volks— hochschule beider basel

04./05.12.2024

Können wir Künstlicher Intelligenz vertrauen?

Malte Helmert, Prof.Dr.

Zusammenfassung

Das Thema "Künstliche Intelligenz" (KI) ist derzeit in aller Munde.
Fast täglich hören wir von neuen Fortschritten auf diesem Forschungsgebiet, von hohen Investitionen, die in diesem Bereich getätigt werden, und von neuen Anwendungsfeldern, in denen in Zukunft künstlich intelligente Systeme die menschliche Expertise ersetzen oder zumindest unterstützen sollen. Spätestens wenn es dabei darum geht, Verantwortung für wichtige Entscheidungen an Computer zu delegieren, kann einem dabei durchaus unwohl werden.
Können wir künstlich intelligenten Systemen vertrauen? Wie kommen sie zu Ihren Entscheidungen? Basieren diese auf Denkprozessen, die denen eines Menschen gleichen? Wenn nicht, basieren sie wenigstens auf einer objektiven Grundlage? Verstehen überhaupt die KI-Entwickler, wie ihre Systeme funktionieren? Oder sind künstlich intelligente Systeme so autonom, dass niemand sie kontrollieren kann?
In diesem Vortrag klären wir zunächst die Frage, was künstliche Intelligenz eigentlich ist. Dabei stellen wir fest, dass eine einheitliche Definition nicht existiert und treffen die wichtige Unterscheidung zwischen deliberativen (modellbasierten) KI-Ansätzen, die auf einer mathematischen Formalisierung realer Probleme arbeiten und streng logischen Grundsätzen folgen, und induktiven (lernenden) KI-Ansätzen, die anhand von Beispieldaten Muster erkennen und generalisieren, die sie dann in neuen, nicht vorprogrammierten Situationen anwenden können.
Anschliessend beleuchten wir die Vertrauenswürdigkeit von KI-Systemen von verschiedenen Seiten. Ist ihr Verhalten vorhersagbar? Welche Aufgaben sollen sie eigentlich lösen? Welchen Einfluss haben die Daten, auf denen sie trainiert werden? Können wir uns gegen Programmierfehler oder böswillige Manipulationen schützen? Welchen Einfluss haben die Computersysteme, auf denen KI-Verfahren ausgeführt werden? Und wenn wir alle technischen Fragen zur Vertrauenswürdigkeit zu unserer Zufriedenheit beantworten können, reicht das dann aus, um den praktischen Einsatz von KI-Systemen in sensiblen Anwendungsgebieten zu rechtfertigen?

Literatur und Internetlinks

1. Stuart Russell und Peter Novig. Künstliche Intelligenz: ein moderner Ansatz. Vierte deutsche Ausgabe, Pearson, 2023.
2. Daniel Kahneman. Schnelles Denken, langsames Denken. Deutsche Ausgabe, Siedler Verlag, 2012.
3. Adnan Darwiche. Human-level intelligence or animal-like abilities? Communications of the ACM 61(10), 2018.
4. Ken Thompson. Reflections on Trusting Trust. Communications of the ACM 27(8), 1984.

Kontakt

Malte Helmert, Prof. Dr.
malte.helmert@unibas.ch

Universität Basel