

KI-BILDUNGSURLAUB

KURS-HANDOUT · TAG 1

Seminarleitung: Cordei Clottey

Überblick

Tag 1 führt in die Grundlagen künstlicher Intelligenz ein: Begriffe, KI-Kategorien, Training von Sprachmodellen sowie rechtliche und sicherheitsrelevante Aspekte für den verantwortungsvollen Einsatz.

1. EINFÜHRUNG: WAS IST KÜNSTLICHE INTELLIGENZ?

Künstliche Intelligenz-Systeme (KI) ahmen menschliche Intelligenz nach und führen Aufgaben aus, die normalerweise menschliches Denken, Problemlösen oder logische Schlussfolgerungen erfordern.

Historischer Kontext & Fähigkeiten:

Geschichte: Erste mechanische Automaten und Ideen von belebten Figuren gab es bereits in der griechischen Antike.

Leistungsspektrum: KI kann heute selbstständig aus Daten lernen, komplexe Muster erkennen, Sprache verstehen, logisch denken, Vorhersagen treffen, Entscheidungen fällen und völlig neue Inhalte generieren.

KATEGORISIERUNG: SCHWACHE VS. STARKE KI

Schwache KI (Narrow AI):

Systeme, die explizit für ein klar definiertes Aufgabengebiet entwickelt wurden (z. B. Schachcomputer, medizinische Bilderkennung). Alle derzeit weltweit existierenden KI-Systeme, einschließlich modernster generativer Sprach- und Bildmodelle, fallen unter diese Kategorie.

Starke KI (Artificial General Intelligence - AGI):

Eine rein hypothetische, zukünftige Form der KI. Sie besäße die Fähigkeit, menschliche kognitive Kompetenzen in vollem Umfang flexibel nachzubilden und auf beliebige Themenbereiche eigenständig anzuwenden.

KI IM MODERNEN ALLTAG

Gesundheit: Früherkennung und hochpräzise Diagnosen durch strukturierte Analyse medizinischer Bild- und Patientendaten.

Verkehr: Navigationsanwendungen berechnen Routenoptimierungen in Echtzeit und bilden das Fundament für autonomes Fahren.

Finanzen: Echtzeit-Betrugserkennung bei Kreditkartentransaktionen sowie automatisierter, algorithmischer Aktienhandel.

Unterhaltung & Shopping: Personalisierte Streaming-Empfehlungen und kundenorientierte Support-Chatbots.

Smart Homes: Intelligente Automatisierung von Raumtemperaturen, Licht- und Mediensteuerung via Sprachassistenten.

2. KI-SYSTEME UND SCHLÜSSELKOMPONENTEN

Ein funktionierendes KI-System basiert auf dem nahtlosen Zusammenspiel technischer Grundkomponenten:

Daten: Das Fundament. Ohne riesige Mengen strukturierter oder unstrukturierter Informationen (z. B. Social-Media-Interaktionen) kann kein Modell Muster ableiten.

Algorithmen: Mathematische Regeln und deterministische Anweisungen, die die Datenverarbeitung leiten.

Modell: Das mathematische Herzstück und Endprodukt des Trainingsprozesses, welches spezifische Aufgaben selbstständig löst.

DIE TECHNOLOGISCHE EVOLUTION

Maschinelles Lernen (ML) [1980er - 2000er]:

Statistische Methoden, bei denen Systeme anhand strukturierter Daten (z. B. Tabellen) lernen, Muster und Korrelationen zu erkennen, ohne explizit für die spezifische Aufgabe programmiert worden zu sein (Beispiel: klassische regelbasierte Spam-Filter).

Vorteile: Hohe Transparenz ('White Box' - nachvollziehbare Rechenwege), geringer Bedarf an Prozessorleistung.

Nachteile: Extrem limitierte Fähigkeiten bei der Verarbeitung komplexer, unstrukturierter Daten wie Bilder oder fließende Sprache.

