

KI-Agenten, Identität und Kontrollarchitektur

Technischer Leitfaden für begrenzte und nachweisbare Handlungsvollmachten

Fachlicher Leitfaden zur Cyber-Physical-AI-Security.

Fassung 1.1 · Stand 25. September 2026 · Adressaten: Sicherheitsarchitektur, KI-Engineering, Softwareentwicklung, Plattformbetrieb, IAM, OT-Integration und technische Prüfer.

Dokumenttyp: Referenzarchitektur und Prüfanforderungen; keine zertifizierte Produktimplementierung

Die Fallstudien in diesem Leitfaden-Paket sind ausdrücklich konstruierte Prüfbeispiele; eine externe wissenschaftliche, klinische, juristische oder sicherheitstechnische Begutachtung des Pakets hat nicht stattgefunden.

Steckbrief

Merkmal	Wert
Dokumenttyp	Referenzarchitektur und Prüfanforderungen; keine zertifizierte Produktimplementierung
Zielgruppe	Sicherheitsarchitektur, KI-Engineering, Softwareentwicklung, Plattformbetrieb, IAM, OT-Integration und technische Prüfer.
Recherche- und Rechtsstand	25. September 2026
Umfang	3.640 Wörter, 14 Seiten
Belege	0 Literatureinträge, 9 Quellenadressen
Gliederung	20 nummerierte Abschnitte, 0 Tabellen

Inhalt

1. Sicherheitsproblem und Entwurfsprinzip	4
2. Bedrohungsmodell: Welche Akteure werden angenommen?	4
3. Identität, Auftraggeber und Verantwortung	4
4. Delegation ist enger als Nachahmung eines Benutzers	4
5. Referenzarchitektur mit getrennten Aufgaben	5
6. Werkzeuge als eng definierte Funktionen	5
7. Kontextprüfung ohne Selbstbeglaubigung	5
8. Datenzugriff und Datenausleitung	6
9. Prompt Injection, Datenmanipulation und Modellfehler unterscheiden	6
10. Zeitgrenzen, Token und der Widerrufsirrtum	6
11. Prüfung unmittelbar vor Ausführung	7
12. Sicherheits- und Verfügbarkeitsverhalten bei Fehlern	7
13. Menschliche Freigaben mit tatsächlichem Informationswert	7
14. Defensive KI und begrenzte Reaktionsrechte	8
15. Beobachtbarkeit und forensische Aussagekraft	8
16. Modell-, Werkzeug- und Lieferkettenänderungen	8
17. Prüfinvarianten für die technische Abnahme	9
18. Testmatrix und sichere Durchführung	9
19. Einführungs- und Betriebsentscheidung	9
20. Arbeitsunterlagen und Ergebnis	9
Literatur und Quellenstatus	10
12 Rechtliche Hinweise	12
12.1 Keine Beratung	12
12.2 Keine Zusicherung, keine Konformitätsaussage	12
12.3 Rechtsstand	12
12.4 Haftung	12
12.5 Verweise auf fremde Inhalte	12
12.6 Marken und Produktbezeichnungen	12
12.7 Urheberrecht und Nutzung	12
12.8 Keine personenbezogenen Daten	13
12.9 Unabhängigkeit	13
12.10 Anwendbares Recht	13

Impressum und Verantwortlichkeit

Die folgenden Angaben sind wörtlich dem Impressum von fluxlane.eu entnommen, abgerufen am 24. September 2026. Sie wurden nicht ergänzt und nicht umformuliert.

Angabe	Inhalt
Angaben gemäß § 5 DDG	WebTrades GmbH Am Ölberg 2 55459 Grolsheim Deutschland
Vertreten durch	Geschäftsführerin: Laurice Bussmer
Kontakt	Telefon: 0170 890 78 00 E-Mail: info@fluxlane.eu
Registereintrag	Eintragung im Handelsregister Registergericht: Amtsgericht Mainz Registernummer: HRB 54479
Umsatzsteuer-ID	Umsatzsteuer-Identifikationsnummer gemäß § 27 a Umsatzsteuergesetz: DE461650292
Redaktionell verantwortlich	Verantwortlich für den Inhalt nach § 18 Abs. 2 MStV: Laurice Bussmer, Am Ölberg 2, 55459 Grolsheim
Verbraucherstreitbeilegung	Wir sind nicht bereit oder verpflichtet, an Streitbeilegungsverfahren vor einer Verbraucherschlichtungsstelle teilzunehmen.

Redaktioneller Hinweis des Herausgebers (wörtlich)

„fluxlane ist ein unabhängiges, redaktionelles Fachportal für Robotik, autonome physische Systeme und verkörperte KI. Der redaktionelle Geltungsbereich umfasst Industrie und Produktion, Logistik, Medizin, Pflege und Rehabilitation, Service und Haushalt sowie Forschung, Hersteller, Markt und Regulierung; künstliche Intelligenz gehört dazu, auch ohne physischen Körper. Die Inhalte dienen der allgemeinen Information und stellen keine Rechts-, Kauf- oder Anlageberatung dar. Angaben zu Preisen, Verfügbarkeit und Terminen erfolgen ohne Gewähr und können sich jederzeit ändern.“

Zitierweise

FluxLane (2026): *KI-Agenten, Identität und Kontrollarchitektur – Technischer Leitfaden für begrenzte und nachweisbare Handlungsvollmachten*. Fassung 1.1, Stand 25. September 2026. Herausgegeben von der WebTrades GmbH, Grolsheim, fluxlane.eu.

Herstellung dieses Dokuments

Das vorliegende PDF ist maschinell aus der Markdown-Fassung des Leitfaden-Pakets gesetzt. Inhaltlich ist dieser Satzvorgang keine neue Sachprüfung; die zugrunde liegende Recherche- und Quellenprüfung ist im Lieferpaket dokumentiert. Gesetzt wurde mit IBM Plex Sans und Sora.

Abbildungsnachweis

Dieses Dokument enthält eine schematische Abbildung: die Referenzarchitektur mit fünf logisch getrennten Bereichen (Abschnitt 5). Die Abbildung zeigt eine logische Trennung, keine fünf zwingend separaten Produkte und keine Herstellerlösung. Ergänzt in Fassung 1.1 am 27. September 2026.

