KI & ETHIK KARTENSET

Know-Center GmbH | Research Center for Data-Driven Business and Big Data Analytics | Sandgasse 34/2 | 8010 Graz | Austria | data-innovation@know-center.at | Tel.+43 316 873 30800 | Fax+43 316 873 1030800 | Firmenbuchgericht Graz | FN 199 685 f | UID: ATU 50367703 | http://www.know-center.at |

1 ZIELSETZUNG

Die Entwicklung und der Einsatz von KI-Systemen nehmen rasant zu – in immer mehr gesellschaftlichen Bereichen und mit wachsender Wirkungstiefe. Doch in vielen Organisationen fehlt bislang ein strukturierter Zugang zu den damit verbundenen ethischen Risiken. Dieses Kartenset wurde entwickelt, um genau hier anzusetzen – mit drei zentralen Ausgangspunkten:

- **Rasanter Anstieg von KI-Nutzung:** Immer mehr Entscheidungen werden durch KI-Systeme unterstützt oder automatisiert – oft ohne dass die Folgen für Einzelne, Gruppen oder Gesellschaften umfassend geprüft werden.
- **Einseitiger Fokus auf bekannte Risiken:** Technische Risiken wie Bias oder Intransparenz sind mittlerweile gut dokumentiert – doch viele ethische Gefahren liegen tiefer: in Organisationen, Prozessen, Zielkonflikten oder systemischen Dynamiken.
- **Fehlende systematische Betrachtung:** Bisherige Risikoanalysen sind häufig unsystematisch, fragmentiert oder nicht teamübergreifend anschlussfähig. Es fehlt an Werkzeugen, um Risiken gemeinsam, nachvollziehbar und ganzheitlich zu reflektieren.

Das Kartenset zielt darauf ab, eine *Vielfalt ethischer Risiken sichtbar zu machen, die über die bekannten Themen wie Transparenz, Bias und Datenschutz hinausgeht*. Es eröffnet neue Perspektiven auf mögliche Gefahren, die durch KI-Systeme entstehen können, und stellt diese in einen breiteren Kontext.

Ein weiteres Ziel ist die *interaktive Nutzung des Sets* zu fördern, um Teams zu einer strukturierten und tiefgehenden Reflexion sowie Diskussion ethischer Fragestellungen anzuregen. Durch den interaktiven Charakter wird das Bewusstsein für die Komplexität und die Vielzahl von Risiken geschärft. Das Kartenset dient auch der *Vermittlung von Wissen*: Es stellt spielerisch und dennoch fundiert Informationen über potenzielle ethische

Risiken zur Verfügung, sodass die Nutzer*innen sich auf einfache und verständliche Weise damit auseinandersetzen können.

Darüber hinaus gibt es den *Impuls und die Struktur für die gezielte Auseinandersetzung mit ethischen Fragen*. In Workshopszenarien können durch die Karten neue Denkanstöße gegeben und der Reflexionsprozess organisiert werden. Nicht zuletzt bietet das Kartenset eine *Anschlussfähigkeit an etablierte Risk-Management-Standards*, wodurch es sich sowohl für erfahrene Risiko-Manager als auch für Neueinsteiger in das Thema eignet.

Das Set richtet sich an interdisziplinäre Teams in der Entwicklung, Anwendung und Regulierung von KI-Systemen – insbesondere an Entwickler:innen, Produktverantwortliche, Ethikbeauftragte, Forscher:innen sowie Fachleute aus Recht, Compliance und Management.

2 INHALT DES KARTENSETS

Die Inhalte des Kartensets basieren auf einer fundierten Auswahl an Quellen, die ethische Risiken von KI-Systemen systematisch erfassen und klassifizieren. Im Rahmen des Projekts wurden drei Referenzen besonders berücksichtigt:

- **AI, Algorithmic and Automation Incidents and Issues Repository (AIAAIC)**[1]
- **MIT AI Risk Repository**[2]
- **Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz – KI-Prüfkatalog (Fraunhofer)**[3]

[1] https://www.aiaaic.org/aiaaic-repository

[2] https://airisk.mit.edu/

[3] https://www.iais.fraunhofer.de/de/publikationen/studien/2021/ki-pruefkatalog.html

Diese Quellen wurden nicht zufällig gewählt. Sie zeichnen sich durch ihre *Praxisnähe, die Anwendungsorientierung ihrer Inhalte, die Vielfalt der abgedeckten Perspektiven sowie durch ihre Glaubwürdigkeit und breite Akzeptanz* in Fachkreisen aus.

