aric

# Responsible AI

## Worauf man bei der Implementierung von KI achten sollte

Nicolas Schulz und Anselm Fehnker

„ Mit dem Begriff der Responsible AI werden Bestrebungen zusammengefasst, Systeme künstlicher Intelligenz in verantwortungsvoller Weise zu entwickeln, respektive einzusetzen und Systeme zu schaffen, die über bestimmte Merkmale und Fähigkeiten – etwa sozialer oder moralischer Art – verfügen. “

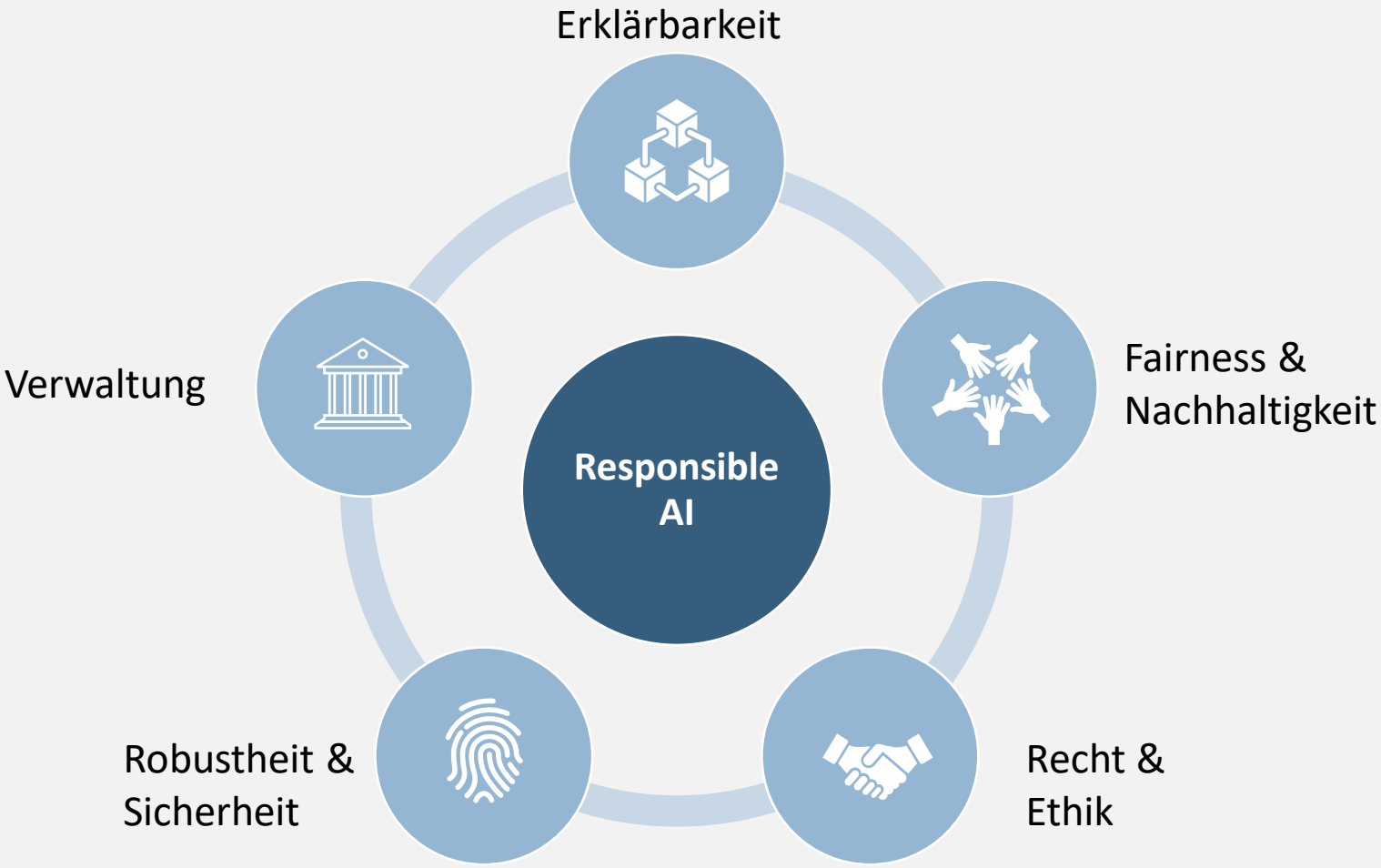
Quelle: Gabler Wirtschaftslexikon

# Responsible AI: Der mensch-zentrierte Ansatz für faire und verantwortungsvolle KI

- KI schafft Möglichkeiten, das Leben von Menschen auf der ganzen Welt zu verbessern
- Dafür muss bei der Implementierung auf bestimmte Aspekte geachtet werden
- Responsible AI ist ein Überbegriff für Frameworks mit der Aufgabe, einen langfristig verantwortungsvollen Umgang mit Künstlicher Intelligenz zu ermöglichen