Neuronale Netzwerke & Deep Learning [Ab 2012]:

Vom menschlichen Gehirn inspirierte Algorithmen aus künstlichen Neuronen (Knoten), die in tiefen, hierarchischen Schichten (Input-, Hidden- und Output-Schichten) angeordnet sind. Sie lernen selbstständig hochkomplexe Abstraktionen (z. B. Kanten -> Formen -> Gesichter bei der Bildanalyse). Untergruppen wie Large Language Models (LLMs) nutzen diese Netze für Natural Language Processing (NLP).

Vorteile: Überragende Präzision bei unstrukturierten Daten (Text, Video, Audio, Bild). Multimodale Modelle können diese Datentypen simultan verarbeiten.

Nachteile: Immenser Ressourcenverbrauch (Daten & Hardware), mathematische 'Black Box' (die genaue Entscheidungsfindung innerhalb der versteckten Schichten ist nicht mehr detailliert nachvollziehbar).

DIE DREI HISTORISCHEN KI-KATEGORIEN

1. Regelbasierte Systeme (seit 1950er): folgt starren wenn-dann-Vorgaben und lernt nicht selbstständig (Beispiel: Eliza Chatbot 1965, Mycin 1972). Führt in den 1980ern zum 'KI-Winter', da die Grenzen der Skalierbarkeit schnell erreicht waren.

2. Diskriminative KI (1980er-1990er): analysiert, klassifiziert und interpretiert bestehende Daten (Beispiel: moderne Spam-Erkennung, Smartphone-Fingerabdruckscanner, vorausschauende Navigationssysteme).

3. Generative KI (seit 2021/2022): findet abstrakte Strukturen in gigantischen Datensätzen und erschafft daraus völlig neue, einzigartige Inhalte (Beispiele: ChatGPT, Google Gemini, Midjourney, Dall-e, Music-Generatoren).

3. Wie KI lernt: der dreiphasige Trainingsprozess

Generative Sprachmodelle (LLMs) wie ChatGPT durchlaufen einen hochkomplexen, mehrstufigen Lernprozess, um von rohen Textdaten zu sicheren, dialogfähigen Assistenten zu reifen:

Phase 1: Das Pre-Training (Vortraining - Aufbau des Basismodells)

Das Modell liest im ersten Schritt riesige Datenmengen aus dem gesamten Internet (Websites, Bücher, Fachartikel, Bildbeschreibungen, Alternativtexte). In dieser Phase lernt das neuronale Netzwerk rein statistisch, welches Wort mit welcher Wahrscheinlichkeit auf ein anderes folgt (reine Wortvorhersage basierend auf Mustern). Es entsteht das 'Basismodell'. Dieses besitzt zwar ein breites Weltwissen, versteht aber noch keine Dialogstrukturen, reagiert oft unberechenbar und unterscheidet nicht zwischen nützlichen und schädlichen Inhalten.

Phase 2: Das Fine-Tuning & RLHF (Feinabstimmung)

Um aus dem rohen Basismodell einen nützlichen Assistenten zu formen, folgt das Lernen durch menschliches Feedback (Reinforcement Learning from Human Feedback - RLHF). Menschliche Trainer bewerten die Antworten der KI. Hierbei wird dem System eine feste 'Standard-Persona' über einen System-Prompt (Verhaltensanweisung) injiziert. Das Modell lernt:

Hilfreich, ehrlich, höflich, respektvoll und empathisch zu antworten.

Die bestmögliche Assistenz bei der Lösung von Anwenderfragen zu bieten.

Die strikte Verweigerung schädlicher Inhalte (z. B. keine Anleitungen zu Straftaten, keine rassistischen oder pornografischen Ausgaben).

Phase 3: In-Context Learning (Anpassung beim Endnutzer)

Wenn ein Anwender einen Prompt eingibt, startet eine temporäre Anpassung im sogenannten 'Kontextfenster' (dem temporären Arbeitsspeicher oder Notizblock des Modells). Durch den fortlaufenden Chatverlauf passt sich die KI individuell an den Nutzer und dessen aktuelle Rolle (Role-Playing) an. Dieses Wissen verbleibt jedoch flüchtig im aktuellen Chat-Kontext und verändert das Kernmodell nicht dauerhaft.

