KI und Ethik

T02.1 EU AI Act – Einordnung von KI-Anwendungen

Anwendung 1:
Ein personalisiertes Empfehlungssystem (z. B. für Netflix oder Spotify) greift auf Nutzerdaten zu, um Inhalte vorzuschlagen. Da solche Systeme in der Regel keine unmittelbaren sicherheitsrelevanten oder grundrechtlichen Risiken bergen, ordnet man sie meist in die **niedrig-risikobehaftete** oder **minimale Risikokategorie** ein.

Anwendung 2:
Ein Spracherkennungssystem zur Analyse von Notrufen in einer Rettungsleitstelle wirkt in einem sicherheitskritischen Umfeld, in dem Fehlinterpretationen direkte Folgen für Menschenleben haben können. Daher fällt diese Anwendung in die **hoch riskante** Kategorie.

Anwendung 3:
Ein automatisiertes Entscheidungsunterstützungssystem zur Vergabe von staatlichen Sozialleistungen beeinflusst fundamentale soziale Rechte und die finanzielle Situation von Bürger*innen. Diese Anwendung hat eine große Tragweite und muss besonders zuverlässig und nachvollziehbar sein – sie wird daher ebenfalls als **hoch riskant** eingestuft.

T02.2 Interpretation der Folien im Kontext von „XAI“ (Explainable AI)

Die gezeigten Folien verdeutlichen zentrale Aspekte der **Erklärbarkeit** von KI-Systemen:

1. **Daten-Typen und XAI-Methoden:**
   - **Images:** Bei Bilddaten kann man mithilfe von Visualisierungen (z. B. Heatmaps, Saliency Maps) aufzeigen, welche Bildbereiche zur Entscheidungsfindung des Modells beigetragen haben.
   - **Text:** Bei Textdaten kann man beispielsweise mittels Highlighting darstellen, welche Textpassagen (Tokens, Wörter) für ein bestimmtes Sentiment oder eine bestimmte Klassifikation verantwortlich sind.
   - **Tabular:** Bei strukturierten Daten werden häufig Feature-Importances oder Entscheidungsbäume genutzt, um nachvollziehbar zu machen, welche Spalten (Features) den größten Einfluss auf die Modellentscheidung hatten.
2. **Wer benötigt Erklärungen und wozu?**

- **Model Builder & ML Ops:**
  - Erklärungen helfen bei Fehlersuche und Optimierung (z. B. Warum ist mein Modell nicht performant? Welche Features sind am wichtigsten?).
  - Aus den Erklärungen können Trainingsdaten verbessert und Features neu gewichtet werden.
- **Endnutzer*innen der ML-Systeme:**
  - Sie brauchen Erklärungen, um das Modellvertrauen aufzubauen und fundierte Entscheidungen zu treffen (z. B. Sollte ich der Empfehlung oder Prognose glauben?).
- **Öffentliche Stakeholder:**
  - Für Regulierung, Ethik-Diskussionen und gesellschaftliche Akzeptanz ist Transparenz essenziell (z. B. Wie wirkt sich das Modell auf unterschiedliche Bevölkerungsgruppen aus?).

3. **Mögliche Handlungsoptionen durch Erklärungen:**
   - **Modell- und Datenanpassung:** Die Einsichten aus XAI können genutzt werden, um das Modell-Design zu verfeinern, Trainingsdaten zu bereinigen oder zusätzliche Features einzubeziehen.
   - **Kontroversen klären:** Erklärungen helfen, bei strittigen oder sensiblen Entscheidungen (z. B. Kreditvergabe, medizinische Diagnosen) Transparenz herzustellen und ggfs. Einspruchsprozesse zu unterstützen.
   - **Verantwortungsbewusste Nutzung:** Stakeholder können auf Basis von Erklärungen Leitlinien entwickeln, um KI ethisch, fair und nachvollziehbar einzusetzen.

Insgesamt zeigen die Folien, dass Explainable AI nicht nur eine technische Herausforderung (z. B. Feature-Attribution), sondern auch eine organisatorische und gesellschaftliche Dimension hat. Unterschiedliche Stakeholder haben unterschiedliche Bedürfnisse und Fragestellungen in Bezug auf Erklärbarkeit und Transparenz von KI-Systemen.

T02.3 Was sind Model Cards?

Model Cards sind standardisierte Dokumentationswerkzeuge für Machine-Learning-Modelle. Sie beinhalten wesentliche Informationen wie:

- **Modellbeschreibung:** Zweck, Einsatzgebiet und Funktionsweise.
- **Leistungsmetriken:** Ergebnisse und Benchmarking.
- **Einschränkungen und Annahmen:** Was das Modell leisten kann und wo es an seine Grenzen stößt.
- **Ethik und Fairness:** Hinweise zu möglichen Verzerrungen (Bias) und ethischen Aspekten.

Diese Dokumente unterstützen Transparenz und helfen Entwicklern sowie Anwender*innen, informierte Entscheidungen bezüglich des Einsatzes eines Modells zu treffen.

T02.4 Bias in einem KI-basierten Emotionsidentifikationssystem

Wenn eine KI entwickelt wird, die Emotionen identifizieren soll, kann ein **Bias** (Verzerrung) zu erheblichen Fehlinterpretationen führen:

- **Kulturelle und demographische Unterschiede:** Das System könnte bestimmte Mimik oder Gestik falsch interpretieren, weil es überwiegend mit Daten einer bestimmten Bevölkerungsgruppe trainiert wurde.
- **Stereotypisierung:** Es besteht die Gefahr, dass das Modell Vorurteile übernimmt und etwa negative Emotionen fälschlicherweise bestimmten Gruppen zuordnet.
- **Fehlentscheidungen:** Eine verzerrte Emotionserkennung kann zu unangemessenen Reaktionen oder falschen Interventionen führen, was besonders in sicherheitsrelevanten oder sensiblen Kontexten problematisch ist.