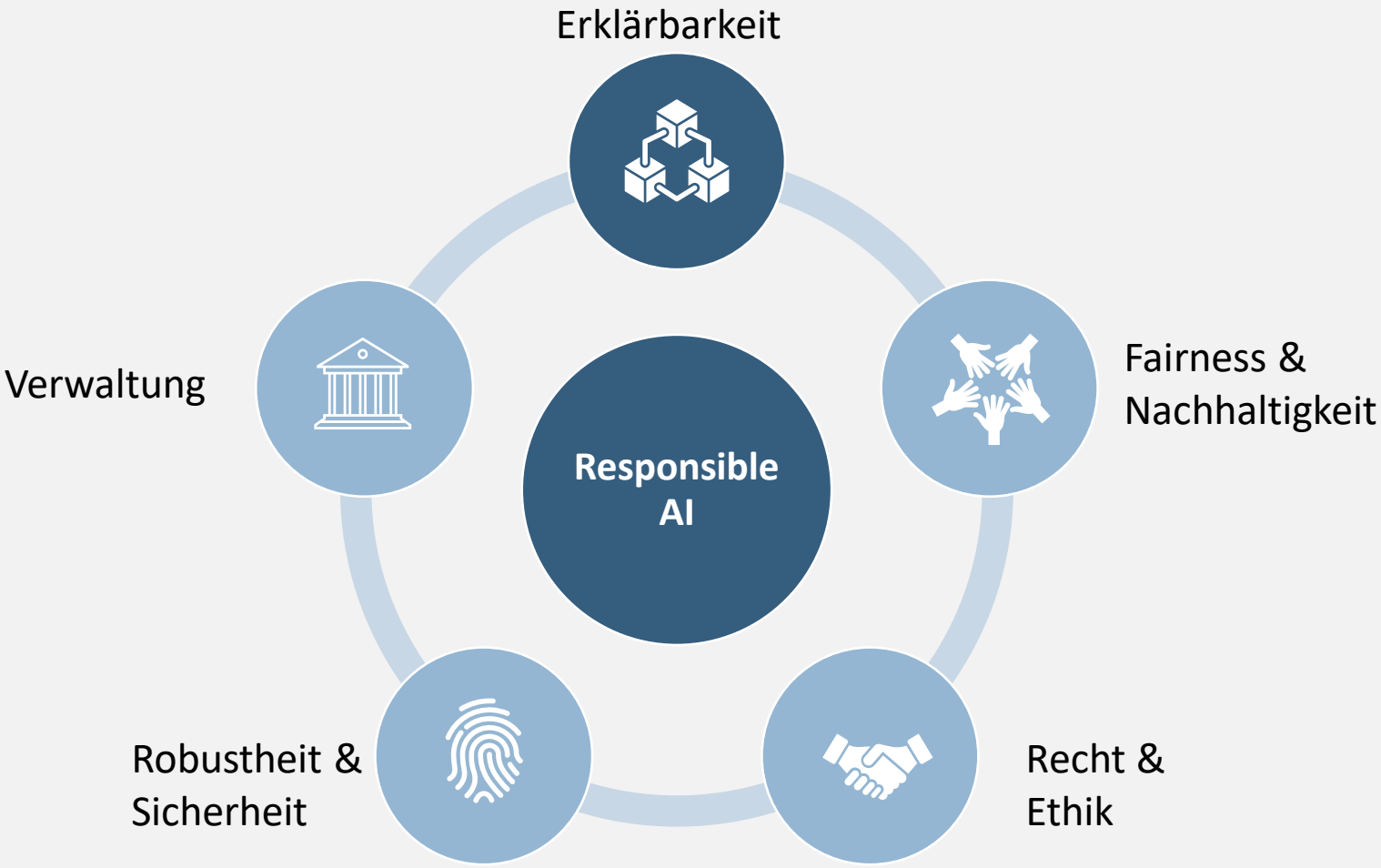
aric

# Um KI-Modelle verantwortungsbewusst einzusetzen, müssen einige Aspekte bedacht werden




Die verschiedenen Aspekte können in Wechselwirkung zueinander stehen (Trade-Off)

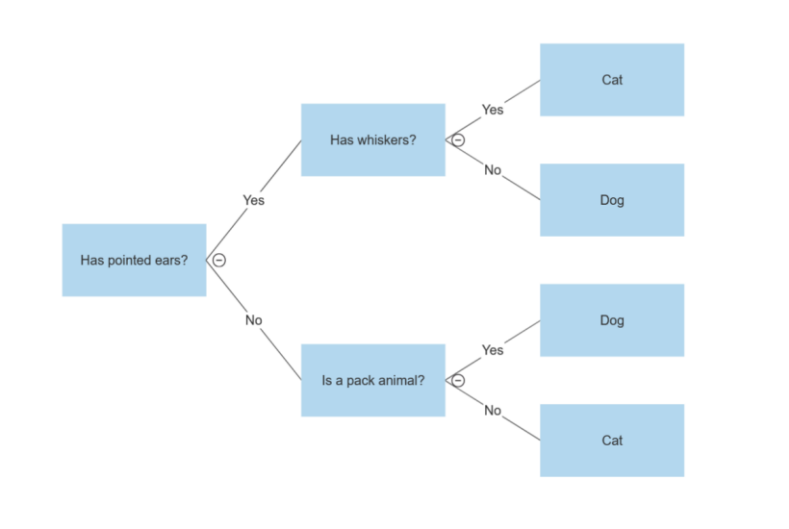
# Erklärbarkeit der Ergebnisse von KI-Modellen



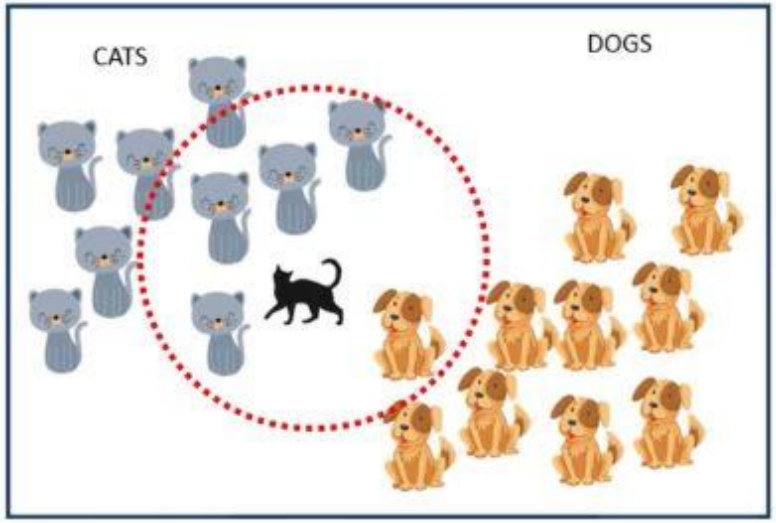

Die verschiedenen Aspekte können in Wechselwirkung zueinander stehen (Trade-Off)