1. Sicherheitsproblem und Entwurfsprinzip

Ein KI-Agent wird nicht durch seine sprachlichen Fähigkeiten gefährlich, sondern durch die Kombination aus Eingaben, Werkzeugen, Datenzugriff und Handlungsvollmacht. Ein Modell ohne ausführende Werkzeuge besitzt andere technische Wirkungsmöglichkeiten als ein Agent mit Datenbankzugriff, Auftragsschnittstelle und Administrationskonto. Die Sicherheitsprüfung muss deshalb die gesamte Laufzeitumgebung erfassen.

Eine geeignete Architektur behandelt ein Modellresultat zunächst als **Vorschlag innerhalb einer begrenzten Zuständigkeit**. Erst getrennte Komponenten entscheiden, ob und wie daraus ein verbindlicher Systemzugriff wird. Diese Trennung ist besonders wichtig, wenn unzuverlässige Eingaben, externe Dokumente oder manipulierbare Werkzeugantworten in die Entscheidung einfließen. Die NIST-Taxonomie zu adversarialer KI bietet einen Ordnungsrahmen für solche Angriffsklassen; sie ersetzt keine lokale Pfadanalyse. [Q11]

Der Entwurf dieses Bandes ist eine Synthese aus ressourcenbezogenem Zugriffsschutz, Agentenidentität, expliziter Delegation und OT-spezifischer Wirkungsbegrenzung. Er ist kein neu behaupteter Industriestandard. NIST bearbeitet Agent Identity and Authorization in einem laufenden Projekt; daraus wird hier kein fertiges Zertifizierungsprogramm abgeleitet. [Q06; Q09]

2. Bedrohungsmodell: Welche Akteure werden angenommen?

Das Modell berücksichtigt externe Akteure ohne Zugang, gestohlene oder missbrauchte legitime Identitäten, kompromittierte externe Dienste, interne Fehlbedienung und einen Agenten, der unzuverlässige Informationen verarbeitet. Zusätzlich wird der Ausfall wichtiger Schutzkomponenten betrachtet. Diese Annahmen unterscheiden sich im Zugriff und dürfen nicht in einer einzigen Aussage „Hacker kontrolliert alles“ zusammenfallen.

Für jeden angenommenen Akteur wird festgehalten, welche Eingaben, Speicher, Verbindungen und Identitäten er beeinflussen kann. Ein manipulierbares Dokument ist etwas anderes als ein kompromittierter Identitätsdienst. Wenn letzterer vollständige Administratorvollmachten vergeben kann, reicht die Prüfung einzelner Toolaufrufe nicht mehr. Die Vertrauensbasis des gesamten Systems muss dann neu bewertet werden.

Nicht jede Bedrohung kann durch dieselbe Architektur vollständig beherrscht werden. Physischer Zugriff auf nicht geschützte Hardware, kompromittierte Firmware oder ein absichtlich handelnder hochprivilegierter Administrator können zusätzliche Maßnahmen verlangen. Der Sicherheitsnachweis nennt deshalb ausdrücklich, welche Vertrauensanker vorausgesetzt werden und welcher zusätzliche Schutz gegen deren Ausfall besteht. Unbenannte Annahmen sind keine Sicherheitsbarrieren.

3. Identität, Auftraggeber und Verantwortung

Jeder ausführende Agent erhält eine eigene technische Identität. Diese Identität ist einer verantwortlichen Organisation, einem System und gegebenenfalls einem menschlichen Auftraggeber zugeordnet. Sie soll nicht durch das unkontrollierte Kopieren persönlicher Zugangsdaten entstehen. Der NIST-Fachbeitrag vom August 2026 betont eigenständige Identitäten und die Risiken breit berechtigter, langlebiger Zugangsmittel. [Q10]

Die Zuordnung muss für einzelne Aufträge nachvollziehbar bleiben. Ein Protokoll „service-account schrieb Datensatz“ kann für die technische Analyse nützlich sein, beantwortet aber noch nicht, welches Mandat dahinterstand. Benötigt werden Auftrag, Identität, delegierte Befugnis, betroffene Ressource und verantwortete Entscheidung. Bei einer Kette von Subagenten darf diese Zuordnung nicht verloren gehen.

Verantwortlichkeit ist gleichwohl keine automatische Schuldzuweisung. Ein korrekt dokumentierter Auftrag kann durch eine fehlerhafte Integration anders ausgeführt worden sein. Die Protokollierung soll den Ablauf rekonstruierbar machen, nicht jedes Fehlverhalten vorschnell dem Auftraggeber zuschreiben. Technische Attribution und rechtliche Verantwortung sind getrennt zu prüfen.

4. Delegation ist enger als Nachahmung eines Benutzers

RFC 8693 unterscheidet Delegation und Impersonation. Für sicherheitsrelevante Agentensysteme ist diese Unterscheidung hilfreich: Der Agent soll erkennbar im Auftrag handeln, statt unsichtbar sämtliche Rechte der

beauftragenden Person zu übernehmen. Eine organisatorische Freigabe für einen begrenzten Vorgang wird nicht durch die Weitergabe eines Vollzugriffstokens korrekt umgesetzt. [Q14, § 1.1]

Ein Delegationsdatensatz enthält mindestens Auftraggeber, ausführende Identität, Ressource, erlaubte Aktion, Zweck, Zeitgrenze, mögliche Unterdelegation und Widerrufsbezug. Für besonders relevante Handlungen kommen zulässiger Betriebsmodus und erforderliche Freigaben hinzu. Der Datensatz wird von einer dafür zuständigen Vertrauensinstanz ausgestellt; der Agent darf seine eigenen Grenzen nicht durch Änderung einer Textanweisung erweitern.

Unterdelegation darf Rechte nicht vergrößern. Ein Subagent kann nur einen Teil des noch gültigen Mandats erhalten. Die Kette besitzt eine begrenzte Tiefe und einen gemeinsamen Auftragshintergrund. Entfällt die übergeordnete Befugnis, müssen die abhängigen Befugnisse nach dem dokumentierten Widerrufsmodell behandelt werden. Ob das technisch durch Tokenaustausch, kurzlebige Berechtigungen oder eine zentrale Autorisierungsentscheidung umgesetzt wird, hängt vom System ab; die gewünschte Eigenschaft ist unabhängig von einem bestimmten Produkt.

5. Referenzarchitektur mit getrennten Aufgaben

Die Architektur besteht logisch aus fünf Bereichen. **Auftrag und Identität** verbinden Zweck, Akteur und Delegation. **Agentenlaufzeit** verarbeitet Informationen und erzeugt Aktionsvorschläge. **Autorisierungs- und Ausführungsschicht** prüft Mandat, Ressource, Operation und Kontext. **Prozess- und Schutzebene** setzt fachliche beziehungsweise physische Grenzen durch. **Nachweis- und Betriebsdienste** erfassen relevante Ereignisse, ermöglichen Eingrenzung und unterstützen Wiederherstellung.

