UNIVERSITÄT MANNHEIM

Risiken bei der Nutzung von KI-Systemen

Merkblatt für Nutzende

Version 1.0 vom 25.08.2025
Dokumentenklassifizierung: TLP green – intern

Ansprechperson: Prof. Dr. Heiko Paulheim
E-Mail: ki.team@uni-mannheim.de

Freigabe und Änderungen

Freigabehistorie

Version	Abnahme/Freigabe durch	Datum
1.0	Heiko Paulheim	25.08.2025

Änderungshistorie

Version	Bearbeiter	Änderungsinformation	Datum
1.0	Jonas Brosig	Initiale Erstellung	25.08.2025

Dokumentenklassifizierung mit TLP

Dieses Dokument ist mit dem Traffic Light Protocol (TLP) als „TLP green – intern“ klassifiziert. Diese Angabe ist auch auf dem Deckblatt zu finden und dort nach der u. a. Tabelle farblich hervorgehoben.

TLP ermöglicht den Autorinnen und Autoren eines Dokumentes die Bedingungen für dessen Weitergabe zu regeln und so die Sicherheit zu erhöhen. Als Empfängerin oder Empfänger dieses Dokumentes müssen Sie die auf dem Deckblatt getroffene Klassifizierung einhalten. Eine genaue Erklärung der Weitergaberegelungen finden Sie in der folgenden Tabelle.

TLP-Klassifizierung	Weitergaberegelung
TLP clear – öffentlich	Dieses Dokument enthält keine vertraulichen Informationen und kann von urheberrechtlichen Aspekten abgesehen ohne Einschränkung weitergegeben und öffentlich zugänglich gemacht werden.
TLP green – intern	Dieses Dokument enthält Informationen, die für den dienstlichen Gebrauch notwendig sind. Es darf an Partner der Universität weitergeben, jedoch nicht veröffentlicht werden.
TLP amber – vertraulich (…)	Dieses Dokument enthält vertrauliche Informationen und darf daher nur einem begrenzten und zuvor definierten Personenkreis (z. B. UNIT, UB, Lehrstuhl X) weitergegeben werden. Eine Weitergabe an Dritte ist nur möglich, wenn der Dritte das Dokument zur Arbeitserfüllung benötigt und ihm diese TLP-Klassifizierung bekannt ist. Der definierte Personenkreis wird bei der Klassifizierung in Klammer ergänzt.
TLP red – streng vertraulich (…)	Dieses Dokument enthält streng vertrauliche Informationen, die nur einem begrenzten und zuvor definierten Personenkreis, meist auch Teilnehmerkreis einer Besprechung, Konferenz oder schriftlichen Korrespondenz (z. B. Rektorat) bereitgestellt werden darf. Eine Weitergabe ist untersagt. Der definierte Personenkreis wird bei der Klassifizierung in Klammer ergänzt.

Inhaltsverzeichnis

Tabellenverzeichnis

1 Allgemeine Hinweise zu Risiken bei der Nutzung von KI-Systemen

KI-Systeme sind immer mehr in der Lage, Sie bei Ihren täglichen Aufgaben zu unterstützen und Ihnen zu helfen, diese schneller und besser zu erledigen. Die Universität Mannheim begrüßt diese Entwicklungen und möchte ihre Mitarbeitenden ausdrücklich zum Einsatz entsprechender Systeme ermutigen.

Bitte beachten Sie dabei, dass mit der Nutzung von KI-Systemen auch Risiken verbunden sind. Verwenden Sie KI-generierte Informationen stets umsichtig und unterziehen sie diese vor Weitergabe einer gewissenhaften Prüfung.

Für einen verantwortungsvollen Einsatz müssen Nutzungsregeln sowie Vorgaben der Anbieter beachtet werden.

Supporthinweis:

Bei allgemeinen Rückfragen zur Nutzung der KI-Systeme wenden Sie sich gerne per E-Mail an das KI-Kernteam der Universität: ki.team@uni-mannheim.de.

Bei Rückfragen betreffend didaktische Aspekte wenden Sie sich bitte an: elearning@uni-mannheim.de.

2 Risiken für Nutzende

Selbst bei einer ordnungsgemäßen Nutzung von KI-Systemen können unerwünschte Nebenwirkungen nicht ausgeschlossen werden.

Sie finden eine Liste mit Handlungsempfehlungen zu den dargestellten Fällen im Anhang zu diesem Dokument. Die Liste soll Sie dabei unterstützen, je nach Fall angemessen zu reagieren.

