Information »Alignment-Problem «

Forscher und Entwickler in der Künstlichen Intelligenz beschäftigen oftmals eine wichtige Frage: Wie kann sichergestellt werden, dass KI-Systeme das tun, was wir wirklich von ihnen wollen?

Diese Herausforderung wird als "Alignment-Problem" bezeichnet und ist von großer Bedeutung in der Entwicklung und Anwendung von KI-Systemen. Es beschreibt allgemein die Schwierigkeit, KI-Systeme so zu gestalten, dass sie mit menschlichen Werten, Zielen und ethischen Prinzipien übereinstimmen. Dieses Problem wird immer wichtiger, je leistungsfähiger KI-Systeme werden.

Schon heute begegnet man dem Alignment-Problem im täglichen Leben. Ein Beispiel sind die Empfehlungssysteme in sozialen Medien. Diese Algorithmen sollen uns interessante Inhalte zeigen, empfehlen aber manchmal Dinge vor, die uns nicht wirklich weiterbringen oder sogar schädlich sein können. Bei immer komplexeren Anforderungen und Entscheidungen in der KI könnte das in der Zukunft gravierende Auswirkungen in kritischen Anwendungsfällen haben.