Thema 1: KI „Value Alignment“ (AI value alignment): Wie kann das Verhalten von KI-Systemen an relevante menschliche Werte angeglichen werden?
Bsp.

- Wie kann „Wahrheit“ als Wert operationalisiert werden (z. B. im Kontext von LLMs oder Chatbots)?
- Wertkonflikte: Wie priorisieren KI-Systeme widersprüchliche menschliche Werte (z. B. Fairness vs. Effizienz)?
- Wie unterscheiden sich Wertvorstellungen von Nutzern in unterschiedlichen Kulturen – und was bedeutet das für KI-Systeme?
- Analyse und Vergleich existierender Ansätze (z.B. Constitutional AI vs RLFH)

Studierende sollten dabei ein konkretes System, einen Anwendungsbereich oder einen Wert wählen. Der prinzipielle Technologieschwerpunkt liegt auf KI-Dialogsystemen (Conversational Agents), wie z.B. LLMs, aber andere KI-Anwendungen sind auch möglich. Die Arbeit kann auf Englisch (präferiert) oder Deutsch verfasst werden.