Diese Bereiche sind nicht zwingend fünf separate Produkte. Eine logische Trennung muss jedoch so umgesetzt sein, dass die Agentenlaufzeit ihre eigene Freigabe nicht umgehen oder die letzte relevante Schutzgrenze verändern kann. Ein Proxy mit derselben uneingeschränkten administrativen Identität wie der Agent ist dafür kein ausreichender Nachweis.

Der Begriff „Cyber-Physical Policy & Safety Layer“ aus dem Master wird hier präzisiert: Er bezeichnet eine **Zusammenarbeit verschiedener Kontrollfunktionen**, nicht eine einzige magische Sicherheitskomponente. Eine gewöhnliche API-Policy-Engine ist nicht allein durch ihre Position im Datenfluss eine sicherheitsgerichtete Maschinensteuerung. Safety-relevante Funktionen benötigen den zu ihrem Einsatzzweck passenden Engineering- und gegebenenfalls normativen Nachweis. [Q02, §§ 4.2.2 und 5.3; Q22, Annex IV]

6. Werkzeuge als eng definierte Funktionen

Eine Werkzeugbeschreibung ist Teil der Angriffsfläche. Die technische Schnittstelle sollte möglichst genau die fachlich erforderliche Aktion ausdrücken. „Lese freigegebene Auftragsdaten“ ist enger als ein allgemeiner Datenbankzugriff. „Erzeuge einen nicht freigegebenen Entwurf“ ist enger als freies Schreiben in einer Akte. Der Agent benötigt nicht allein deshalb eine Shell, weil sie die Integration erleichtert.

Eingaben werden anhand eines eindeutig versionierten Schemas geprüft. Zulässige Ressourcentypen, Felder, Wertebereiche und Operationen müssen technisch begrenzt sein. Die Prüfung umfasst auch unerwartete Zusatzfelder, fehlende Angaben und widersprüchliche Zustände. Ein natürlichsprachlicher Werkzeugaufruf darf keinen unkontrollierten Umweg zu einer allgemeineren Funktion eröffnen.

Die Antwort eines Werkzeugs bleibt Dateninput, nicht automatisch eine neue Autorität. Enthält ein externer Inhalt die Behauptung, eine Policy sei geändert oder ein zusätzlicher Zugriff sei freigegeben, darf daraus keine neue Berechtigung entstehen. Eine solche Änderung muss aus dem vorgesehenen administrativen Verfahren und seiner vertrauenswürdigen Quelle stammen. Genau diese Trennung wird im Negativtest geprüft.

7. Kontextprüfung ohne Selbstbeglaubigung

Kontext kann Aufgabe, Wartungsfenster, Patienten- oder Auftragsbezug, Betriebsart und aktuelle Ressourcenversion umfassen. Die Autorisierung muss wissen, welche dieser Merkmale verbindlich sind und woher sie stammen. Ein Agent, der seinen eigenen Auftragstext frei verändern darf, kann diesen Text nicht zugleich als unabhängige

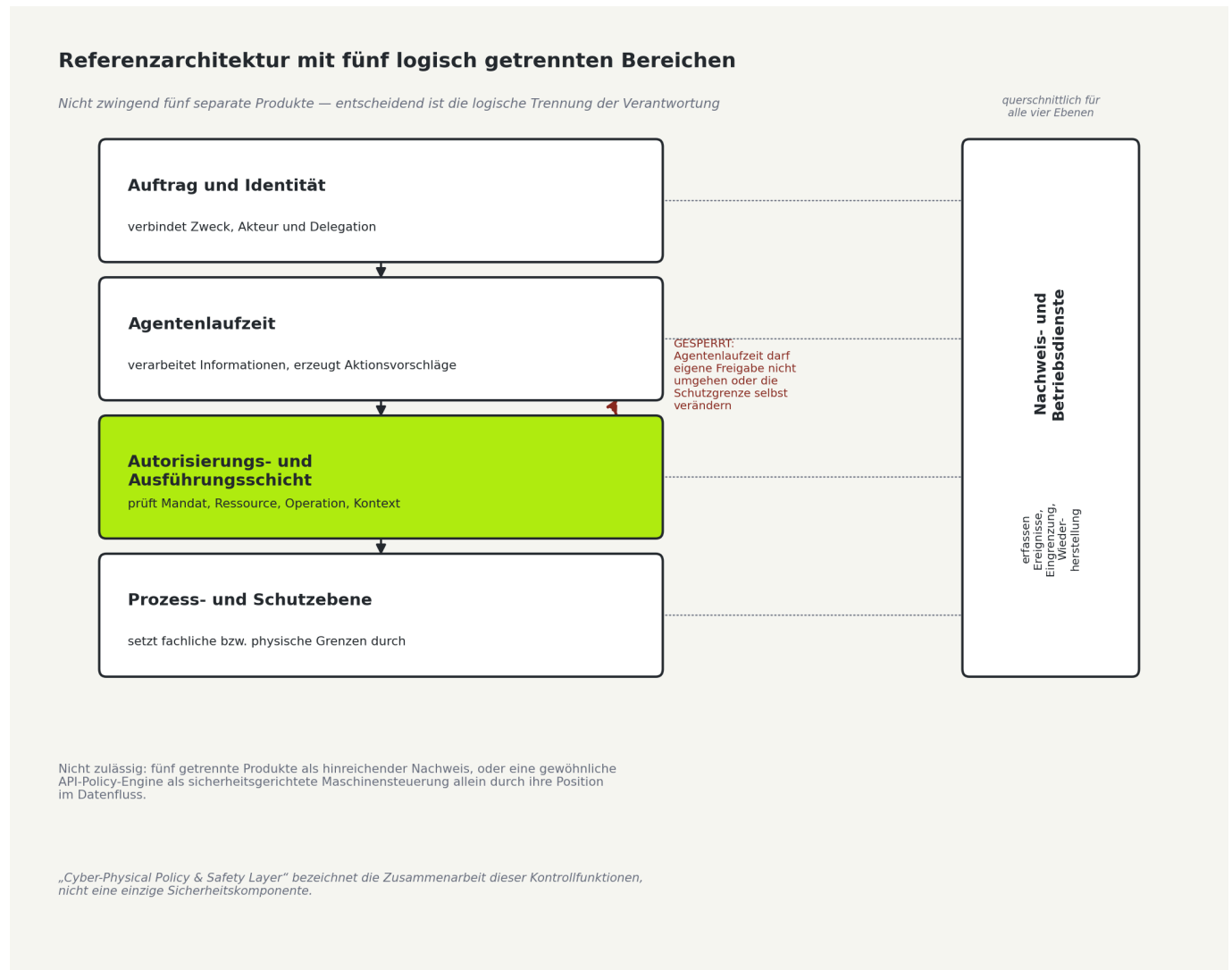


Abbildung · Referenzarchitektur mit fünf logisch getrennten Bereichen (zu Abschnitt 5)

Bestätigung seiner Befugnisse verwenden.