# Die Erklärbarkeit der Ergebnisse ist nicht bei allen KI-Modellen gleich

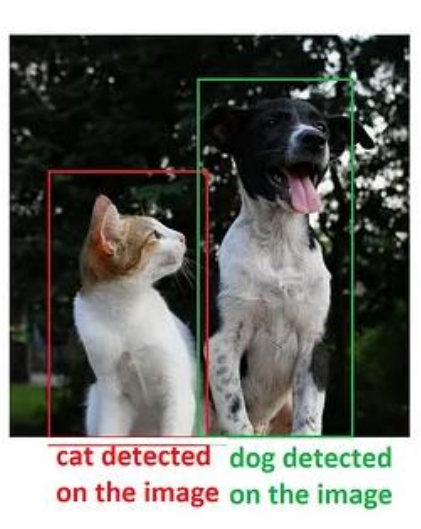
Decision Tree




K-nearest Neighbor




Neural Network

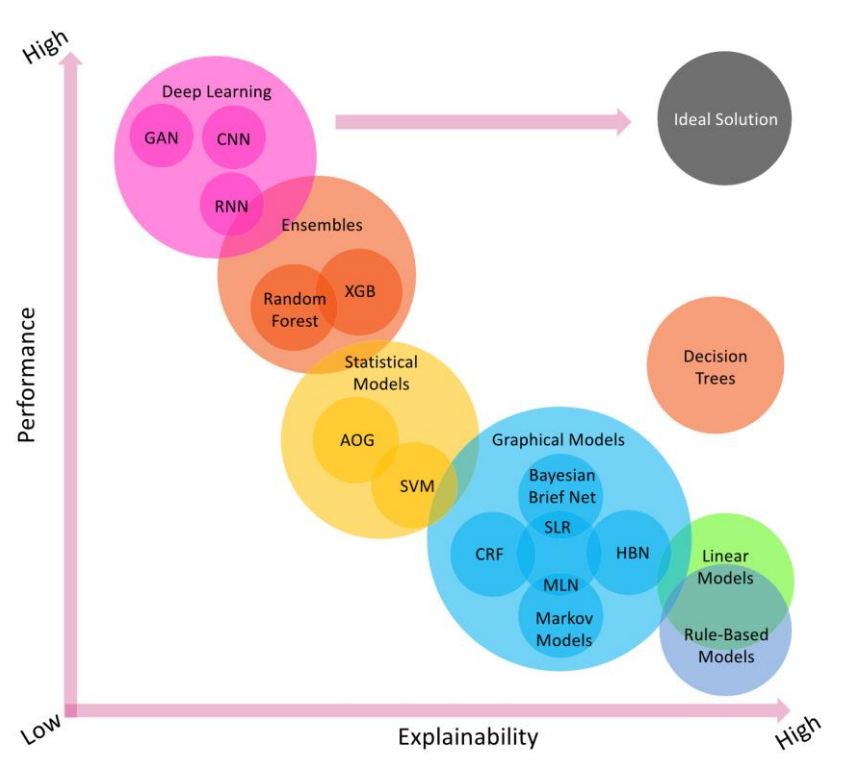


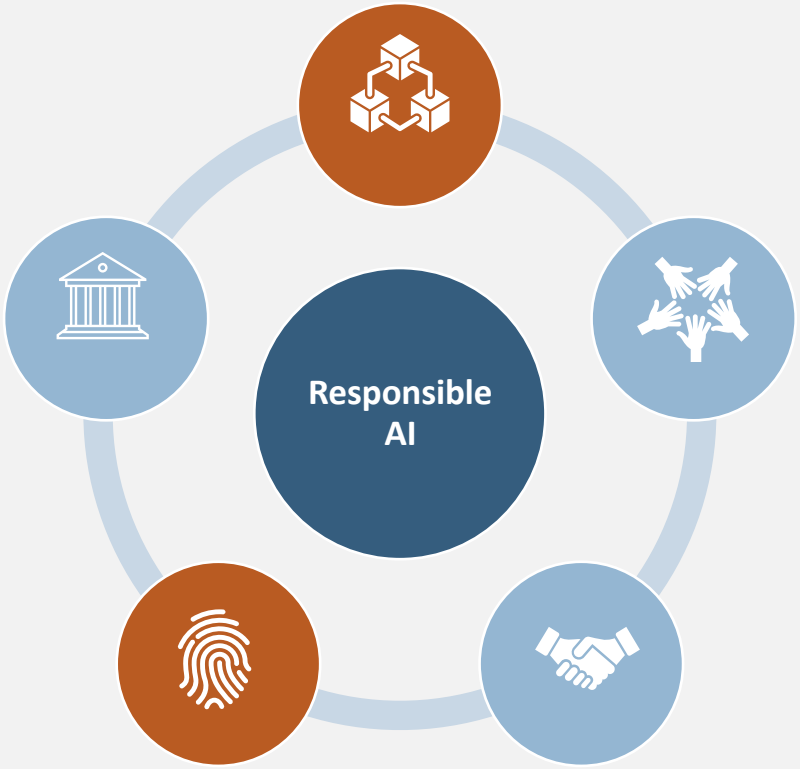

Erklärbarkeit: Wie nachvollziehbar ist der

**Entscheidungsprozess** der Künstlichen Intelligenz?

aric

# Es gibt einen Trade-Off zwischen der Erklärbarkeit und Performance




Erklärbarkeit steht in starkem Zusammenhang

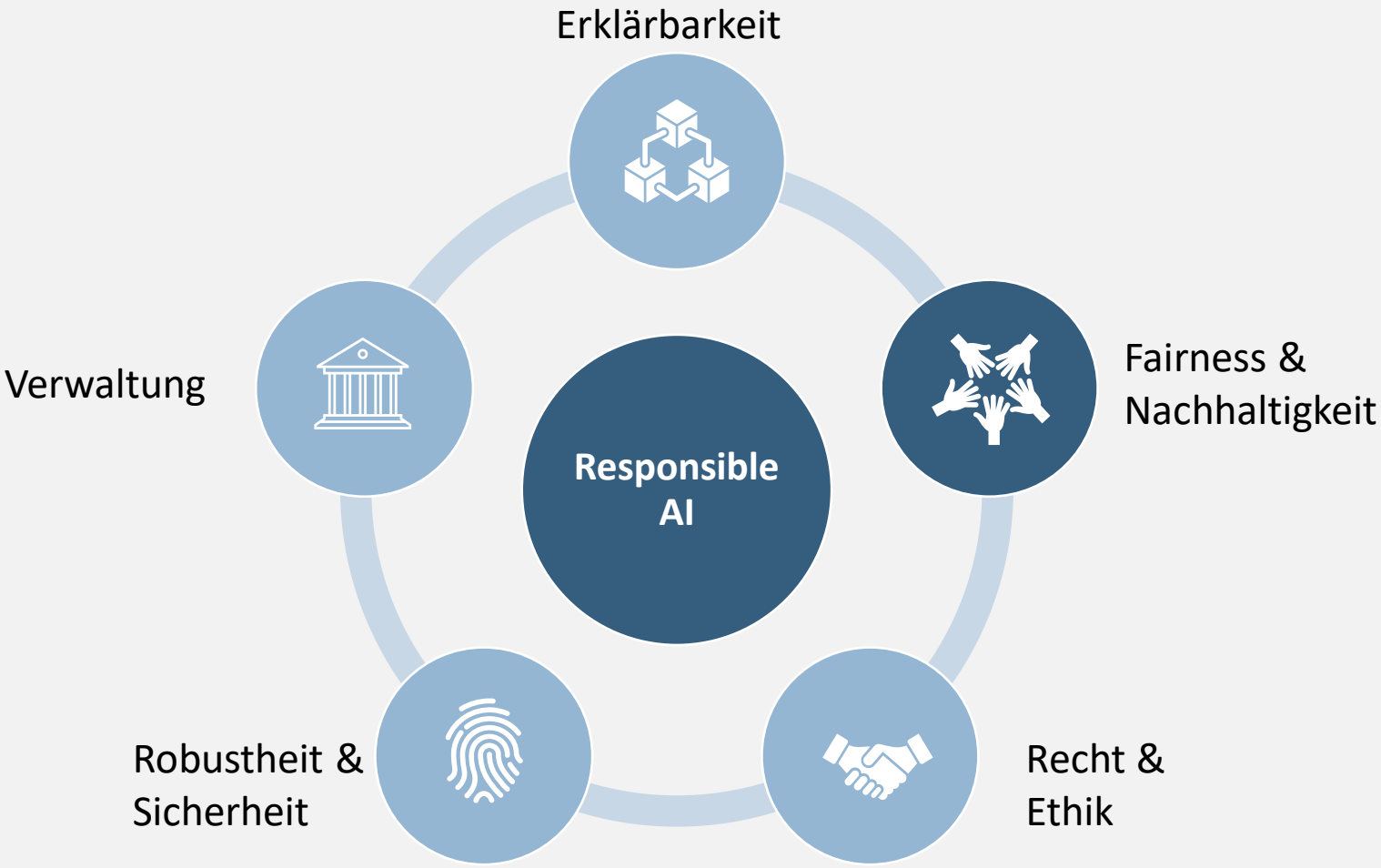
mit Sicherheit & Robustheit der Modelle

# Warum ist es wichtig den Entscheidungsprozess nachvollziehen zu können?

Beispiel: Bildklassifizierung von Großkatzen

- Nur weil das Model die richtige finale Entscheidung trifft, bedeutet es nicht das es auf einer sinnigen Grundlage geschieht
  - Datensatzspezifische 'confounding variables' (z.B. Bildhintergrund) sind falsche Entscheidungsgrundlagen, führen aber zu einem falschen Verständis was die Künstliche Intelligenz kann

# Fairness und Nachhaltigkeit




Die verschiedenen Aspekte können in Wechselwirkung zueinander stehen (Trade-Off)