Dabei ist zu betonen: Es existieren zahlreiche weitere Frameworks, Leitlinien, Methoden und Datenbanken zu ethischen Aspekten von KI. Die hier gewählten Quellen stehen exemplarisch für die aktuelle Praxis – und erlauben gleichzeitig eine gute Übertragbarkeit in konkrete Workshop- und Anwendungskontexte.

Das Kartenset gliedert sich in zwei zentrale Kategorien: *Ursachenkarten* und *Schadenskarten*. Diese zweigeteilte Struktur folgt der Logik: *Was kann schiefgehen – und was sind die Folgen?*

Ursachenkarten (40 Stück)

Die Ursachenkarten stehen für potenzielle *Auslöser* ethischer Risiken. Sie helfen, systematisch nach Schwachstellen in Daten, Modellen, Systemarchitekturen oder organisatorischen Rahmenbedingungen zu suchen. Die Ursachen sind in vier Untergruppen zu je zehn Karten unterteilt: *Datenbezogene, Modellbezogene, Systembezogene sowie Organisatorische Ursachen.*

- **Datenbezogene Ursachen** 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 von KI-Systemen auswirken können.
- **Modellbezogene Ursachen** beschreiben Probleme, 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** 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** 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.

Abbildung 1: Aufbau der Ursachenkarten

Schadenskarten (9 Stück)

Die Schadenskarten beschreiben grundlegende Schadensarten, die Individuen, Organisationen, Gesellschaft oder Umwelt betreffen können – unabhängig davon, wodurch sie verursacht wurden. Sie zeigen mögliche Folgen auf, die durch den Einsatz von KI-Systemen entstehen können.

Abbildung 2: Aufbau der Schadenskarten

3 ANWENDUNGSKONTEXT

Das Kartenset ist für den Einsatz in interdisziplinären Teams konzipiert und eignet sich besonders für Workshops, Projektreviews, Ethik-Boards oder Risikoanalysen im Rahmen von KI-Entwicklung, -Einsatz oder -Beschaffung. Es ist sowohl in frühen Projektphasen (z. B. bei der Systemkonzeption) als auch in späteren Evaluations- oder Auditprozessen einsetzbar. Das Set orientiert sich am ISO 31000 Risk Management Process und unterstützt zentrale Phasen dieses Prozesses:

1. **Risk Identification:** Die Ursachenkarten unterstützen bei der strukturierten Sammlung typischer Schwachstellen in Daten, Modellen, Systemen und Organisationen. Schadenskarten machen mögliche Auswirkungen sichtbar.
2. **Risk Analysis:** Die Karten erleichtern eine Bewertung von Risiken. Durch die Kombination aus Ursache und Wirkung können Risikoausmaß und Handlungsbedarf abgeschätzt werden.

- **Risk Evaluation:** Kartenpaare mit hoher Eintrittswahrscheinlichkeit und schwerer Auswirkung markieren hochkritische Risiken. So lassen sich Prioritäten für Maßnahmen ableiten.
- **Risk Treatment:** Karten liefern konkrete Impulse zur Ableitung von Handlungsansätzen: Was muss reduziert, überwacht oder organisatorisch geregelt werden, um eine gewisse Ursache abzuschwächen oder auch zu eliminieren.

Das Kartenset ist bewusst niedrigschwellig gestaltet – es kann ohne spezielle Vorkenntnisse verwendet werden, lässt sich aber auch gut mit bestehenden Risiko- oder Governance-Methoden verzahnen. Besonders wirkungsvoll ist es in moderierten Formaten mit kollaborativem Charakter.