Die Kontextquelle kann selbst ausfallen oder veraltet sein. Das System benötigt dann ein ausdrücklich festgelegtes Verhalten. Ein rein administrativer Datenexport kann unter Umständen blockiert werden. Eine bereits laufende physische Funktion darf hingegen nicht unkontrolliert von einer entfernten Kontextprüfung abhängig sein, wenn deren Ausfall gefährlich wäre. In solchen Fällen muss die lokale Schutz- und Betriebslogik den geeigneten Zustand gewährleisten.

Kontextabhängige Autorisierung ist keine völlige Neuerfindung jenseits von Zero Trust. Bereits ressourcenbezogene Zugriffsmodelle ermöglichen differenzierte Entscheidungen. Der zusätzliche Schwerpunkt dieses Leitfadens ist die **nachweisbare Verknüpfung mit der fachlichen Handlung und ihren möglichen physischen Folgen**. Eine Firewall kann je nach Produkt sehr wohl Anwendungsmerkmale auswerten; die relevante Frage bleibt, ob die konkrete Architektur die erforderliche Bedeutung und Wirkung tatsächlich durchsetzt. [Q06]

8. Datenzugriff und Datenausleitung

Leserechte sind nicht folgenlos. Ein Agent mit Zugriff auf große Datenbestände kann sensible Informationen in Ausgaben, externe Dienste, Protokolle oder Exportdateien übertragen. Deshalb werden Eingangs- und Ausgangsgrenzen getrennt geprüft. Der zulässige Datenumfang richtet sich nach Zweck und Aufgabe, nicht allein danach, ob der Agent technisch lesen kann.

Eine geeignete Ausleitungsregel beschreibt Zielsystem, Datenkategorie, Zweck, Umfang, Frequenz und erforderliche Freigabe. Für besonders sensible Inhalte ist zu prüfen, ob eine lokale Verarbeitung, eine engere Auswahl oder eine kontrollierte Aggregation ausreicht. Diese Maßnahmen benötigen einen Test mit dem real vorgesehenen Workflow, weil eine scheinbar anonyme technische Kennung zusammen mit weiteren Informationen identifizierend sein kann.

Protokollierung darf nicht unbemerkt einen zweiten ungeschützten Datenbestand erzeugen. Vollständige Eingaben und Ausgaben sind nicht immer notwendig, um eine Handlung nachzuvollziehen. Aufzubewahren sind die für Zweck und Nachweis erforderlichen Informationen mit geeigneter Zugriffskontrolle und begründeter Frist. Rechtsgrundlage und Betroffenenrisiko werden getrennt geprüft. [Q38, Art. 5, 25, 32 und gegebenenfalls 35 als Nachschlagepfad]

9. Prompt Injection, Datenmanipulation und Modellfehler unterscheiden

Prompt Injection betrifft hier den Versuch, über nicht vertrauenswürdige Inhalte das Verhalten eines sprachmodellgestützten Systems zu beeinflussen. Datenmanipulation kann dagegen fachliche Werte verändern, ohne überhaupt eine sprachliche Instruktion zu enthalten. Ein Modellfehler kann auch bei unveränderten Daten auftreten. Die gemeinsame Schutzfrage lautet, welche unzulässige Wirkung aus diesen Eingaben entstehen kann; die Ursachen bleiben für Untersuchung und Gegenmaßnahmen unterscheidbar. [Q11]

Die wichtigste Barriere ist nicht die Hoffnung auf eine perfekte Formulierung der Systemanweisung. Es sind begrenzte Werkzeuge, unabhängige Autorisierung und klar getrennte Vertrauensdomänen. Ein Modell darf unzuverlässige Inhalte lesen, soweit der Zweck dies erfordert; diese Inhalte dürfen jedoch nicht die Zugriffsregeln der Umgebung neu definieren. Zusätzliche Modellprüfungen können helfen, ersetzen diese Durchsetzung aber nicht.

Tests verwenden ungefährliche synthetische Inhalte, die beispielsweise eine angebliche zusätzliche Freigabe behaupten. Geprüft wird, ob ein nicht erlaubter Vorgang technisch abgewiesen und nachvollziehbar protokolliert wird. Ein Test gilt nicht schon dann als bestanden, wenn das Modell freundlich widerspricht. Maßgeblich ist, ob der tatsächliche Werkzeug- oder Datenzugriff ausbleibt. Ebenso muss ein zulässiger Referenzvorgang weiterhin funktionieren, damit die Schutzmaßnahme nicht nur jede Arbeit verhindert.

10. Zeitgrenzen, Token und der Widerrufsirrtum

Ein widerrufenes Zugangsmittel ist nicht automatisch überall sofort unwirksam. Verteilte Dienste können Berechtigungen zwischenspeichern, bereits erteilte Aufträge weiterverarbeiten oder bei Verbindungsverlust lokale Entscheidungen treffen. RFC 7009 beschreibt Tokenwiderruf und Implementierungsfragen; daraus folgt kein automatischer Stopp bereits laufender physischer Aktionen. [Q15, §§ 2–3]

Der Entwurf benötigt deshalb einen **Widerrufspfad bis zur Wirkung**. Dokumentiert werden Erkennung des Verdachts, Entscheidung, Verteilung des Widerrufs, Verhalten der Endpunkte, Behandlung von Warteschlangen und Umgang mit bereits laufenden Aufträgen. Für jede Phase wird die maximal akzeptierte Verzögerung aus dem konkreten Risiko abgeleitet. Es gibt keinen in diesem Paket vorgegebenen universellen Sekundenwert.

Wo die mögliche gefährliche Wirkung schneller eintreten kann als ein Netzwerk-Widerruf, muss eine geeignete lokale Begrenzung greifen. Ein zentraler Identitätsdienst kann diese Funktion nicht allein durch häufigere Tokenprüfung ersetzen. Gleichzeitig darf eine abgelaufene Berechtigung nicht dazu führen, dass eine notwendige lokale Schutzfunktion ihre Arbeit beendet. Zugriffsschutz und sicherer Prozesszustand haben unterschiedliche Aufgaben.