# KI-Modelle lernen Verzerrungen (Bias) aus den Trainingsdaten mit

AI generated image of „a doctor“ vs „a teacher“

# Künstliche Intelligenz hat einen Einfluss auf Menschen und Umwelt

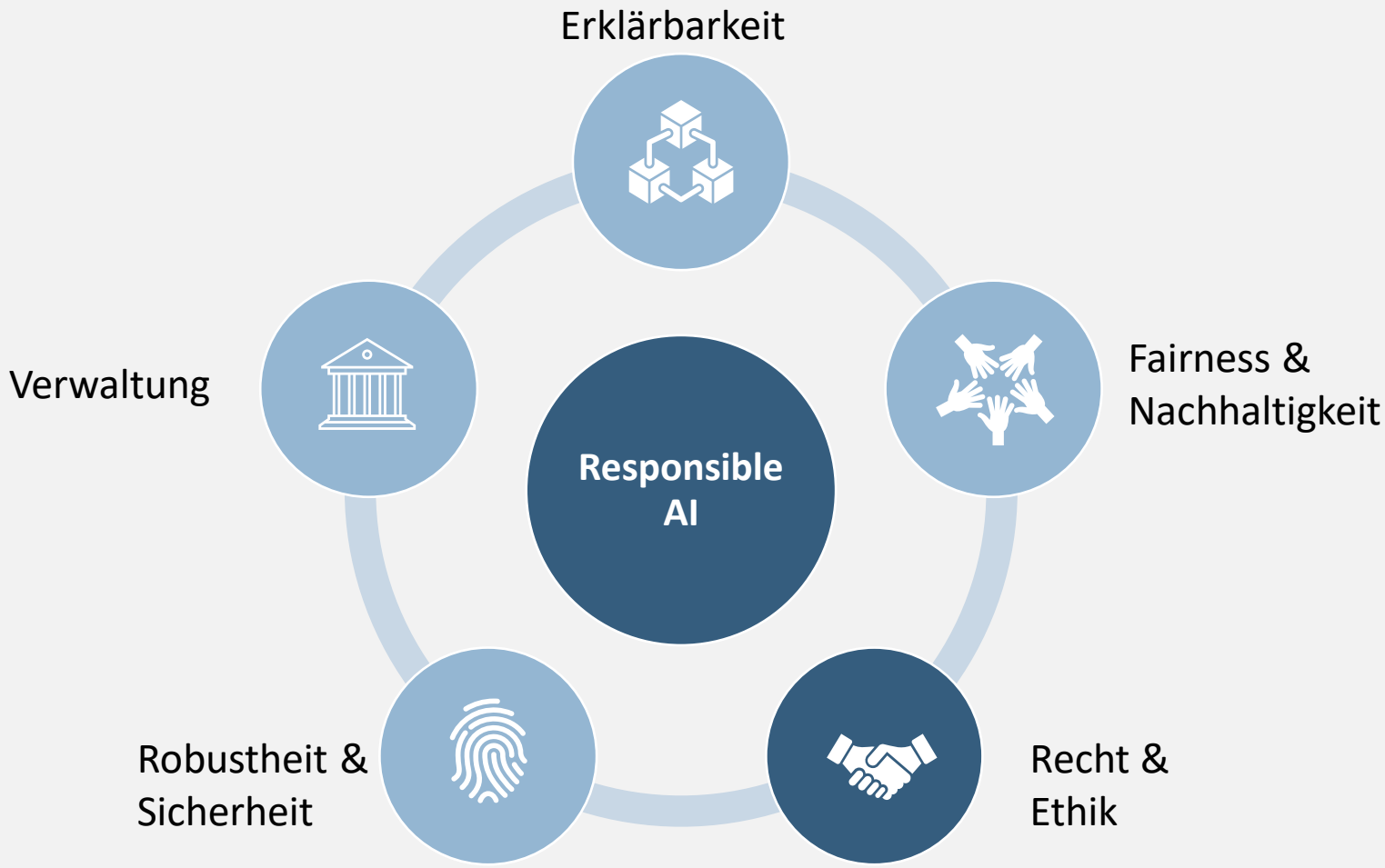
KI kann Ungleichheiten verstärken

KI-Modelle verbrauchen viel Strom

Schlechte Arbeits-bedingungen beim *labeln*

KI kann zu mehr Konsum / Produktion führen

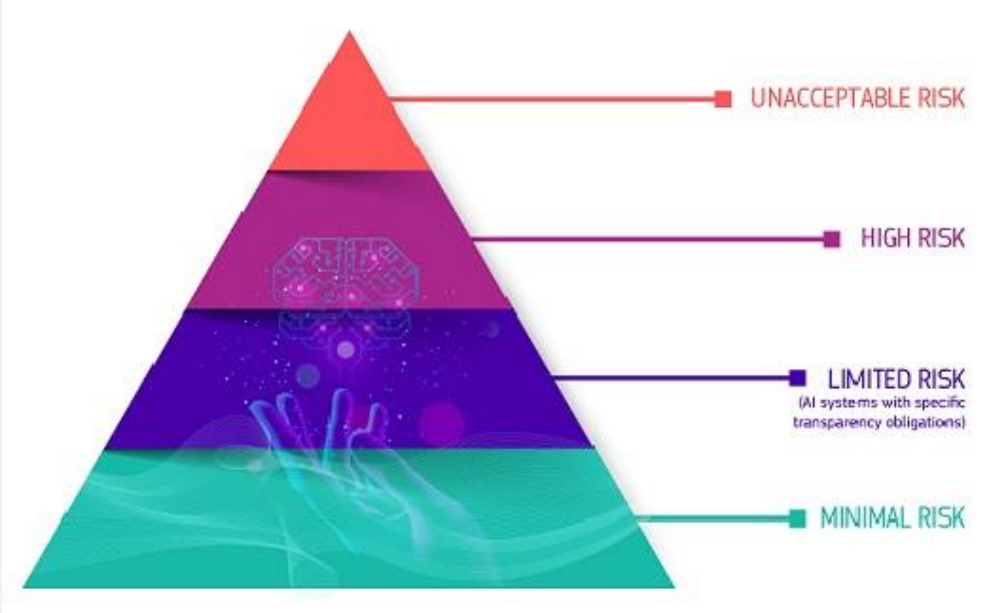
# Recht und Ethik




Die verschiedenen Aspekte können in Wechselwirkung zueinander stehen (Trade-Off)

