AI Ethical Risk Karten

Christof Wolf-Brenner
cbrenner@know-center.at

Controller Institut

Dipl.-Ing. Christof Wolf-Brenner, MA

Innovation Management
MA, Campus02

Computer Science
PhD, Graz University of Technology

IT & Business Informatics
BSc+Dipl.-Ing., Campus02

Philosophy
BA, University of Graz

08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24

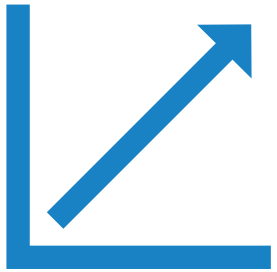
Accountant
Styria Media Group

SAP Consultant
AVL Group

Consultant
Know-Center GmbH

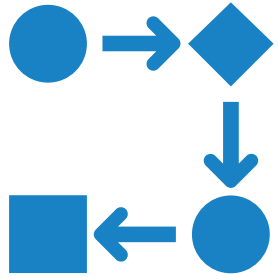
Wie, wann, wo, von wem, warum... werden ethische Risiken von KI-Systemen heute identifiziert?

Warum?



Rasanter Anstieg von KI-Nutzung

Einseitiger Fokus auf bekannte Risiken



Fehlende systematische Betrachtung

Ziel des Kartensets

- **Vielfalt ethischer Risiken sichtbar machen**, auch jenseits bekannter Themen (Transparenz, Bias, Datenschutz).
- **Interaktive Nutzung fördern**, um Teams zu einer strukturierten, tiefgehenden Reflexion und Diskussion ethischer Fragen anzuregen.
- **Vermittlung von Wissen über potenzielle ethische Risken**: Kartensets dieser Art ermöglichen es spielerisch Informationen zu vermitteln.
- **Setzung von Impulsen und Strukturierung**: Mittels Kartensets können in Workshopszenarien Impulse für Überlegungen gegeben und das Vorgehen strukturiert werden.
- **Anschlussfähigkeit** an Risk Management Standards

Theoretische Grundlagen des Kartensets

- AI, algorithmic and automation incidents and issues repository (AIAAIC)
- MIT AI Risk Repository
- Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz – KI Prüfkatalog (Fraunhofer)

Anmerkung: Es gibt unzählige Frameworks, Kataloge, Methoden, Datenbanken etc. zu diesem Thema. Diese drei wurden von mir im Rahmen des Projektes nach eingehender Recherche aufgrund der folgenden Kriterien ausgewählt.

- **Praxisnähe und Anwendungsorientierung**
- **Vielfalt und Abdeckung unterschiedlicher Perspektiven**
- **Glaubwürdigkeit und Akzeptanz**

Struktur des Kartensets

URSACHEN
(Auslöser)

- Verzerrungen in Daten
- Intransparente Modelle
- Fehlende Tests
- Rechtliche Grauzonen
- uVm.

WIRKUNGEN
(Schaden)

- Psychische Verletzung
- Finanzieller Schaden
- Manipulation
- Einschränkung der Handlungsfreiheit
- uVm

Struktur des Kartensets

URSACHEN UND WIRKUNGEN (SCHADEN)

1. Fördert systematisches Denken
2. Erleichtert präventives Handeln
3. Vermeidet verkürzte Perspektiven
4. Fördert interdisziplinäre Diskussionen
5. Unterstützt strukturiertes Mapping

Struktur des Kartensets

Schadenskarten (Wirkung)	9 Karten – zeigen mögliche Folgen auf, die durch den Einsatz von KI-Systemen entstehen können. Diese Karten beschreiben grundlegende Schadensarten, die Individuen, Organisationen, Gesellschaft oder Umwelt betreffen können – unabhängig davon, wodurch sie verursacht wurden.
Datenbezogene Ursachen	*10 Karten – thematisieren Probleme in Bezug auf Datenerhebung, -qualität oder -struktur.* Diese Karten zeigen, wie **Fehler, Verzerrungen oder Lücken in Datensätzen** entstehen und sich negativ auf die Leistung und Fairness von KI-Systemen auswirken können.