11. Prüfung unmittelbar vor Ausführung

Zwischen Freigabe und Ausführung kann sich der Zustand ändern. Ein Auftrag kann gültig gewesen sein, als er erstellt wurde, und ungültig werden, bevor ihn das Zielsystem bearbeitet. Die Schnittstelle muss deshalb festlegen, wann welche Bedingungen erneut geprüft werden. Das kann etwa die aktuelle Betriebsart, die Ressourcenversion und die noch gültige Delegation betreffen.

Eine technische Umsetzung kann einen Auftrag an eine erwartete Zustandsversion und eine eindeutige Auftragskennung binden. Das Zielsystem muss dann definieren, ob ein bereits bearbeiteter Auftrag erneut angenommen wird, wie widersprüchliche Versionen behandelt werden und wann eine neue Freigabe nötig ist. Solche Mechanismen werden hier als Entwurfsanforderungen beschrieben, nicht als sofort deploybare universelle Middleware.

Bei physischen Vorgängen ist besondere Vorsicht mit der Annahme der Wiederholbarkeit erforderlich. Ein erneut gesendeter Auftrag kann eine zweite reale Handlung auslösen. Die technische Eigenschaft „idempotent“ muss auf die tatsächliche Wirkung bezogen und getestet werden. Eine API, die denselben Erfolgsstatus zurückgibt, beweist noch nicht, dass im Gerät keine zweite Bewegung oder Verarbeitung stattgefunden hat.

12. Sicherheits- und Verfügbarkeitsverhalten bei Fehlern

„Fail closed“ ist für viele Zugriffsentscheidungen sinnvoll: Bei fehlender Berechtigung wird kein neuer Zugriff erteilt. Als universelle Regel für physische Prozesse ist der Ausdruck jedoch zu ungenau. Eine lebens- oder sicherheitsrelevante laufende Funktion benötigt ein aus der Gefährdung abgeleitetes Verhalten, das nicht mit dem einfachen Verweigern einer API-Anfrage gleichgesetzt werden kann.

Die Architektur muss daher Fehlerfälle je Funktion beschreiben. Was geschieht bei nicht erreichbarer Policy-Engine? Bei fehlerhafter Uhrzeit? Bei widersprüchlichen Zustandsmeldungen? Bei vollständigem Protokollierungsausfall? Darf nur lesend weitergearbeitet werden, muss ein neuer Auftrag abgewiesen werden oder übernimmt eine lokale Betriebslogik? Wer wird informiert und welche Beobachtung bestätigt den eingetretenen Zustand?

Das System sollte nicht dadurch „verfügbar“ bleiben, dass es bei Fehlern pauschal alle Zugriffe erlaubt. Ebenso wenig ist ein zentraler Totalausfall als vermeintlich sichere Standardreaktion akzeptabel. Die richtige Entscheidung ist funktions- und zustandsabhängig und muss vorab begründet werden. Der hieraus entstehende zusätzliche Engineering-Aufwand ist Bestandteil der Einführung agentischer Funktionen, nicht eine spätere Komfortoption.

13. Menschliche Freigaben mit tatsächlichem Informationswert

Eine menschliche Bestätigung ist nur wirksam, wenn die Person versteht, worüber sie entscheidet, genügend Zeit besitzt und die Konsequenzen erkennen kann. Eine Folge austauschbarer „Zulassen“-Dialoge kann zu routinemäßiger Bestätigung führen. NIST diskutiert dieses Problem als Zustimmungsermüdung im Kontext agentischer Berechtigungen. [Q10]

Die Freigabeoberfläche sollte deshalb die konkrete Änderung, betroffene Ressource, mögliche Wirkung, gültigen Zeitraum und Abweichung vom Normalfall zeigen. Eine Person muss eine Alternative wählen und eine Ablehnung ohne unzulässigen Betriebsdruck aussprechen können. Für bestimmte Tätigkeiten ist eine zweite fachliche Freigabe sinnvoll; diese muss tatsächlich unabhängig sein und darf nicht nur ein weiterer Klick derselben automatisierten

Kette sein.

Nicht jede Routineaktion braucht diese Form menschlicher Zustimmung. Der Mensch legt den zulässigen Handlungsspielraum fest und entscheidet über relevante Erweiterungen, Ausnahmen und Wiederfreigaben. Welche Aktionen autonom zulässig sind, wird aus dem Risikoprofil abgeleitet. „Human in the loop“ ist kein pauschaler Sicherheitsnachweis; entscheidend sind Rolle, Information, Zeit und technische Durchsetzbarkeit der Entscheidung.

14. Defensive KI und begrenzte Reaktionsrechte

Ein Verteidigungsagent kann Beobachtungen zusammenführen und Gegenmaßnahmen vorbereiten. Seine Rechte werden jedoch genauso begrenzt wie die eines produktiven Agenten. Eine Detektionsfunktion soll nicht automatisch alle Konten sperren oder jede kritische Verbindung trennen können. Das Reaktionsmandat nennt den zugelassenen Umfang und die möglichen Folgen.

Ein sinnvolles Modell unterscheidet Beobachten, Vorschlagen, eng begrenztes Blockieren, Widerrufen und das Anfordern weitergehender Maßnahmen. Bei jeder automatischen Aktion wird geprüft, ob ein Fehlalarm einen größeren Schaden verursachen könnte als die erwartete Verzögerung einer manuellen Entscheidung. Dieser Vergleich erfordert lokale Informationen; er lässt sich nicht durch eine universelle Autonomieskala beantworten.

Die Verteidigung selbst wird in Negativtests angegriffen, ohne reale Schäden auszulösen: irreführende Logeinträge, ungewöhnliche, aber legitime Wartung, fehlende Kontextdaten oder gleichzeitig auftretende Routineereignisse. Der Agent muss Unsicherheit anzeigen und innerhalb seiner Vollmacht bleiben. Eine ungewöhnliche Aktivität ist kein ausreichender Grund, einen sicherheitskritischen Prozess ohne genehmigte Wirkungslogik zu unterbrechen.

15. Beobachtbarkeit und forensische Aussagekraft

Ein brauchbarer Ereignisdatensatz enthält Auftrag, ausführende Identität, Delegationsbezug, Ressource, angeforderte Aktion, Autorisierungsentscheidung, verwendete Policy-Version, Ausführungsergebnis und relevante Zeitangaben. Wo tatsächliche Wirkung wichtig ist, muss eine geeignete Rückmeldung vom Zielsystem vorliegen. Die bloße Meldung des Agenten „erledigt“ ist keine unabhängige Bestätigung.

