

Leitfaden: Daten sicher für KI vorbereiten

Anonymisieren, pseudonymisieren und synthetische Daten nutzen

Ziel: Daten so vorzubereiten, dass personenbezogene und vertrauliche Informationen bei der Nutzung von KI Tools möglichst geschützt werden. Anonymisierung ist dabei nicht dasselbe wie Pseudonymisierung. Manchmal ist es sinnvoller, echte Daten gar nicht zu verwenden, sondern synthetische Daten.

⚠ Wichtig:

Datenaufbereitung ersetzt nicht die Prüfung, ob Informationen überhaupt in das verwendete KI System eingegeben werden dürfen. Interne Vorgaben, Vertraulichkeit, Geschäftsgeheimnisse, Verträge und Datenschutz bleiben relevant.

1. Drei Begriffe, die du unterscheiden solltest

Verfahren	Was passiert?	Personenbezug?
Anonymisierung	Eine Person ist mit vernünftigerweise einsetzbaren Mitteln nicht mehr identifizierbar. Auch Kombinationen und Kontext dürfen keine realistische Zuordnung ermöglichen.	Nein, wenn die Anonymisierung tatsächlich wirksam ist.
Pseudonymisierung	Direkte Identifikatoren werden ersetzt, zum Beispiel „Max Mustermann“ durch „Kunde A“. Eine Zuordnung kann aber weiterhin möglich sein.	Ja. Die Daten bleiben grundsätzlich personenbezogen.
Synthetische Daten	Künstlich erzeugte Daten bilden typische Eigenschaften oder Strukturen nach, ohne einen realen Einzelfall verwenden zu müssen.	Nicht automatisch. Es muss geprüft werden, ob reale Personen rekonstruierbar oder wiedererkennbar sind.

2. Anonymisierung: Daten sind wirklich nicht mehr zuordenbar

Bei einer echten Anonymisierung reicht es nicht, nur den Namen zu löschen. Entscheidend ist, ob eine Person direkt oder indirekt noch identifiziert werden kann. Einzelne Angaben können harmlos wirken, in Kombination aber eindeutig werden.

Beispiel: „57 Jahre, einziger Niederlassungsleiter in Standort X, Eintritt 02.05.2024“ kann eine Person auch ohne Namen erkennbar machen.

Typische Maßnahmen

- Direkte Identifikatoren entfernen.
- Seltene oder sehr genaue Merkmale verallgemeinern.
- Genaue Datumsangaben und Orte gegebenenfalls gröber machen.
- Freitexte auf verräterische Kontextinformationen prüfen.
- Prüfen, ob öffentliches oder internes Zusatzwissen eine Zuordnung ermöglicht.

Merksatz:

Name weg bedeutet noch lange nicht anonym.

3. Pseudonymisierung: Identität ersetzen, aber nicht beseitigen

Bei der Pseudonymisierung werden identifizierende Angaben ersetzt. Aus „Anna Beispiel, Muster GmbH“ kann etwa „Person A, Unternehmen B“ werden. Das reduziert Risiken und kann für viele Arbeitsprozesse sinnvoll sein. Es ist aber keine Anonymisierung, wenn die Person über eine Zuordnungstabelle, weitere Daten oder den Kontext wieder ermittelt werden kann.

Geeignete Platzhalter: [Person A], [Unternehmen B], [Ort C], [Projekt X], [Datum 1]. Eine vorhandene Zuordnungsliste darf nicht zusammen mit den pseudonymisierten Daten an das KI Tool übermittelt werden.

Merksatz:

Pseudonymisierte Daten bleiben grundsätzlich personenbezogene Daten.

4. Was solltest du vor der KI Nutzung prüfen?

Prüfbereich	Beispiele
Direkte Identifikatoren	Name, Adresse, E-Mail, Telefonnummer, Nutzername
Kennnummern	Kunden-, Personal-, Vertrags-, Rechnungsnummern, IBAN, interne IDs
Besonders sensible Inhalte	Gesundheitsdaten, Diagnosen, biometrische oder andere besonders geschützte Angaben
Indirekte Identifikatoren	Alter, Funktion, genauer Standort, seltene Merkmale, exakte Zeitpunkte
Unternehmensbezug	Projektnamen, Kundennamen, interne Produktnamen, vertrauliche Vorgänge
Technische Daten	IP-Adressen, Gerätekennungen, Dateinamen, Pfade, Metadaten
Freitext	Signaturen, Zitate, Fallbeschreibungen, Kommentare und andere Kontextinformationen

5. Synthetische Daten: echte Fälle oft gar nicht nötig

Synthetische Daten werden künstlich erzeugt, damit Struktur, Größenordnung oder typische Eigenschaften realer Daten erhalten bleiben, ohne einen konkreten realen Datensatz verwenden zu müssen. Für Übungen, Prompttests, Analysen, Schulungen oder Prototypen ist das häufig die sauberste Lösung.