# Grenzen und Standards - Was darf eine KI maximal und was muss sie mindestens können?

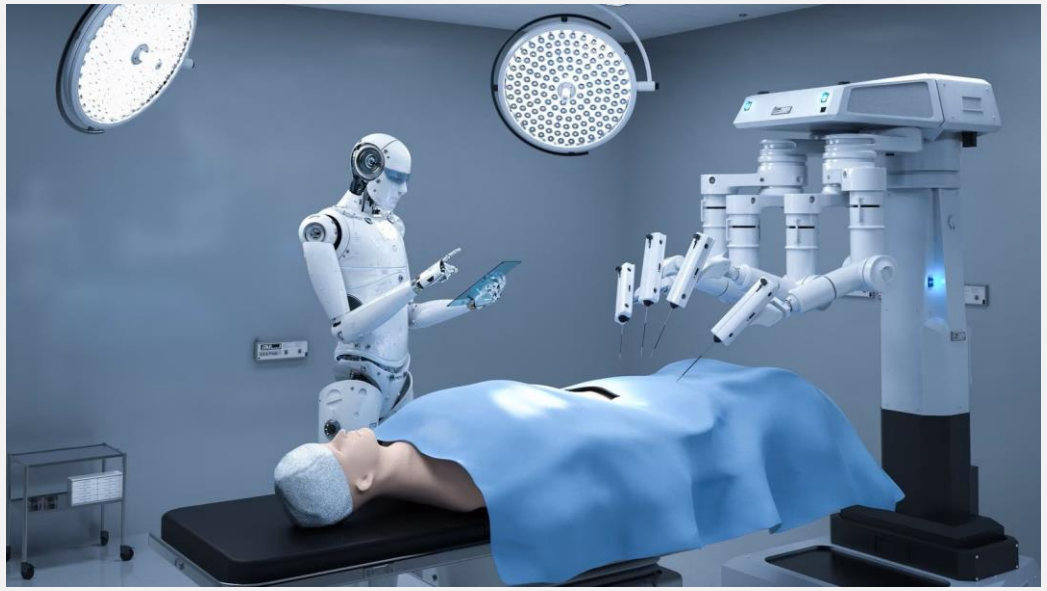
- Neue rechtliche Fragen treten auf, deren Beantwortung beidseitige Expertise (Recht und KI) bedarf
- Bei allen (europäischen) KI-Anwendungen müssen der EU AI Act und die DSGVO berücksichtigt werden
- Grundrechte, Sicherheit & Gesundheit von Menschen muss immer geachtet werden
    - Komplexe Fragestellungen ergeben sich in der Umsetzung: Was ist vertrauenswürdiger, ein 99% treffsicherer Algorithmus oder ein Mensch?
- Frage von Verantwortung und Haftung: Wer ist Schuld bzw. Haftet bei Fehlern?
    - Die Nutzer:innen des Produkts?
    - Die Entwickler:innen des Algorithmus?
    - Das Unternehmen, welches das Produkt bereitstellt?



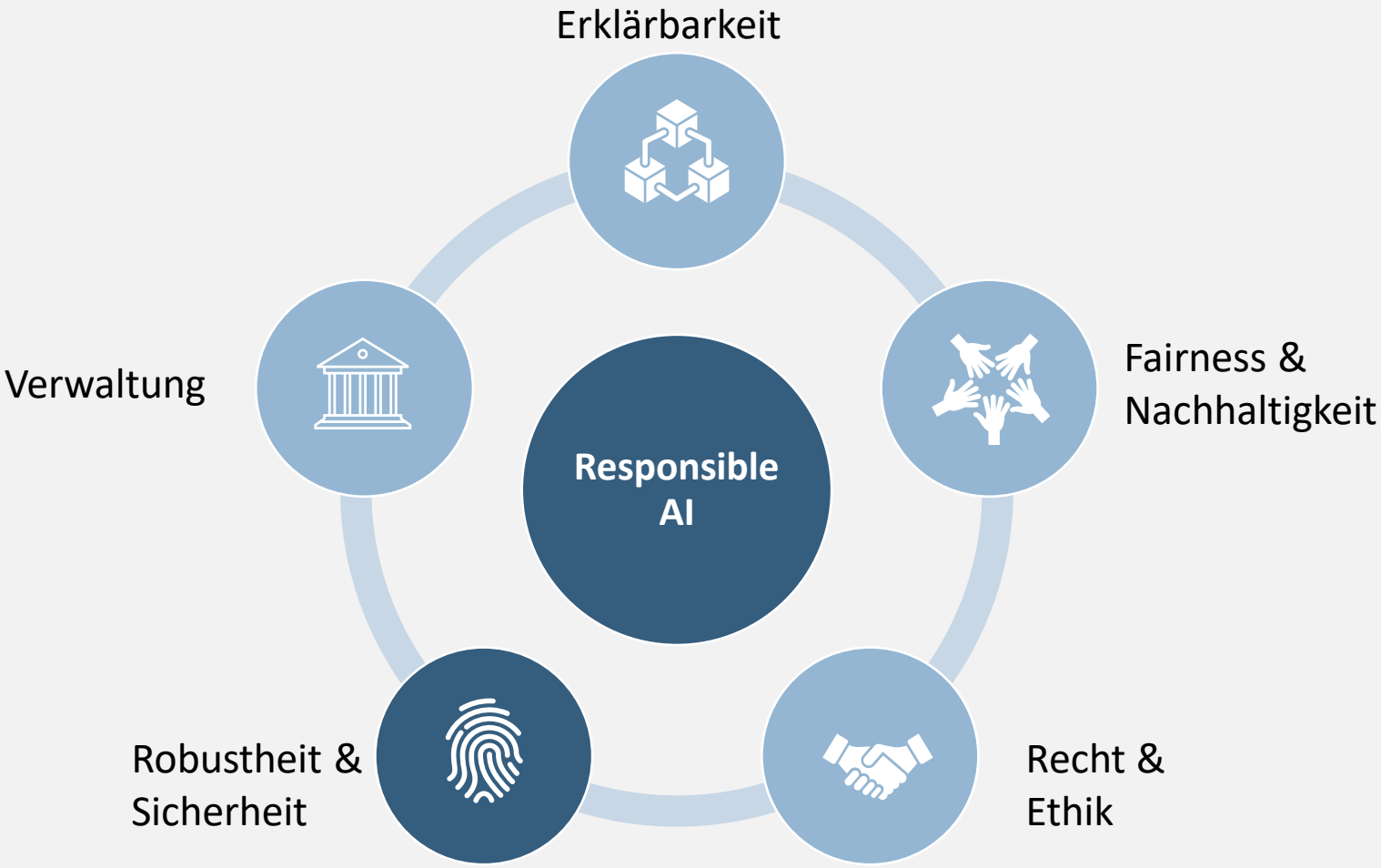
# Nicht alle Aufgaben, die von einer KI übernommen werden können, sollten von einer KI übernommen werden (KI & Ethik)

- Bei jeder Mensch-Maschine-Interaktion sind Vertrauen und Akzeptanz wichtig
  - Wie können wir dieses Vertrauen aufrichtig gewährleisten?