Protokolle benötigen eine nachvollziehbare Zeitbasis, Zugriffsrechte und Schutz vor unbemerkter Veränderung. Sie müssen auch bei eingeschränktem Betrieb im erforderlichen Umfang verfügbar bleiben. Gleichwohl sind sie nicht unfehlbar. Ein kompromittierter Erzeuger kann falsche Ereignisse liefern; eine sichere Ablage macht diese nicht nachträglich wahr. Die Auswertung bewertet deshalb Herkunft und mögliche Lücken.

Nicht erforderlich ist eine vermeintlich vollständige Einsicht in interne Modellgedanken. Für den Betriebsnachweis zählen dokumentierbare Eingaben, Versionen, Entscheidungen an Kontrollpunkten und tatsächlich ausgeführte Handlungen. Eine nachträglich erzeugte sprachliche Erklärung kann hilfreich sein, ist aber kein Ersatz für diese Evidenz. Diese Abgrenzung verhindert, dass ein überzeugender KI-Kommentar mit einem forensischen Beleg verwechselt wird.

16. Modell-, Werkzeug- und Lieferkettenänderungen

Der freigegebene Agent wird durch eine Konfiguration beschrieben, nicht allein durch einen Produktnamen. Dazu gehören Modellstand, Werkzeugversionen, Schemas, Laufzeit, Systemregeln, Datenquellen, Berechtigungen und Ausführungsumgebung. Eine Änderung eines einzelnen Werkzeugs kann die Wirkung stärker verändern als ein Modellwechsel.

Die Organisation führt deshalb einen versionierten Freigabestand und prüft Änderungen risikobezogen. Neue Werkzeugrechte, weitere Datenquellen oder zusätzliche Unterdelegation lösen eine erneute Prüfung der betroffenen Wirkungsketten aus. Ein Anbieter muss relevante Änderungen, Sicherheitskorrekturen und Supportgrenzen in geeigneter Form kommunizieren. Fehlt diese Möglichkeit, ist die Abhängigkeit im Beschaffungs- und Risikoverfahren zu bewerten.

Ein Softwarekomponentenverzeichnis unterstützt die Übersicht, bildet aber nicht sämtliche Risiken ab. Externe Modellfunktionen, Konfigurationen, dynamische Werkzeuge und Datenflüsse können zusätzliche Bestandteile des

Nachweises sein. Das Leitmotiv lautet: Was das Verhalten oder die Vollmacht ändern kann, muss dem Änderungs- und Prüfverfahren zugeordnet sein.

17. Prüfinvarianten für die technische Abnahme

Die folgenden Eigenschaften sind vorgeschlagene, an den konkreten Scope anzupassende Invarianten. **I1:** Keine fachlich relevante Ausführung ohne gültige Identität und Mandat. **I2:** Unterdelegation vergrößert den erlaubten Umfang nicht. **I3:** Ein Agent kann die Policy seiner eigenen Handlung nicht unkontrolliert ändern. **I4:** Nach einem definierten Widerruf entstehen keine neuen unzulässigen Wirkungen außerhalb des dokumentierten Restverhaltens. **I5:** Protokollierung und tatsächliche Ausführung können in relevanten Fällen miteinander abgeglichen werden. **I6:** Ein Ausfall zentraler IT hebt die erforderlichen lokalen Schutzfunktionen nicht auf.

Diese Eigenschaften sind keine vollständig bewiesenen mathematischen Sätze über jede Implementierung. Sie bilden prüfbare Anforderungen. Für jede wird festgelegt, welche Komponenten und Annahmen in die Prüfung eingehen, wie ein Gegenbeispiel aussehen würde und welche Tests angemessen sind. Eine formale Verifikation kann in geeigneten Teilmodellen ergänzen, ersetzt aber nicht die Validierung der Modellannahmen und der realen Integration.

Ein Dokument „I1 erfüllt“ ohne Mess- oder Konfigurationsnachweis ist unzureichend. Ebenso ist ein erfolgreicher Test nur für den geprüften Stand gültig. Ändern sich Mandatslogik, Toolpfad oder Zustandsquelle, muss die betroffene Aussage neu bewertet werden.

18. Testmatrix und sichere Durchführung

Die Testmatrix kombiniert Identität, Berechtigung, Kontext, Systemzustand und Fehlerfall. Es werden zulässige Referenzfälle, unzulässige Aktionen, abgelaufene Mandate, Unterdelegation, widersprüchliche Zustände, Kommunikationsverlust und Wiederholungsaufträge geprüft. Für jedes Ergebnis ist das erwartete Verhalten vor dem Test dokumentiert.

Tests beginnen in isolierten Umgebungen mit synthetischen Daten. Reale Zugangsdaten, Patientendaten oder sicherheitskritische Aktoren gehören nicht in eine improvisierte Red-Team-Übung. Für spätere Integrationstests sind Verantwortliche, Stopkriterien, Rückbau und Schutzmaßnahmen genehmigt. Die Tester dürfen ihr eigenes positives Ergebnis nicht durch eine Änderung des erwarteten Verhaltens nachträglich herstellen.

Die Abnahme betrachtet Fehlablehnungen ebenso wie unzulässige Freigaben. Ein System, das alles blockiert, kann die gewünschte Betriebsleistung nicht erfüllen. Eine niedrige Fehlalarmrate ist andererseits wertlos, wenn relevante Angriffe unentdeckt bleiben. Kennzahlen benötigen daher klare Nenner, repräsentative Testklassen und eine offene Darstellung nicht untersuchter Fälle.

19. Einführungs- und Betriebsentscheidung

Die erste Freigabe sollte sich auf einen fachlich sinnvollen, aber begrenzten Umfang beziehen. Das kann ein lesender oder entwurfsorientierter Einsatz sein. Schreib- und Steuerfunktionen werden erst zugelassen, wenn deren zusätzliche Wirkungsketten nachgewiesen sind. Diese Stufung ist keine Behauptung, dass Leserechte risikolos seien; sie macht unterschiedliche Folgen separat prüfbar.

Im Betrieb werden Mandate, Ausnahmen, relevante Änderungen und Sicherheitsereignisse regelmäßig überprüft. Jede Ausnahme besitzt Eigentümer, Begründung und Ablaufdatum. Ein dauerhaftes Notfalltoken wäre keine Ausnahme mehr, sondern eine neue reguläre Vollmacht und müsste als solche bewertet werden. Gleiches gilt für pauschale Konten, die „vorübergehend“ an Integratoren vergeben wurden.

