechselt werden.

5 Deployment-Topologien

Drei Konfigurationen, die sich darin unterscheiden, wo die Control Plane läuft. Der Connector ist in allen dreien vorhanden.

Topologie A — Managed Control Plane

Die Control Plane läuft als verwalteter, in der EU gehosteter Dienst. Der Connector läuft im Netzwerk des Kunden.

- **Überquert die Grenze:** transformierte Ergebnisse, Tool-Schemata, Policy-Entscheidungen, Nutzungsmetadaten
- **Überquert sie nie:** Systemzugangsdaten, unmaskierte Nutzdaten, Pseudonym-Mappings
- **Der Kunde betreibt:** nur den Connector
- **Passt für:** die meisten Organisationen; der schnellste Weg in die Produktion

Topologie B — Self-hosted Control Plane

Die Control Plane läuft in der eigenen Cloud-Umgebung oder im Rechenzentrum des Kunden. Nichts läuft über die Infrastruktur des Anbieters.

- **Überquert die Grenze:** nichts Operatives; nur Software-Updates und Lizenzvalidierung
- **Der Kunde betreibt:** Control Plane und Connector
- **Passt für:** Organisationen, in denen die Auftragsverarbeitung durch Dritte selbst der Blocker ist — oder in denen ein Argument zu ausländischen Herausgabepflichten strukturell statt vertraglich geschlossen werden muss

Topologie C — Air-gapped

Topologie B ohne jeglichen Egress. Updates kommen als signierte Artefakte über den bestehenden Offline-Prozess des Kunden. Der Modellzugriff ist auf Modelle beschränkt, die innerhalb der Umgebung gehostet werden.

- **Überquert die Grenze:** nichts
- **Passt für:** Verteidigung, kritische Infrastruktur und regulierte Umgebungen mit hartem No-Egress-Gebot
- **Kosten:** manueller Update-Prozess, keine Telemetrie, längere Incident-Diagnose

VERGLEICH

	A · Managed	B · Self-hosted	C · Air-gapped
Zugangsdaten im Kundennetzwerk	Ja	Ja	Ja
Unmaskierte Nutzdaten verlassen das Netzwerk	Nein	Nein	Nein
Konfiguration extern gehalten	Ja	Nein	Nein
Drittanbieter als Auftragsverarbeiter	Ja	Nein	Nein
Betriebsaufwand beim Kunden	Niedrig	Mittel	Hoch
Update-Pfad	Automatisch	Automatisch	Manuell, signiert
Zeit bis zur Produktion	Wochen	Wochen bis Monate	Monate

Die Unterscheidung, auf die es in den meisten Beschaffungsgesprächen ankommt, ist enger, als sie zunächst wirkt. In allen drei Topologien bleiben Zugangsdaten und unmaskierte Nutzdaten im Netzwerk. Was sich unterscheidet, ist, ob *Konfiguration und Metadaten* von einem Dritten gehalten werden. Für einen regulierten Käufer ist das eine berechtigte Frage — aber eine kleinere als „verlassen unsere Daten das Haus“, und wer sie präzise stellt, verkürzt das Gespräch.

6 Ablauf einer Anfrage

Eine einzelne Anfrage in Topologie A, von Anfang bis Ende:

- 1 Der Nutzer authentifiziert den KI-Client am Gateway-Endpunkt über den Identity Provider der Organisation. Die Sitzung wird an eine benannte Identität gebunden.
- 2 Der Client sendet `tools/list`. Das Gateway löst die Identität gegen das Rollenmodell auf und liefert nur erlaubte Tools zurück.

- 3 Der Nutzer stellt eine Frage in natürlicher Sprache. Das Modell wählt ein Tool und konstruiert die Argumente.
- 4 Der Client sendet `tools/call`. Das Gateway bewertet den Aufruf gegen die Policy: Darf diese Identität dieses Tool, mit diesen Argumenten, gegen dieses Ziel, genau jetzt?
- 5 Bei Genehmigung wird der Aufruf für den zuständigen Connector eingereicht.
- 6 Der Connector — der bereits eine ausgehende Verbindung hält — holt den genehmigten Aufruf ab, löst das Credential lokal auf und führt ihn gegen das interne System aus.
- 7 Das Ergebnis wird im Connector transformiert: Maskierung auf Feldebene, Pseudonymisierung, DLP-Prüfung.
- 8 Das transformierte Ergebnis geht über den bestehenden Kanal zurück und wird an den Client weitergereicht.
- 9 Das Modell erzeugt eine Antwort. Surrogate darin werden vom Connector in echte Werte zurückübersetzt, bevor sie angezeigt werden.
- 10 Das Gateway protokolliert die vollständige Transaktion: Identität, Tool, Argumente, Ziel, Policy-Entscheidung, Ergebnisgröße, Zeitstempel.

In den Schritten 4 und 7 lebt das Produkt. In Schritt 6 hält die Credential-Grenze.

7 Audit, Aufbewahrung und der Betriebsrat

Audit-Logging ist einfach zu bauen und im deutschsprachigen Raum schwer richtig hinzubekommen — aus einem Grund, der organisatorisch ist, nicht technisch.

Ein vollständiges Audit-Protokoll jeder KI-Interaktion jedes namentlich benannten Mitarbeiters ist aus Sicht eines Betriebsrats ein System, das zur Überwachung von Leistung und Verhalten geeignet ist — womit in Deutschland Mitbestimmungsrechte nach §87 BetrVG greifen. Ein Rollout, der das ignoriert, wird nach der technischen Einführung gestoppt: an dem Punkt, an dem eine Änderung am teuersten ist.