Risiken
Unerwünschte Ausgaben und Bias: KI-Modelle können unerwünschte oder voreingenommene Inhalte generieren, die auf den Trainingsdaten basieren.
Fehlende Qualität und Faktizität: Die generierten Inhalte können fehlerhaft oder erfunden sein, was als „Halluzinieren“ bezeichnet wird.
Fehlende Aktualität: Modelle ohne Echtzeitzugriff können keine aktuellen Informationen liefern.
Fehlende Reproduzierbarkeit und Erklärbarkeit: Die Ausgaben sind oft nicht reproduzierbar und schwer nachvollziehbar.
Fehlende Sicherheit von generiertem Code: Generierter Code kann Sicherheitslücken enthalten.
Fehlerhafte Reaktion auf spezifische Eingaben: Kleine Änderungen in den Eingaben können zu großen Unterschieden in den Ausgaben führen.

Automation Bias: Nutzende könnten den generierten Inhalten zu viel Vertrauen schenken.
Der Trainingsdatensatz kann rechtswidrig verarbeitete personenbezogene Daten enthalten oder Rechte am geistigen Eigentum Dritter verletzen.
Die Ausgaben können rechtswidrig personenbezogene Daten enthalten oder Rechte am geistigen Eigentum Dritter verletzen.
Die ausgegebenen Inhalte können strafbar (z. B. beleidigend, volksverhetzend) sein.

1: Risiken bei ordnungsgemäßer Nutzung

3 Anhang: Hinweisliste mit *Handlungsempfehlungen*

Disclaimer: Die nachstehende Liste verfolgt einen allgemeinen Ansatz und erhebt keinen Anspruch auf Vollständigkeit.

1. Prompts klar und strukturiert formulieren
→ Beschreiben Sie Ihr Ziel, Ihre Zielgruppe und den gewünschten Stil präzise. Nutzen Sie klare Anweisungen, damit die KI besser versteht, was Sie benötigen.

2. Quellen prüfen und Fakten abgleichen
→ Verlassen Sie sich nicht blind auf KI-Antworten. Fordern Sie Quellen ein und überprüfen Sie Inhalte mit vertrauenswürdigen Datenbanken oder Webseiten.

3. Mehrere Sichtweisen einfordern
→ Bitten Sie die KI um alternative Erklärungen oder Gegenargumente, damit Sie keine einseitige Perspektive übernehmen.

4. Ergebnisse als KI-Ausgaben kennzeichnen
→ Machen Sie sich und anderen bewusst: Es handelt sich um Inhalte einer KI, nicht um gesicherte Fakten. Nutzen Sie sie als Unterstützung, nicht als endgültige Wahrheit.

5. Ergebnisse kritisch überprüfen (menschliche Validierung)
→ Prüfen Sie wichtige oder sensible Inhalte selbst oder lassen Sie sie von Fachleuten gegenlesen, bevor Sie diese weiterverwenden.

6. Auf Aktualität und Datenstand achten
→ Fragen Sie nach dem Wissensstand der KI und ergänzen Sie fehlende Informationen durch eigene Recherche, vor allem wenn Aktualität entscheidend ist.

7. Datenschutz beachten
→ Geben Sie keine sensiblen oder personenbezogenen Daten in die KI ein. Wenn Sie Inhalte veröffentlichen, achten Sie auf Anonymisierung und rechtliche Rahmenbedingungen.

8. Eigene Nutzung dokumentieren
→ Notieren Sie sich wichtige Eingaben und Ausgaben, damit Sie später nachvollziehen können, wie ein Ergebnis zustande gekommen ist.

9. Inhalte technisch überprüfen (bei Code oder Daten)
→ Wenn Sie Code generieren, testen Sie diesen in einer sicheren Umgebung und prüfen Sie auf Fehler oder Sicherheitslücken, bevor Sie ihn einsetzen.

10. Fachleute einbeziehen
→ Ziehen Sie Expert*innen hinzu, wenn es um komplexe Themen geht (z. B. Recht, Datenschutz, IT-Sicherheit), damit Sie fundierte Entscheidungen treffen.

11. Auf problematische Inhalte achten
→ Seien Sie aufmerksam: KI kann strafbare, diskriminierende oder unangemessene Inhalte erzeugen. Verwenden Sie solche Ausgaben nicht weiter. Sie können entsprechende Problem über vorhandene Feedback-Funktionen und an die Funktionsadresse *ki.team@uni-mannheim.de* melden. Bitte sichern Sie hierzu Prompt, Ausgabe sowie deren Zeitpunkt (Screenshot/Log).

12. Eigenes Bewusstsein schärfen und auf dem aktuellen Stand bleiben
→ Bedenken Sie: KI ist ein Hilfsmittel, kein Ersatz für kritisches Denken. Informieren Sie sich regelmäßig über Chancen und Risiken im Umgang mit KI (z. B. durch Teilnahme an Schulungen).