Beispiel: Statt eine echte Kundenliste mit Name, Umsatz, Bestellhistorie und Beschwerden hochzuladen, erzeugst du 100 fiktive Kunden mit realistischen Umsatzklassen, Kaufhäufigkeiten und Beschwerdekategorien. Die KI kann damit dieselbe Analyseaufgabe bearbeiten, ohne die echten Kundendaten zu benötigen.

Aber: Synthetisch heißt nicht automatisch anonym. Werden synthetische Daten aus echten Datensätzen erzeugt, muss geprüft werden, ob seltene Datensätze zu genau übernommen wurden oder Rückschlüsse auf reale Personen möglich sind. Einfach Namen auszutauschen und den Rest unverändert zu lassen, erzeugt keine verlässlich synthetischen Daten.

 **Regel:**

Wenn echte Daten für die Aufgabe nicht erforderlich sind, nutze lieber synthetische.

6. Word, Excel, PDF und Screenshots

Word und Excel

Vor dem Export Dokumenteigenschaften, Kommentare, Änderungsverfolgung, Autorinformationen und versteckte Inhalte prüfen.

In Microsoft Office steht dafür die Dokumentprüfung unter „Datei > Informationen > Auf Probleme überprüfen > Dokument überprüfen“ zur Verfügung.

Excel zusätzlich

Versteckte Tabellenblätter, ausgeblendete Zeilen und Spalten, Formeln, Verknüpfungen, Kommentare und benannte Bereiche kontrollieren. Sichtbar gelöscht bedeutet nicht zwingend aus der Datei entfernt.

PDF

Einen schwarzen Balken über Text zu legen ist keine sichere Schwärzung. Der darunterliegende Text kann je nach Verfahren weiterhin vorhanden, markierbar oder extrahierbar sein. Für echte Schwärzungen eine Redaktionsfunktion verwenden, die den zugrunde liegenden Inhalt entfernt.

Screenshots

Nicht nur den eigentlichen Inhalt prüfen. Browser Tabs, Profilbilder, Benachrichtigungen, Dateipfade, URLs, E-Mail-Adressen und andere Bildbereiche können ebenfalls Informationen verraten.

7. Metadaten nicht vergessen

Dateien können Informationen enthalten, die auf der sichtbaren Seite nicht erscheinen: Autor, Firma, Bearbeiter, Erstellungszeit, Kommentare, Versionsinformationen oder GPS Daten. Prüfe diese Informationen lokal, bevor du eine Datei weitergibst. Lade vertrauliche Dateien nicht nur zur Metadatenprüfung bei einem beliebigen Online Dienst hoch.

8. Die 5 Fragen vor dem Upload

1. Enthält die Datei personenbezogene oder besonders sensible Daten?
2. Kann jemand auch ohne Namen aus Kontext und Merkmalskombinationen erkannt werden?
3. Habe ich anonymisiert oder lediglich pseudonymisiert?
4. Enthält die Datei vertrauliche Unternehmensinformationen oder Geschäftsgeheimnisse?
5. Brauche ich überhaupt echte Daten oder reichen synthetische Daten?

Wenn du eine dieser Fragen nicht sicher beantworten kannst, verwende die Daten nicht unverändert in einem KI Tool.

9. Kurzentcheidung

Situation	Vorgehen
Echte Personen sind für die Aufgabe irrelevant.	Synthetische Daten bevorzugen.
Struktur realer Daten ist nötig, Identität aber nicht.	Anonymisierung prüfen oder synthetische Daten mit vergleichbarer Struktur verwenden.
Eine spätere Zuordnung zu Personen muss möglich bleiben.	Pseudonymisierung kann geeignet sein. DSGVO und interne Vorgaben gelten weiter.
Vertrauliche oder besonders sensible Inhalte sind enthalten.	Nicht einfach hochladen. Freigabe, Rechtsgrundlage, Toolkonfiguration und interne Regeln prüfen.

💡 Merke:

Name ersetzt = pseudonymisiert.

Nicht mehr identifizierbar = anonymisiert.

Gar kein echter Datensatz = häufig die bessere Lösung.

Stand: September 2026.

Fachlicher Hinweis: Der Leitfaden dient der praktischen Orientierung und ersetzt keine rechtliche Einzelfallprüfung.