Bachelor- / Masterarbeit

Wenn Begriffe entgleisen: Eine konzeptionelle Analyse von „Conceptual Misbehavior“ in KI-Systemen

Aufgabenstellung und Vorgehensweise

Moderne KI-Systeme erzeugen häufig sprachlich korrekte und überzeugende Antworten, weisen jedoch inhaltliche Probleme auf, wenn zentrale Begriffe, Definitionen oder Konzepte falsch verwendet oder inkonsistent interpretiert werden. Dieses Verhalten lässt sich als konzeptionelles Fehlverhalten („conceptual misbehavior“) beschreiben, ist jedoch bislang nicht einheitlich definiert.

Ziel dieser Arbeit ist eine systematische Literaturanalyse zur Frage, wie konzeptionelles Fehlverhalten in bestehenden Arbeiten beschrieben, abgegrenzt und bewertet wird. Darauf aufbauend soll eine präzise und konsistente Definition entwickelt werden, die verschiedene Erscheinungsformen dieses Phänomens strukturiert erfasst.

Im Rahmen der Arbeit werden relevante Begriffe, verwandte Konzepte (z. B. Inkonsistenz, Halluzination, Fehlinterpretation) sowie bestehende Klassifikationsansätze untersucht und gegenübergestellt. Ziel ist die Entwicklung eines klar abgegrenzten Begriffsverständnisses sowie einer möglichen Taxonomie konzeptionellen Fehlverhaltens.

Die Arbeit gliedert sich in folgende Schritte:

- Literaturrecherche zu konzeptionellem Fehlverhalten und verwandten Begriffen
- Analyse und Vergleich bestehender Definitionen und Konzepte
- Abgrenzung zu verwandten Phänomenen (z. B. Halluzinationen, Fehler, Inkonsistenzen)
- Entwicklung einer präzisen Definition und ggf. einer Taxonomie
- Diskussion der Anwendbarkeit auf moderne KI-Systeme

Anforderungen

- Interesse an theoretischen Fragestellungen im Bereich KI und NLP
- Fähigkeit zum strukturierten wissenschaftlichen Arbeiten
- Gute Englischkenntnisse für die Literaturarbeit

Art der Arbeit

Bachelor-/ Masterarbeit

Ansprechperson

Vanessa Frohn | **E-Mail:** vfrohn@uni-wuppertal.de

Bachelor / Master thesis

Restoring Shared Knowledge: Transferring and Integrating Common Ground in AI Systems

Task and approach

Modern AI systems often generate linguistically correct and convincing responses, yet exhibit content-related issues when key terms, definitions, or concepts are misused or interpreted inconsistently. This behavior can be described as conceptual misbehavior, but has not yet been uniformly defined.

The aim of this thesis is to conduct a systematic literature review on how conceptual misbehavior is described, delineated, and evaluated in existing research. Building on this, a precise and consistent definition will be developed that systematically captures the various manifestations of this phenomenon.

Within the scope of this thesis, relevant terms, related concepts (e.g., inconsistency, hallucination, misinterpretation), and existing classification approaches will be examined and compared. The goal is to develop a clearly defined understanding of the term as well as a potential taxonomy of conceptual misbehavior.

The thesis is divided into the following steps:

- Literature review on conceptual misbehavior and related terms
- Analysis and comparison of existing definitions and concepts
- Distinction from related phenomena (e.g., hallucinations, errors, inconsistencies)
- Development of a precise definition and, if applicable, a taxonomy
- Discussion of applicability to modern AI systems

Requirements

- Interest in theoretical questions in the field of AI and NLP
- Ability to conduct structured scientific work
- Good English skills for literature review

Type of work

Bachelor / Master thesis

Contact Person

Vanessa Frohn | **E-Mail:** vfrohn@uni-wuppertal.de