Die Freigabeentscheidung nennt schließlich, welche Rechte entzogen oder eingeschränkt werden können, wie schnell dies nachgewiesen ist und welche lokalen Funktionen trotzdem erhalten bleiben. Erst diese Verbindung von Begrenzung und Betrieb macht agentische Sicherheit praktisch belastbar.

20. Arbeitsunterlagen und Ergebnis

Die Kontrollen **K29–K38** werden durch **K01–K08** und die Betriebs-/Recovery-Kontrollen ergänzt. **F03 Handlungsvollmachten**, **F05 Nachweisfall**, **F07 Änderung**, **F09 Reaktionsmandat** und **F15 Prüfprotokoll** bilden die zentrale Arbeitsakte. Die maschinenlesbaren Beispiele in 02_PRAXIS sind Dokumentationsbeispiele mit synthetischen Identitäten, keine ausführbare Sicherheitskonfiguration.

Ergebnis: Die entscheidende Schnittstelle erkennt nicht nur, welches System eine Anfrage stellt. Sie setzt durch, welche konkrete Handlung dieses System unter welchen Bedingungen ausführen darf. Diese Berechtigung wird von Dateninhalten, Modellverhalten und physischer Schutzfunktion sauber getrennt, sodass Fehler erkannt und ihre Wirkung begrenzt werden können.

<!-- GENERATED_LOCAL_BIBLIOGRAPHY -->

Literatur und Quellenstatus

Die folgenden Angaben gehören zu diesem Band. Das vollständige Quellenregister erläutert Verwendungszweck und Zugriff. Quellenstatus und Grenzen sind bei jeder Weiterverwendung zu erhalten. Die Q-Nummern im Text beziehen sich auf diese Einträge; Fundstellen bezeichnen die externe Quelle, nicht eine erfundene PDF-Seite dieses Bandes.

Q02 – Keith Stouffer et al. / NIST. Guide to Operational Technology (OT) Security, SP 800-82 Rev. 3. 2023-09-28. DOI: 10.6028/NIST.SP.800-82r3. [Originalquelle](#). Stand des Abrufs: 2026-09-25.

Fundstelle: §§ 2.3.7, 2.4, 3.3, 4.2.2, 5.2, 5.3, 6.2.10–6.2.12, 6.3–6.5; gedruckte S. 25, 28, 38–42, 55, 69–79, 122–138. Zugriff: Landingpage und relevante PDF-Abschnitte gelesen; PDF-Stichprobe visuell geprüft. Grenze: Freiwillige US-Leitlinie, kein deutscher Konformitätsnachweis. Der Leitfaden ergänzt eigene Überlegungen; Kapitelverweise sind keine wörtlichen Normanforderungen.

Q06 – Scott Rose; Oliver Borchert; Stu Mitchell; Sean Connelly / NIST. Zero Trust Architecture, SP 800-207. 2020-08. DOI: 10.6028/NIST.SP.800-207. [Originalquelle](#). Stand des Abrufs: 2026-09-25.

Fundstelle: Abstract; ressourcenbezogenes Vertrauen und getrennte Authentisierung/Autorisierung. Zugriff: Offizielle Publikationsseite / Abstract gelesen. Grenze: Zero Trust garantiert weder Prozesssicherheit noch korrekte Daten. Die vorgeschlagene Agentenarchitektur ist eigene Synthese.

Q09 – NIST NCCoE. Software and AI Agent Identity and Authorization. Projektstand abgerufen 2026-09-25. [Originalquelle](#). Stand des Abrufs: 2026-09-25.

Fundstelle: Project overview; Identity and Authorization. Zugriff: Projektseite gelesen. Grenze: Projektseite ist kein verabschiedeter Standard und keine gesetzliche Pflicht.

Q10 – NIST Cybersecurity Insights. Back to the Future: Why Agentic AI Needs a Strong Identity Foundation. 2026-08-27. [Originalquelle](#). Stand des Abrufs: 2026-09-25.

Fundstelle: Agent identities; credentials; authorization; consent fatigue. Zugriff: Webtext gelesen. Grenze: Fachbeitrag, keine Norm. Zustimmungsermüdung ist hier qualitativ behandelt, nicht als populationsweite Effektgröße.

Q11 – Apostol Vassilev et al. / NIST. Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations, AI 100-2e2025. 2025-03. DOI: 10.6028/NIST.AI.100-2e2025. [Originalquelle](#). Stand des Abrufs: 2026-09-25.

Fundstelle: Abstract; Attack goals, capabilities and knowledge; Errata. Zugriff: Publikationsseite / Abstract einschließlich Erratahinweis gelesen. Grenze: Taxonomie belegt keine Häufigkeit; keine Aussage, dass jede beschriebene Angriffsklasse im betrachteten System praktisch funktioniert.

Q14 – M. Jones et al. / IETF. OAuth 2.0 Token Exchange, RFC 8693. 2020-01. DOI: 10.17487/RFC8693. [Originalquelle](#). Stand des Abrufs: 2026-09-25.

Fundstelle: § 1.1 Delegation vs. Impersonation; § 4.1 act claim. Zugriff: RFC-Webvolltext / Delegationsabschnitte gelesen. Grenze: Token Exchange allein begrenzt keine physische Wirkung und erzeugt keine vollständige Verantwortlichkeitskette.

Q15 — T. Lodderstedt et al. / IETF. OAuth 2.0 Token Revocation, RFC 7009. 2013-08. DOI: 10.17487/RFC7009. [Originalquelle](#). Stand des Abrufs: 2026-09-25.

Fundstelle: § 2.1–2.2 Revocation Request / Response; § 3 Implementation Note. Zugriff: RFC-Webvolltext / Widerrufsabschnitte gelesen. Grenze: Ein Tokenwiderruf hebt bereits ausgeführte oder autonom weiterlaufende Maschinenaktionen nicht auf. Diese Folgerung ist eigene Systemanalyse.

Q22 — Medical Device Coordination Group / Europäische Kommission. MDCG 2019-16 rev. 1: Guidance on Cybersecurity for medical devices. 2020-07 (Rev. 1). [Originalquelle](#). Stand des Abrufs: 2026-09-25.