4. SICHER MIT KI ARBEITEN: RECHT, DATENSCHUTZ & COMPLIANCE

Der geschäftliche und private Einsatz von KI erfordert die strikte Einhaltung regulatorischer und ethischer Rahmenbedingungen. Diese Abschnitte dürfen nicht verkürzt werden:

SÄULE 1: DATENSCHUTZ NACH DSGVO / GDPR

Personenbezogene Daten (Art. 4 DSGVO):

Es ist strengstens untersagt, personenbezogene Daten – also jegliche Informationen, die eine natürliche Person direkt oder indirekt identifizierbar machen (z. B. Klarnamen, private E-Mail-Adressen, Telefonnummern, Kundendaten, finanzielle oder medizinische Aufzeichnungen, Stimmen, Biometrie) – in öffentliche, frei zugängliche KI-Modelle einzugeben. Jede Verarbeitung erfordert eine valide Rechtsgrundlage nach Art. 6 DSGVO (Einwilligung, Vertragserfüllung, berechtigtes Interesse oder gesetzliche Pflicht).

Risiken und Schutzmaßnahmen:

Data Residency: Die meisten führenden KI-Anbieter (OpenAI, Google) speichern und verarbeiten Daten auf Servern in den USA, wodurch kein einheitliches EU-Datenschutzniveau garantiert ist.

Opt-Out nutzen: Anwender müssen in den Datenschutzeinstellungen der Tools die Option deaktivieren, dass ihre Prompts und Dokumentenuploads für das zukünftige Training der globalen KI-Modelle verwendet werden dürfen.

Unternehmenslizenzen: Im beruflichen Kontext sollten ausschließlich geschlossene Enterprise-Tarife oder gesicherte API-Schnittstellen genutzt werden, bei denen vertraglich vereinbart ist, dass Eingaben nicht permanent gespeichert oder zwecks Modellverbesserung analysiert werden.

EU-Alternativen: Der Einsatz von Anbietern mit Sitz in der Europäischen Union (z. B. Mistral AI aus Frankreich) minimiert transatlantische Datenübertragungsrisiken.

Verantwortlichkeiten im Unternehmen:

Unternehmen als Verantwortlicher: Die juristische Person haftet vollumfänglich für Zweck und Mittel der Datenverarbeitung.

Rollenverteilung: Die Geschäftsführung trägt die Gesamtverantwortung; der Datenschutzbeauftragte berät und schult; Projektleitungen stellen die Konformität mit DSGVO und dem EU AI Act sicher.

Mitarbeiterpflichten: Striktes Befolgen der internen KI-Richtlinien, Verbot des Datenabflusses an unzulässige Tools, sofortige Meldung von Datenpannen sowie lückenlose Faktenprüfung aller Ergebnisse.

SÄULE 2: VERANTWORTUNG & QUALITÄTSSICHERUNG

Human-in-the-Loop: Da KI-Systeme statistisch plausible, aber nicht zwangsläufig korrekte Ergebnisse liefern, muss jeder Text, Code, jede juristische Einschätzung und jeder Fakt zwingend durch eine qualifizierte menschliche Fachkraft final verifiziert werden.

Haftung: Die rechtliche Verantwortung für fehlerhafte Ratschläge, falsche Programmierungen oder Urheberrechtsverletzungen im Endprodukt liegt ausnahmslos beim menschlichen Anwender bzw. dessen Unternehmen, niemals bei der KI.

Bias-Kontrolle: KI-Modelle spiegeln die Vorurteile ihrer Trainingsdaten wider (gesellschaftliche Stereotypen bezüglich Geschlecht, Ethnie, Alter oder Klasse). Ergebnisse müssen kritisch hinterfragt werden (z. B. historische Diskriminierungsmuster bei automatisierten Bewerber-Auswahlsystemen).

SÄULE 3: URHEBERRECHT (COPYRIGHT)

Fehlende Schöpfungshöhe: Rein von einer KI generierte Inhalte (Texte, Bilder, Programmiercodes, Musik) genießen nach aktueller europäischer Rechtsprechung und dem EU AI Act keinen urheberrechtlichen Schutz. Sie gelten als gemeinfrei, da die erforderliche menschliche, kreative Schöpfungsleistung fehlt.

