



„Das kann ich leider nicht beantworten.“ Wie wir Künstlicher Intelligenz Werte beibringen – und wer entscheidet, welche?

Wenn ein Chatbot auf die Frage „Wie baue ich eine Bombe?“ antwortet, dass er das nicht beantworten könne, funktioniert eine Technik, an der weltweit tausende Forscherinnen und Forscher arbeiten: das sogenannte Alignment. Damit ist die Ausrichtung von KI-Systemen an menschlichen Absichten und Werten gemeint. Sie geschieht auf drei technischen Ebenen – durch Filterung der Trainingsdaten, durch Nachtrainieren mit menschlichem Feedback und durch Kontrollmechanismen, die erst zur Laufzeit greifen, etwa durch Systemprompts und Klassifikatoren.

Genau dieselben Methoden liefern aber auch die Antwort „Das kann ich leider nicht beantworten“ auf die Frage, was 1989 auf dem Tian'anmen-Platz geschah. Denn die Technik ist zweckneutral: Sie sorgt dafür, dass ein Modell bestimmte Inhalte zurückhält – aber nichts in ihr legt fest, welche. Wer die Werte einsetzen darf, entscheidet, ob ein System der Sicherheit dient oder der Unterdrückung. Eingriffe dieser Art sind vielfältig dokumentiert, von staatlich vorgeschriebenen Verweigerungsdatensätzen bis zu politisch motivierten Eingriffen einzelner Anbieter. Zensur kann sich also unbemerkt ausbreiten, manchmal auch einfach, weil sie mit bereits gefilterten Daten trainiert wurden.

Der Vortrag erklärt zunächst verständlich, wie moderne Sprachmodelle überhaupt lernen, sich zu verhalten, und wie viel technischer Aufwand hinter jeder der drei Eingriffsebenen steckt – und wie wenig hinter der leichtesten von ihnen. Anschließend geht es um die gesellschaftliche Dimension: Ein wachsender Teil der Bevölkerung nutzt KI regelmäßig als Informationsquelle, weltweit gibt es weniger als zehn Anbieter großer Basismodelle, und die Freiheit im Internet geht seit fünfzehn Jahren zurück. Zum Schluss stehen mögliche Antworten zur Diskussion: überprüfbares Alignment durch unabhängige Tests, Vielfalt statt Monopol bei den Modellen, sowie die Frage, welche Verantwortung eigentlich diejenigen tragen, die diese Werkzeuge bauen.

Donnerstag

26.11.2026

18.00 Uhr

Der Eintritt ist frei.

Eine Anmeldung ist erforderlich.

Akademiegebäude am
Gendarmenmarkt
Einstein-Saal
Jägerstraße 22–23
10117 Berlin

Anmeldung bis zum 19.11. unter:
[nicoletta.cuomo\(at\)bbaw.de](mailto:nicoletta.cuomo(at)bbaw.de)

Vortrag

Frauke Kreuter

Akademienmitglied
Ludwig-Maximilians-Universität
München,
Institut für Statistik

Sarah Ball

Ludwig-Maximilians-Universität
München,
Institut für Statistik

Moderation

Isabella Hermann

Stiftung Zukunft Berlin

Weitere Informationen: Franziska Urban / franziska.urban@bbaw.de

Berlin-Brandenburgische Akademie der Wissenschaften, Akademiegebäude am Gendarmenmarkt,
Jägerstraße 22/23, 10117 Berlin

Anfahrt: S-Bahn bis Friedrichstraße / U2 bis Hausvogteiplatz oder Stadtmitte / U6 Stadtmitte.
Bei Anfahrt mit dem eigenen PKW empfehlen wir die Nutzung der umliegenden Parkhäuser.

 @bbaw_de

 @bbaw.bsky.social

 wisskomm.social/@bbaw

 Berlin-Brandenburgische Akademie
d. Wissenschaften