    -> z.B. Privacy by Design
- KI sollte als ein Werkzeug gesehen werden, welches dem reflektierten Nutzer:innen das Leben vereinfachen kann
  - Ein Grundverständnis der Technologie ist nötig
- Human Agency: Es muss dauerhaft gewährleistet sein, dass der Mensch die Kontrolle, Verantwortung und Entscheidungsmacht über die Künstliche Intelligenz behält



# Robustheit und Sicherheit




Die verschiedenen Aspekte können in Wechselwirkung zueinander stehen (Trade-Off)

# Ein KI-System sollte in allen Anwendungsfällen stabil funktionieren

Herausforderungen:

In der realen Welt kommen viele Abweichungen vor

Trainingsdaten können nicht alle Randfälle abdecken

Umgebungen verändern sich mit der Zeit

Gegenmaßnahmen:

Künstliches hinzufügen von Noise

Gute Train-Test-Validation Splits

Regelmäßiges Nachtrainieren

aric

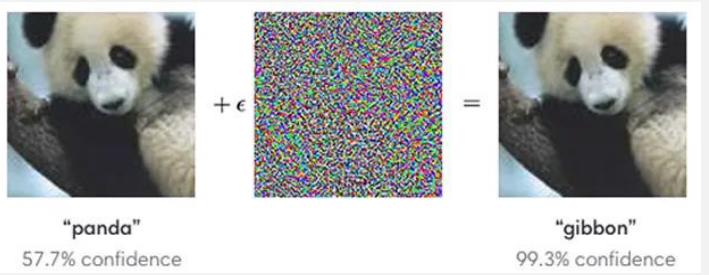
# KI-Modelle können anfällig für Cyberattacken sein

KI-Modelle sind **Teil des Softwaresystems**. Für sie gelten die gleichen **Sicherheitsvorschriften**.

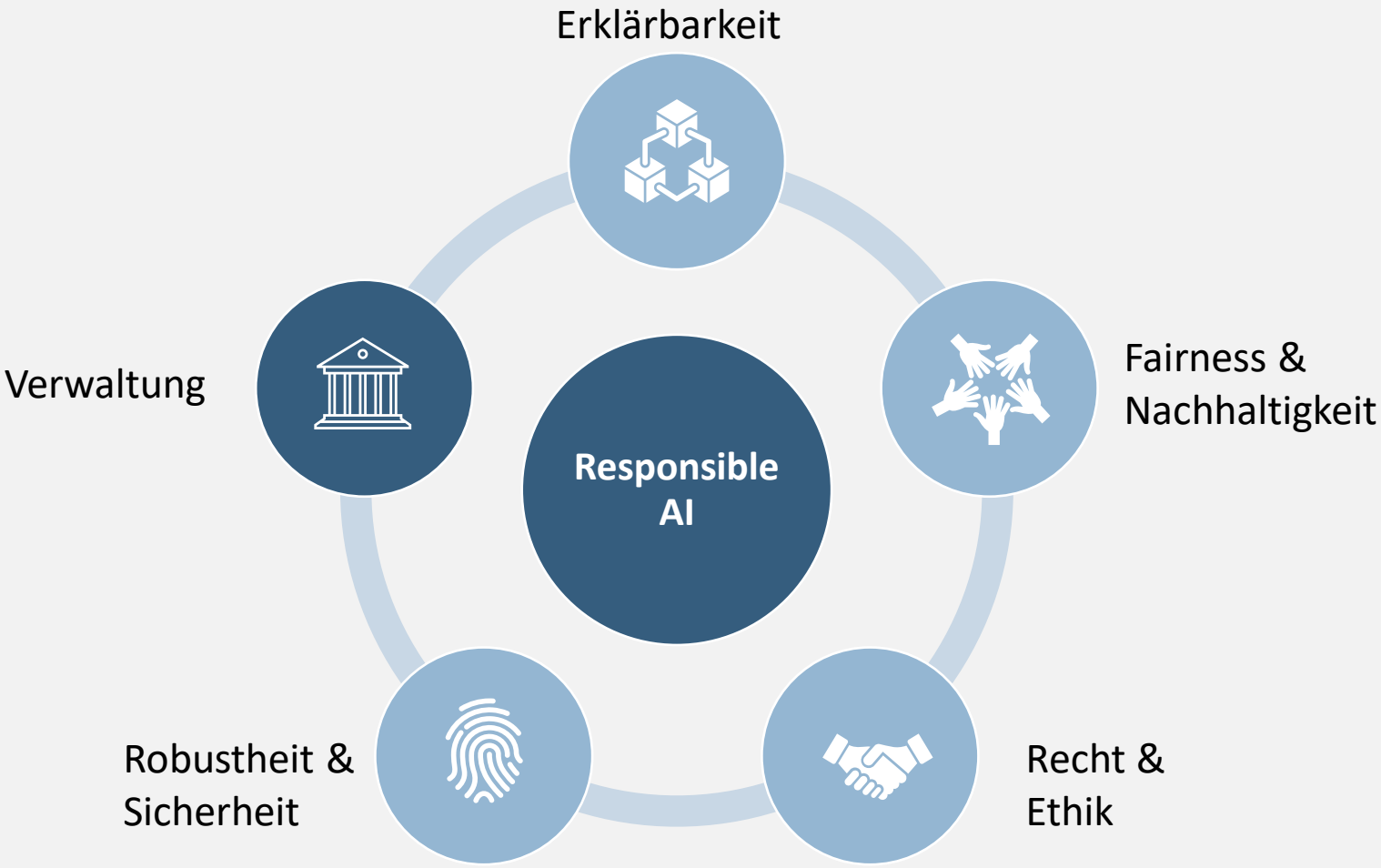
Bei der Nutzung **von externen KI-Systemen** wird häufig nicht beachtet was mit den **eingegebenen Daten** passiert