4 NUTZUNGSEMPFEHLUNG

Das Kartenset ist flexibel einsetzbar – es eignet sich für Einzelpersonen ebenso wie für Teams, für kurze Impulsformate ebenso wie für mehrstündige Workshops. Je nach Ziel und Kontext kann es frei kombiniert, selektiv eingesetzt oder mit anderen Methoden (z. B. Impact Assessment, Checklisten, Ethik-Canvas) ergänzt werden. Für eine systematische und zielführende Anwendung wird folgendes Vorgehen in einem Workshop-Setting empfohlen.

4.1 Systemverständnis schaffen (20 min)

Ziel: Gemeinsames Bild vom betrachteten KI-System, seinem Zweck, Kontext und den betroffenen Stakeholdern entwickeln.

Kernfragen:

- Wozu dient das System, wer nutzt es, wer ist betroffen?
- Welche Daten fließen ein, welche Entscheidungen werden getroffen?
- Welche Akteure sind beteiligt oder potenziell gefährdet?

4.2 Schadenstypen priorisieren (15 min)

Ziel: Relevante ethische Schäden mittels der 10 Schadenskarten identifizieren und in Bezug auf zentrale Stakeholder priorisieren.

Kernfragen:

- Welche Arten von Schäden wären besonders schwerwiegend?
- Für wen wären diese Folgen kritisch? (z. B. Nutzer:innen, Gesellschaft)
- Welche Schäden müssen unbedingt vermieden werden?

Methode: Auswahl durch Karten-Sichtung, Diskussion oder Voting.

4.3 Ursachen identifizieren (20 min)

Ziel: Wahrscheinliche Ursachen im Kontext des betrachteten KI-Systems und dessen Einsatzkontext mittels der 40 Ursachenkarten identifizieren

Kernfragen:

- Wo liegen die Schwachstellen im aktuellen System?
- Welche Ursachen könnten auftreten oder sind vielleicht sogar besonders wahrscheinlich?
- Welche treten in der Praxis häufig auf?

Methode: Auswahl und/oder Priorisierung relevanter Ursachenkarten

4.4 Ursache-Wirkung-Pfade bilden (20 min)

Ziel: Sichtbarmachen der Zusammenhänge zwischen Ursachen und Schadenstypen.

Kernfragen:

- Welche Ursache(n) kann/können zu welchem Schaden führen?
- Welche Ketten oder Verknüpfungen sind plausibel oder wahrscheinlich?
- Gibt es Kombinationen, die besonders besorgniserregend sind?

Methode: Karten zusammenlegen oder auf Whiteboard verbinden (z. B. mit Linien/Pfeilen).

4.5 Risiken bewerten (15 min)

Ziel: Kombinationen aus Ursache & Schaden bewerten und priorisieren.

Kernfragen:

- Welche Kombination ist besonders wahrscheinlich und zugleich besonders gravierend?
- Welche Risiken gelten im Gesamtbild als kritisch?
- Welche Risiken sollten unbedingt adressiert werden?

Hinweis: Nutzt die bereits erfolgte Bewertung von Ursachen (Wahrscheinlichkeit) und Schäden (Schwere) aus 1. und 2.

4.6 Hotspots identifizieren (10 min)

Ziel: Muster und Häufungen erkennen – systemische Schwachstellen benennen.

Kernfragen:

- Welche Ursachen kommen besonders oft in kritischen Kombinationen vor?
- Welche Schadenstypen oder Ursachen treten besonders häufig auf?
- Welche Risikopfade sind besonders zentral oder systemrelevant?

4.7 Maßnahmen ableiten (20 min)

Ziel: Konkrete Handlungsansätze zur Risikominderung auf Ursachenebene entwickeln.

Kernfragen:

- Was können wir an der Ursache tun – vermeiden, reduzieren, kontrollieren?
- Was braucht es organisatorisch, technisch oder prozessual?
- Was ist kurzfristig machbar, was langfristig nötig?

Methode: Brainstorming, Clustering, ggf. Zuordnung nach Zuständigkeit.