Drei Designentscheidungen adressieren das:

- **Logs nach Zweck trennen.** Sicherheits-Audit-Datensätze (Identität, Tool, Ziel, Entscheidung) werden nach definierter Richtlinie in der Umgebung des Kunden aufbewahrt. Produkt-Telemetrie an den Anbieter ist pseudonymisiert oder aggregiert und enthält keine personenbezogenen Verhaltensdetails.
- **Logs mit Nutzdaten lokal halten.** Argumente und Ergebnisse von Tool-Aufrufen können personenbezogene Daten enthalten. Sie gehören in die Umgebung des Kunden, unter dessen Aufbewahrungsrichtlinie — nicht in die des Anbieters.
- **Aufbewahrung explizit definieren, pro Log-Klasse.** „Wir loggen alles unbegrenzt“ ist keine Antwort, die eine Datenschutzprüfung übersteht.

Anbieter in diesem Markt sollten ein Löschkonzept und eine Muster-Betriebsvereinbarung als Dokumente übergeben können. Begriffe wie *revisionssicher* oder *betriebsratsfest* ohne diese Artefakte dahinter überstehen die erste juristische Prüfung nicht.

8 Was ein Gateway nicht löst

Prompt Injection. Inhalte aus einem internen System können Anweisungen enthalten, auf die das Modell reagiert. Ein Gateway kann einschränken, welche Tools aufrufbar sind, und protokollieren, was passiert ist; es kann das Modell nicht dazu bringen, feindliche Inhalte in einem Dokument zu ignorieren. Die Gegenmaßnahmen sind architektonisch — Bestätigungspflicht für zustandsändernde Operationen, enge Schreib-Scope —, nicht präventiv.

Überberechtigte Quellsysteme. Ist der zugrunde liegende Service-Account breit berechtigt, kann das Gateway ihn einengen — aber es erbt, was das Quellsystem erlaubt. Least Privilege muss auch im Quellsystem etabliert werden.

Aufbewahrung und Training auf Modellseite. Was der Anbieter mit empfangenen Daten tut, regeln Vertrag und Anbieterkonfiguration, nicht das Gateway. Das Gateway reduziert, was gesendet wird; was danach passiert, kontrolliert es nicht.

Der Confused Deputy. Ein Gateway, das Aufrufe im Namen von Nutzern ausführt, ist ein hochwertiges Ziel: Es hält per Design breite Konnektivität. Seine eigene Kompromittierung ist das dominante Risiko der Architektur — weshalb seine Authentifizierung, sein Umgang mit Geheimnissen und seine Update-Kette mehr Prüfung verdienen als jede andere Komponente.

Compliance. Ein Gateway liefert technische und organisatorische Maßnahmen. Rechtsgrundlage, Risikoklassifizierung, Verarbeitungsverzeichnis, Transparenzpflichten und die Betreiberpflichten aus dem AI Act bleiben Verantwortung des Betreibers. Werkzeuge unterstützen die Dokumentation; die Compliance erzeugen sie nicht.

9 Checkliste für die Evaluierung

Fragen, die sich jedem Anbieter dieser Kategorie stellen lassen — auch diesem:

01 Wo genau wird maskiert, und können Sie zeigen, dass es auf unserer Seite der Grenze passiert?

02 Was ist der Standard für ein unklassifiziertes Feld — durchlassen oder verweigern?

03 Wie werden Freitextfelder behandelt, und welches Restrisiko wird ausdrücklich benannt?

04 Wo liegen die Systemzugangsdaten, und was hält die Control Plane stattdessen?

05 Wird der Berechtigungsentzug zum Aufrufzeitpunkt durchgesetzt — oder hängt er davon ab, dass der Client seine Tool-Liste aktualisiert?

06 Was genau wird in der Managed-Topologie an Sie übertragen? Verlangen Sie eine Feldliste, keine Kategorie.

- 07 Wie lautet die Aufbewahrungsrichtlinie pro Log-Klasse, und können wir sie selbst festlegen?
- 08 Können Sie ein Löschkonzept und eine Muster-Betriebsvereinbarung als Dokumente vorlegen?
- 09 Welche Zertifizierungen existieren heute, welche sind „in Arbeit“? Verlangen Sie das Zertifikat.
- 10 Was passiert mit einer laufenden Sitzung, wenn der Connector die Verbindung verliert?

Anhang — Unterstützte Quelltypen

DATENBANKEN

PostgreSQL, MySQL/MariaDB, Microsoft SQL Server, Oracle, MongoDB

BUSINESS-SYSTEME

SAP, Microsoft Dynamics/Navision, Salesforce, ServiceNow

GENERISCH

HTTP/REST-APIs mit konfigurierbarer Authentifizierung

NACHGELAGERTE MCP-SERVER

Jeder konforme MCP-Server, fördert hinter derselben Policy-Schicht

CLIENTS

ChatGPT, Claude, Langdock, Open WebUI und jeder MCP-fähige Client

Dieses Dokument beschreibt Architektur und ist keine Rechtsberatung. Compliance-Bewertungen für ein konkretes Deployment erfordern die Prüfung durch qualifizierte Rechtsberatung und den zuständigen Datenschutzbeauftragten.