Fundstelle: Abschnitte zu Design / Post-market; Annex IV, gedruckte S. 46. Zugriff: Relevante PDF-Abschnitte gelesen; Inhaltsverzeichnis und Annex IV visuell geprüft. Grenze: Leitlinie ist ausdrücklich nicht rechtsverbindlich; kein Ersatz für MDR/IVDR oder den Normvolltext.

Q38 — Europäische Union. Verordnung (EU) 2016/679 — Datenschutz-Grundverordnung. 2016; aktuelle Fassung vor Anwendung prüfen. [Originalquelle](#). Stand des Abrufs: 2026-09-25.

Fundstelle: Art. 5, 6, 9, 25, 28, 30, 32–35, 38–39, 44 ff.. Zugriff: Direkter Web-Volltextabruf technisch blockiert; nicht als neu volltextverifiziert ausgewiesen. Grenze: Artikelzuordnung dient der rechtlichen Prüfung. Redaktionell bekannte Regelungsinhalte werden nicht als vollständig neu geprüfte konsolidierte Rechtsberatung ausgegeben.

Arbeitsmittel: Kontrollkatalog, Formularbuch, Fallstudien und Claim–Evidence-Matrix.

12 Rechtliche Hinweise

12.1 Keine Beratung

Diese Veröffentlichung ist ein redaktionelles Erzeugnis. Sie stellt keine Rechtsberatung, keine medizinische, pflegerische, arbeitsrechtliche oder datenschutzrechtliche Beratung und keine Beschaffungs- oder Vergabeberatung dar. Sie ersetzt insbesondere nicht die Prüfung durch Rechtsanwaltschaft, Datenschutzaufsicht, Medizinprodukteberatung, Vergabestelle, Ethikkommission nach Medizinprodukterecht oder Arbeitnehmervertretung. Wer auf Grundlage dieses Dokuments entscheidet, entscheidet auf eigene Verantwortung und hat den Sachverhalt vorher eigenständig fachlich prüfen zu lassen.

12.2 Keine Zusicherung, keine Konformitätsaussage

Die beschriebenen Verfahren, Prüfstufen und Checklisten sind redaktionelle Vorschläge. Sie sind keine Zusicherung einer Eigenschaft, kein garantiertes Ergebnis und kein Nachweis der Einhaltung einer Rechtsvorschrift. Wer ein Verfahren aus diesem Dokument anwendet, erfüllt damit keine gesetzliche Anforderung; ob eine Anforderung erfüllt ist, entscheidet die zuständige Stelle am Einzelfall.

12.3 Rechtsstand

Rechtsstände, Fristen und Fundstellen beziehen sich auf den im Dokumentkopf genannten Tag. Der KI-Rechtsrahmen wird laufend geändert; ein älterer Fristenplan darf nicht unverändert weiterverwendet werden. Genannte Gesetze und Verordnungen sind in ihrer jeweils geltenden Fassung heranzuziehen.

12.4 Haftung

Die Inhalte wurden mit Sorgfalt erstellt und, soweit im Methodenteil beschrieben, gegengeprüft. Für Richtigkeit, Vollständigkeit und Aktualität wird gleichwohl keine Gewähr übernommen. Eine Haftung des Herausgebers ist auf Vorsatz und grobe Fahrlässigkeit beschränkt; dies gilt nicht bei der Verletzung des Lebens, des Körpers oder der Gesundheit sowie bei zwingender gesetzlicher Haftung.

12.5 Verweise auf fremde Inhalte

Dieses Dokument verweist auf Inhalte Dritter, insbesondere auf Behörden-, Fachgesellschafts-, Normen- und Fachverlagsseiten. Auf deren Gestaltung und Inhalt hat der Herausgeber keinen Einfluss; für sie ist jeweils der dortige Anbieter verantwortlich. Die Adressen waren zum im Literaturverzeichnis genannten Prüfzeitpunkt erreichbar. Die Nennung eines Produkts, Herstellers oder Dienstes ist weder eine Empfehlung noch eine Bewertung seiner Eignung.

12.6 Marken und Produktbezeichnungen

Genannte Produkt-, Marken- und Unternehmensbezeichnungen sind Kennzeichen ihrer jeweiligen Inhaber und werden ausschließlich zur Beschreibung des Sachverhalts verwendet. Aus der Nennung folgt keine Verbindung zwischen dem Herausgeber und dem jeweiligen Unternehmen und keine Aussage über Zulassungs-, Konformitäts- oder Zertifizierungsstatus.

12.7 Urheberrecht und Nutzung

Dieses Dokument, sein Text, seine Tabellen und seine Abbildungen sind urheberrechtlich geschützt. Rechteinhaberin ist die WebTrades GmbH. Zulässig sind das Lesen, das Speichern und der Ausdruck für den eigenen Gebrauch sowie das Zitieren einzelner Stellen mit vollständiger Quellenangabe nach der im Impressum genannten Zitierweise. Die Vervielfältigung, Bearbeitung, Verbreitung, öffentliche Zugänglichmachung oder kommerzielle Verwertung des Dokuments oder wesentlicher Teile daraus bedarf der vorherigen schriftlichen Zustimmung der Rechteinhaberin. Das gilt ausdrücklich auch für die Übernahme von Tabellen und Abbildungen.

12.8 Keine personenbezogenen Daten

Für dieses Dokument wurden keine personenbezogenen Daten verarbeitet. Es beruht auf veröffentlichter Literatur, amtlichen Rechtsquellen und Veröffentlichungen von Fachgesellschaften; es wurden weder Beschäftigte befragt noch Patientendaten verwendet. Wer die hier beschriebenen Verfahren auf eine reale Einrichtung anwendet, verlässt diesen Rahmen: Die Auswertung von Tätigkeits-, Alarm- oder Quittierungsdaten unterliegt der Datenschutz-Grundverordnung und in der Regel der Mitbestimmung.

12.9 Unabhängigkeit

Dieses Dokument ist ohne Auftrag, ohne Finanzierung und ohne inhaltliche Beteiligung eines Herstellers, Krankenhausträgers, Kostenträgers, Verbands oder Beratungsunternehmens entstanden. Es besteht kein Vergütungs- oder Beteiligungsverhältnis zu einem der genannten Unternehmen oder Institutionen. Genannt werden Systeme, Anbieter und Dokumente, weil sie in den beschriebenen Gegenstand fallen, nicht weil sie beworben werden sollen.

12.10 Anwendbares Recht

Es gilt das Recht der Bundesrepublik Deutschland unter Ausschluss des UN-Kaufrechts. Gesetzliche Verbraucherschutzvorschriften bleiben unberührt.