Struktur des Kartensets

Modellbezogene Ursachen	*10 Karten – beschreiben Risiken, die durch die Struktur, das Verhalten oder die Entwicklung von KI-Modellen entstehen.* Hier geht es um **Grenzen in der Modellarchitektur, fehlerhaftes Lernen oder unzureichende Kontrolle**.
Systembezogene Ursachen	*10 Karten – beziehen sich auf Probleme im Gesamt-KI-System, in dem Modelle eingebettet sind.* Diese Karten adressieren Schwächen in **Benutzerschnittstellen, Prozessen, Rollenverteilung oder Integration.**
Organisatorische Ursachen	*10 Karten – beleuchten strukturelle, kulturelle oder strategische Defizite innerhalb von Organisationen.* Sie zeigen, wie **interne Anreize, fehlende Verantwortung, mangelnde Business-Ethik oder einseitige Perspektiven** zu systematischen Fehlentwicklungen in KI-Anwendungen führen können.

Beispiel Kartenaufbau (Schadenskarten)

Titel

Beschreibung

Karten-ID

Kategoriensymbol

Beispiel Kartenaufbau (Schadenskarten)

Titel

Beschreibung

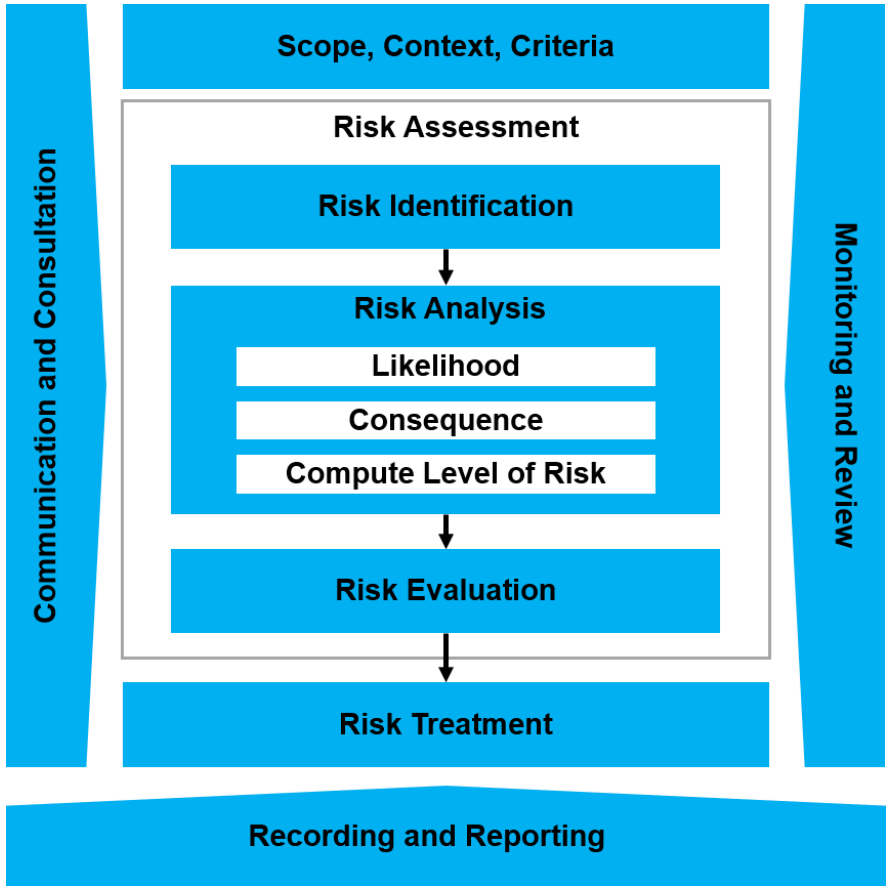
Karten-ID

Beispiel

Kategoriensymbol

Vorgehensweise im Risikomanagement

ISO 31000 Risk management process



RISIKOIDENTIFIKATION

Die Karten bieten strukturelle Inspiration.