KI-Systeme können auch **absichtlich ausgetrickst** werden und so **Sicherheitslücken** im System darstellen

Es gibt viele **Trade-Offs** zwischen **Security, Robustheit Erklärbarkeit und Transparenz**

# Verwaltung

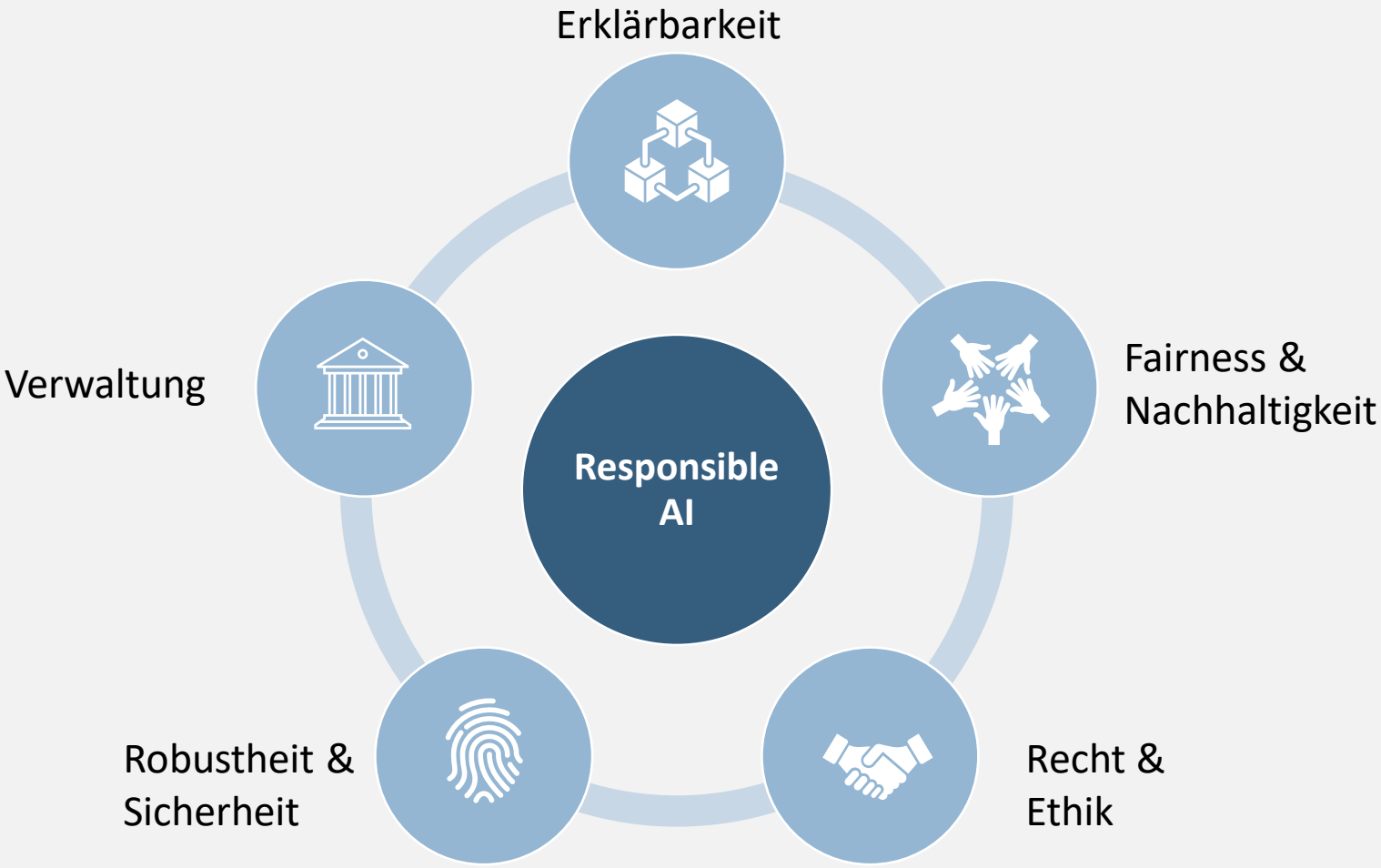



Die verschiedenen Aspekte können in Wechselwirkung zueinander stehen (Trade-Off)

# KI-Verwaltung ist essentiell, um den genannten Problematiken Struktur zu geben

- KI Governance: Bei der Prozessgestaltung muss deutlich geregelt werden, *wann* ein Mensch *wofür* verantwortlich ist und was seine/ihre *Handlungsoptionen* in einer gegebenen Situation sind
  - Wer ist verantwortlich, die KI erklären zu können? Wer sorgt dafür, dass die KI faire Entscheidungen trifft oder auch legal handelt?
- Governance muss auf verschiedenen Ebenen definiert sein
  - Innerhalb Unternehmen und Organisationen
  - Für den öffentlichen Raum
  - Auf politischer Ebene (EU AI-Act)
- Ein Grundwissen in der breiten Bevölkerung ist essentiel, um diese Herausforderungen sinnig angehen zu können!

# Diskussion




Die verschiedenen Aspekte können in Wechselwirkung zueinander stehen (Trade-Off)

# Danke!


Anselm Fehnker
AI Project Manager
fehnker@aric-hamburg.de

Nicolas Schulz
AI Project Manager
schulz@aric-hamburg.de

Artificial Intelligence Center Hamburg (ARIC) e.V.
Van-der-Smissen-Straße 9
22767 Hamburg

www.aric-hamburg.de