T02.5 KI und die Arbeitswelt

1. Auswirkungen auf den Arbeitsmarkt und betroffene Branchen

Die Einführung von Künstlicher Intelligenz verändert den Arbeitsmarkt signifikant:

- **Disruption traditioneller Berufsfelder:** Automatisierung kann Tätigkeiten in Bereichen wie Fertigung, Kundenservice, Logistik und Verwaltungsaufgaben ersetzen.
- **Branchen mit hoher Automatisierung:** Besonders betroffen sind Berufe, in denen repetitive, standardisierte Aufgaben dominieren. Gleichzeitig entstehen neue Berufsbilder und Tätigkeiten, insbesondere im IT- und Datenanalysesektor.

2. Einfluss der Automatisierung auf Arbeitsplatzdynamik und Qualifikationsanforderungen

- **Veränderung der Arbeitsplatzstruktur:** Routineaufgaben werden zunehmend automatisiert, während kreative, zwischenmenschliche und strategische Kompetenzen an Bedeutung gewinnen.
- **Neue Qualifikationsprofile:** Arbeitnehmer*innen müssen sich in den Bereichen digitale Kompetenzen, Datenanalyse und technisches Verständnis weiterbilden, um mit den

neuen Technologien Schritt zu halten.
- **Flexibilisierung der Arbeitswelt:** Es entstehen hybride Arbeitsmodelle und interdisziplinäre Teams, die eng mit KI-Systemen zusammenarbeiten.

3. Rolle staatlicher Regulierung und politischer Maßnahmen

- **Schutzmaßnahmen:** Staatliche Regulierungen und politische Maßnahmen können dazu beitragen, den Übergang in eine KI-geprägte Arbeitswelt sozial verträglich zu gestalten.
- **Weiterbildung und Umschulung:** Investitionen in Bildung und Umschulungsprogramme sind notwendig, um die Arbeitskräfte auf die veränderten Anforderungen vorzubereiten.
- **Arbeitsmarktsicherheit:** Politische Maßnahmen können soziale Sicherheitsnetze stärken und dafür sorgen, dass die Vorteile der Automatisierung breit verteilt werden.

T02.6 Human-Centered-Design im Kontext ethischer AI-Entwicklung

Der Begriff **Human-Centered-Design** bezeichnet einen Entwicklungsansatz, bei dem die Bedürfnisse, Werte und Perspektiven der Menschen im Mittelpunkt stehen. Im Kontext der ethischen AI-Entwicklung bedeutet dies:

- **Partizipation:** Nutzer*innen und betroffene Gruppen werden frühzeitig in den Entwicklungsprozess eingebunden.
- **Transparenz und Verständlichkeit:** KI-Systeme werden so gestaltet, dass ihre Entscheidungsprozesse nachvollziehbar sind.
- **Fokus auf den Menschen:** Technologien sollen die menschliche Autonomie unterstützen und nicht ersetzen, wobei ethische Überlegungen wie Fairness und Inklusion berücksichtigt werden.

T02.7 KI – Urheberschaft und Systemsicherheit

a) Watermarks in KI-Anwendungen

Watermarks sind digitale Kennzeichnungen, die in Text- oder Bildgeneratoren eingebettet werden können. Sie dienen dazu:

- **Authentizität und Herkunft:** Nachzuverfolgen, ob Inhalte von einer KI erzeugt wurden.
- **Missbrauchsprävention:** Manipulationen zu erkennen und Urheberrechtsverletzungen vorzubeugen.

- **Transparenz:** Endnutzer*innen können so informierte Entscheidungen treffen, ob sie die Inhalte als von Menschen oder KI generiert einstufen.

b) Adversarial-Attacks

Adversarial-Attacks beziehen sich auf gezielte Manipulationen von Eingabedaten, um KI-Systeme zu täuschen. Dabei werden kleine, oft für den Menschen unmerkliche Veränderungen vorgenommen, die dazu führen:

- **Fehlklassifikation:** Das Modell trifft falsche Entscheidungen.
- **Sicherheitsrisiken:** Besonders in sicherheitskritischen Systemen können solche Angriffe verheerende Folgen haben.

c) Die vier Wellen von KI nach Kai-Fu Lee

Kai-Fu Lee beschreibt die Entwicklung der KI in vier „Wellen“:

1. **Internet AI:** Nutzung großer Datenmengen und Nutzerinteraktionen, um personalisierte Dienste bereitzustellen.
2. **Business AI:** Einsatz von KI zur Optimierung betrieblicher Prozesse und Entscheidungsunterstützung in Unternehmen.
3. **Perception AI:** Integration von KI in sensorbasierte Anwendungen (z. B. Bilderkennung, Spracherkennung) für eine verbesserte Wahrnehmung der Umgebung.
4. **Autonomous AI:** Systeme, die selbstständig und in Echtzeit Entscheidungen treffen – beispielsweise in autonomen Fahrzeugen oder Robotik.

Diese Wellen verdeutlichen die zunehmende Komplexität und den wachsenden Einfluss von KI-Technologien in unterschiedlichen Lebens- und Arbeitsbereichen.