- Ursachenkarten: decken typische Schwachstellen in Daten, Modellen, Systemen und Organisationen ab.
- Schadenskarten: machen die potenziellen Wirkungen greifbar.

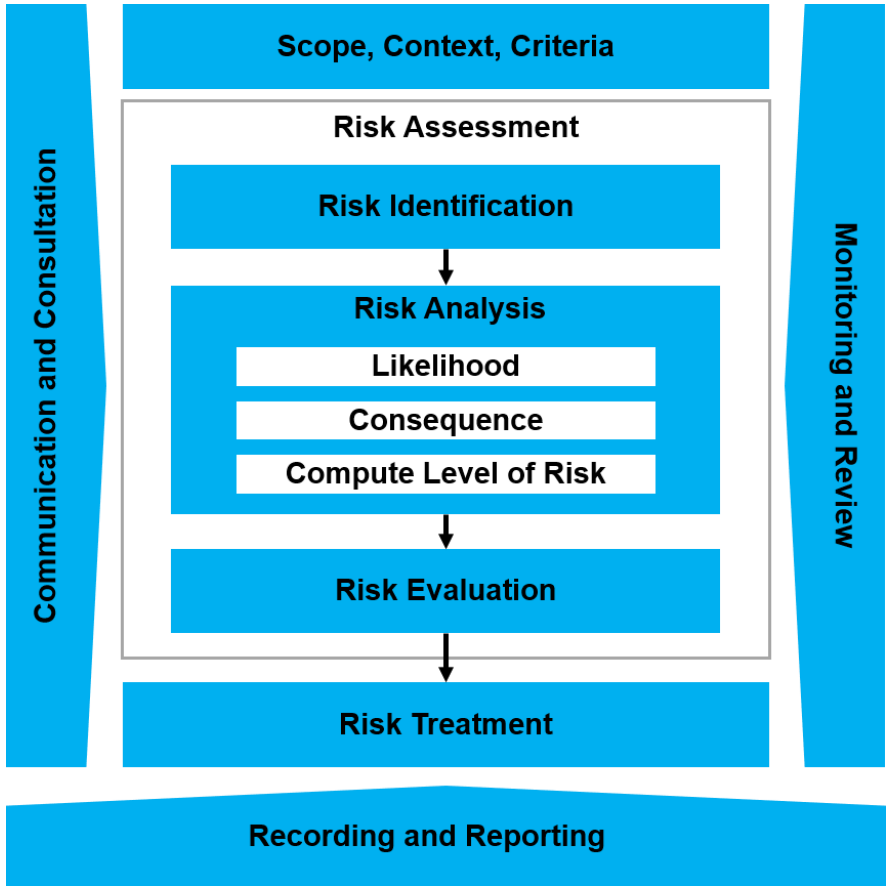
RISK ANALYSIS

Die Karten erleichtern eine gezielte Bewertung.

- Was ist wahrscheinlich? Fokus auf häufige oder wahrscheinliche Ursachen.
- Was ist kritisch? Fokus auf besonders schwere Auswirkungen.

Vorgehensweise im Risikomanagement

ISO 31000 Risk management process



RISK EVALUATION

Die Kartenpaare helfen bei der Priorisierung.

- Kombinationen aus häufiger Ursache & schwerwiegender Wirkung markieren **superkritische Risiken**.

RISK ANALYSIS

Die Karten liefern konkrete Behandlungsansätze.

- Ursachenkarten zeigen: **Was muss ich beheben, reduzieren oder kontrollieren?**

→ MIRO

Feedback und Evaluierung

20250328 - Methodenevaluierung
AI Ethics Risk Cards

https://forms.office.com/e/12c4Hw51mM