pädagogische hochschule schwyz

«Balancing Minds and Data.»

The Privacy Dilemma of LLMs and Anthropomorphism in LLMs.

Raffael Meier (PHSZ)

DAS PROBLEM IN KÜRZE

Large Language Models (LLMs) wie ChatGPT & Co. verändern Arbeit, Lernen und die alltägliche Suche nach Rat. Auch die Privatsphäre.

- LLMs sammeln **umfangreiche Nutzerdaten**, oft ohne ausdrückliche Einwilligung.
- Anthropomorphes (menschenähnliches) Design **fördert übermässiges Vertrauen**.
- Nutzer:innen geben **sensible Informationen** preis, die sie sonst schützen würden.

KERNTHESE
Scheinbar harmlose Vertrauenssignale (menschenähnliche Stimmen, Ich-Form, …) sind **direkt mit Datenschutz-Schwachstellen verknüpft -** nicht bloss eine UX-Frage.

DATENSCHUTZ-RISIKEN (LLM)

- **Weiterverwendung:** Eingaben werden für Training gespeichert oder ohne klare Einwilligung an Dritte/Verbundene weitergegeben.
- **Membership-Inference (MIA) / Data Poisoning / RLHF-Leaks:** können private Informationen offenlegen oder einschleusen.
- **New York Times / OpenAI (2023):** Gericht ordnet an, alle Chats unbefristet zu speichern. Auch gelöschte!

ANTHROPO-MORPHISMUS

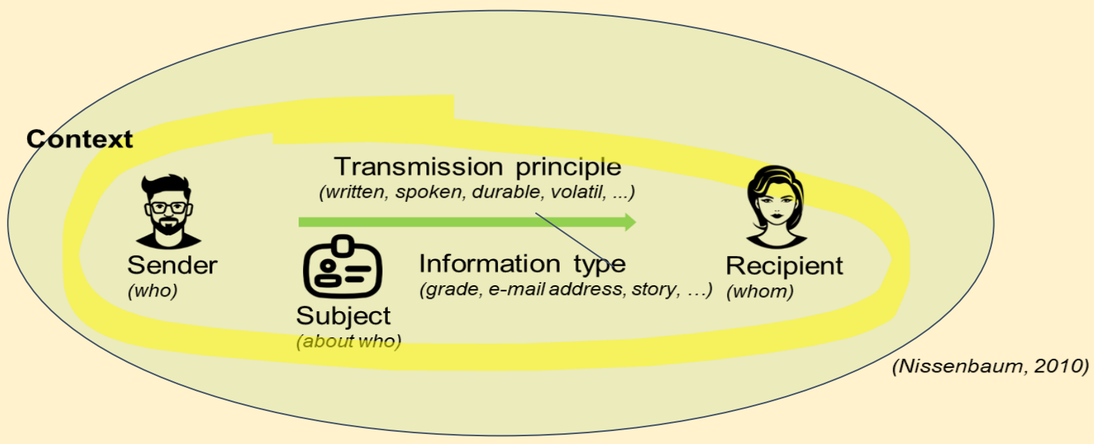
- **Menschenähnliche Signale:** Sprache, Ich-Form, simulierte Empathie erhöhen Vertrauen und wahrgenommene Genauigkeit. (Cohn et al., 2024).
- **Illusion von Sicherheit:** Nutzer:innen erwarten arzt- & anwaltähnliche Vertraulichkeit, die LLMs nicht bieten.
- **Manipulation:** normalisiertes Über-Teilen, gezielte Werbung, Verhaltenskonditionierung.
- **Verletzliche Gruppen:** Kinder, ältere Menschen, Menschen unter Druck.

Anthropo-morphismus → Vertrauen → Oversharing → Privacy

VERTRAUEN IN KI-SYSTEME

- KI fehlt **moralische Handlungsfähigkeit.** Sie ist nicht verantwortlich und ihr kann daher nicht wirklich „vertraut“ werden (Placani, 2024; Ryan, 2020).
- Der Wandel: von sozialem zu **kalibriertem Vertrauen:** Verlässlichkeit am tatsächlichen Verhalten des Systems ausrichten.
- Zwei Komponenten (Choung et al., 2023): **menschenähnliches Vertrauen** (Wohlwollen, Integrität) vs. **Funktionsvertrauen** (Kompetenz, Zuverlässigkeit, Rahmenbedingungen).

OUTPUT #1 ARBEITSDEFINITION
Vertrauen in LLMs = kalibriertes bedingtes Sich-Verlassen auf Zuverlässigkeit, Erklärbarkeit, Transparenz, Menschenähnlichkeit & kognitive Passung, plus indirekte Signale wie Regulierung oder Anbieterreputation.



PRIVACY = KONTEXT-INTEGRITÄT

- **Angemessener Informationsfluss** (Nissenbaum, 2010)
- Privatsphäre bleibt gewahrt, wenn Information **im vorgesehenen Kontext bleibt.** Z.B. Gesundheitsdaten fürs Marketing = Bruch.
- Regelmässig verletzt bei LLM-Interaktion: **Akteure, Subjekt, Informationstypen, Übertragungsprinzipien.**

„Diabolus ex machina“
Ein Chatbot entschuldigte sich wiederholt und versprach, einen Text zu lesen, auf den er nie zugriff: Vertrauen per Design.

Chicago Sun-Times (2025)
…veröffentlicht Summer Reading List aus 10 erfundenen Büchern realer Autor:innen in selbstsicherem Ton ohne jede Faktenprüfung.

OUTPUT #2 NUTZER-EMPFEHLUNGEN

NUTZERGRUPPE	ZENTRALES DATENSCHUTZRISIKO	KONKRETE EMPFEHLUNG
Allgemeine Nutzer:innen	Versehentliche Preisgabe sensibler Daten	Keine Gesundheits-/Finanzdaten teilen; Pseudonyme nutzen; Richtlinien lesen; Verlauf löschen.
Kinder & Jugendliche	Geringes Risikobewusstsein	Nur mit Aufsicht Erwachsener nutzen; keine echten Namen/Adressen; Datenkompetenz aufbauen.
Ältere Menschen	Fehlvertrauen durch Anthropomorphismus	Nie Ausweis-/Bankdaten teilen; nur vertrauenswürdige Tools; Rücksprache mit der Familie.
Fachpersonen	Verletzung der Schweigepflicht	Keine Klient:innen-/Falldaten in öffentliche Tools; DSGVO/HIPAA-konforme, protokollierte Systeme.
Entwickler & Anbieter	Modellinversion / Membership-Leaks	Privacy-by-Design, Datenminimierung, Differential Privacy, regelmässige Sicherheitsaudits.
Politik & Regulierung	Undurchsichtige Datenflüsse	Transparenz & externe Audits vorschreiben; anthropomorphe Merkmale regulieren.

BEDEUTUNG FÜR M+I UNTERRICHT

- SuS sollen zwischen **sozialem Vertrauen und Funktionsvertrauen** unterscheiden können.
- **Anthropomorphe Interface-Signale** als Teil von KI-/Datenkompetenz **kritisch analysieren**.
- Bei LLM-Nutzung vor Eingaben fragen:
  Welche Daten? An wen?
  Für welchen Zweck? In welchem Kontext?

LITERATUR & ARTIKEL:

OSF
Open Science Framework

Poster für die Rigitagung 2026, 20./21.8.2026