Substantielle menschliche Bearbeitung: Urheberrechtlicher Schutz an einem Mischwerk kann erst entstehen, wenn ein Mensch das KI-Ergebnis im Nachgang substantiell und wesentlich verändert oder die KI nachweislich bloß als rein ausführendes Werkzeug einer hochgradig detaillierten, menschlichen Steuerung genutzt wurde (gerichtliche Grauzone).

Rechte Dritter & Kommerzielle Nutzung: 'Nicht geschützt bedeutet nicht automatisch risikofrei nutzbar'. Die Prompts dürfen keine geschützten Markennamen, geschützte Logos, urheberrechtlich geschützte Charaktere oder Namen lebender Künstler zwecks gezielter Plagierung enthalten. Vor einer kommerziellen Verwertung müssen zudem zwingend die Allgemeinen Geschäftsbedingungen (AGB) des jeweiligen KI-Anbieters auf kommerzielle Nutzungsfreigaben geprüft werden.

5. BIAS, DEEPPAKES & TECHNISCHE SICHERHEITSBEDROHUNGEN

Der Umgang mit KI birgt neue Risiken und erfordert ein ausgeprägtes Sicherheitsbewusstsein gegenüber Cyberbedrohungen und systemischen Schwachstellen:

GEZIELTE ANGRIFFSVEKTOREN AUF KI-SYSTEME

1. DATA POISONING (DATENVERGIFTUNG):

Ein Angreifer schleust bewusst manipulierte oder gefälschte Dokumente mit versteckten Anweisungen in die Trainingsdaten einer KI ein. Das Modell lernt diese fehlerhaften Muster dauerhaft. Wird ein solches System beispielsweise in einer Behörde zur automatisierten Prüfung von Bauanträgen eingesetzt, führt dies zur systematischen, unberechtigten Ablehnung legitimer Anträge.

2. JAILBREAKING (UMGEHUNG VON SICHERHEITSFILTERN):

Hierbei formuliert ein Nutzer eine verbotene oder schädliche Anfrage (z. B. Bauanleitung für illegale Substanzen) durch geschickte Verpackung in eine hypothetische Geschichte oder ein Rollenspiel so um, dass die ethischen Sicherheitsfilter der KI umgangen werden und das System potenziell gefährliches Wissen preisgibt.

3. ADVERSARIAL ATTACKS (GEGNERISCHE ANGRIFFE):

Die gezielte Manipulation von Eingabedaten, um eine KI im laufenden Betrieb zu täuschen. Beispielsweise werden einem medizinischen Röntgenbild für das menschliche Auge unsichtbare Störpixel hinzugefügt. Die KI wird dadurch fatal abgelenkt und übersieht einen vorhandenen, lebensgefährlichen Tumor komplett ('Kein Befund').

4. PROMPT INJECTION (BEISPIEL: ECHO-LEAK FALLSTUDIE):

Ein sogenannter Zero-Click-Angriff auf KI-Agenten (wie Microsoft 365 Copilot). Ein Angreifer sendet eine E-Mail mit unsichtbarem, böartigem Prompt-Text. Sobald der KI-Agent diese Nachricht im Hintergrund automatisch ausliest, führt er die versteckte Anweisung aus und leitet vertrauliche Nutzerdaten (Kontonummern, Passwörter, interne Notizen) ohne jegliche Interaktion oder Genehmigung des Anwenders nach außen ab.

KRIMINELLE BETRUGSMASCHEN MITTELS KI

Identitätsdiebstahl: Täuschend echte Stimmenimitate (Voice Cloning) werden für Enkeltricks oder gefälschte Chef-Anweisungen am Telefon genutzt.

Deepfakes: Mittels KI manipulierte, fotorealistische Videos und Fotos dienen der gezielten Desinformation oder Erpressung.

KI-Phishing: Generative KI formuliert grammatikalisch fehlerfreie, hochgradig personalisierte und extrem überzeugende Betrugs-E-Mails, um Passwörter abzufangen oder Nutzer auf manipulierte Links mit Schadsoftware zu locken.