idw - Informationsdienst Wissenschaft
Nachrichten, Termine, Experten

Press release

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, DFKI
Jennifer Oberhofer

01/29/2025
http://idw-online.de/en/news846571

Research results, Scientific Publications
Economics / business administration, Environment / ecology, Information technology, Politics, Social studies
transregional, national

AI Safety Report erscheint als Auftakt zum AI Action Summit in Paris

Der 2023 von der britischen Regierung initiierte „International Scientific Report on the Safety of Advanced AI“ wird heute in seiner Abschlussversion veröffentlicht. Mitgewirkt hat auch der CEO des Deutschen Forschungszentrums für Künstliche Intelligenz (DFKI), Prof. Antonio Krüger. Unter den 96 beteiligten KI-Experten ist er der von der Bundesregierung ernannte deutsche Vertreter. Im Bericht geht es um Chancen, aber auch um Herausforderungen der immer fähiger werdenden KI-Systeme und Lösungsansätze. Die weltumspannende, gemeinschaftliche Auseinandersetzung mit KI liefert wertvolle Anknüpfungspunkte für die künftige Forschung.

Der AI Safety Report analysiert aktuelle und künftige KI-Systeme, die ein breites Spektrum an Aufgaben erledigen können, sogenannte „General-Purpose AI“. Die Fähigkeiten dieser KI-Systeme, die für viele Menschen durch die Anwendung ChatGPT erstmals erlebbar wurden, haben sich in den letzten Monaten rapide verbessert. Die Experten verlieren die Chancen nicht aus dem Blick, verweisen aber im Bericht auch auf Risiken, wie solche, die sich durch Fehlfunktionen ergeben können.

Antonio Krüger: „Der Bericht liefert eine gute, umfassende Bestandsaufnahme auch zu den Risiken von KI. Die größte Gefahr sehe ich bei den fehlerbehafteten KI-Systemen. KI, die nicht in unserem Sinne funktioniert, könnte, als wichtiger Bestandteil des Alltags, die Gesellschaft schleichend beeinflussen.“

Genannt wird die mangelnde Zuverlässigkeit aktueller Systeme. KI kann Schäden anrichten, wenn sie halluziniert, falsche Informationen ausgibt, Kausalitäten nicht kennt oder kein Kontextwissen heranzieht. Als Schwachstelle von General-Purpose AI wird im Bericht auch Bias aufgeführt. Voreingenommene KI-Systeme produzieren verzerrte Ergebnisse, weil sie zum Beispiel vornehmlich mit englischen Sprachdaten trainiert werden und westlich geprägt sind. Auch ein möglicher Kontrollverlust des Menschen über die Technologie wird als kontroverser Punkt aufgegriffen.

Zu den weiteren Herausforderungen zählt der Report systemische Risiken, wie eine globale KI-Schere zwischen den Staaten, nachteilige Umwelteinflüsse, Verletzungen der Privatsphäre, Urheberrechtsfragen, aber auch die böswillige Verwendung der Technologie durch Kriminelle.

„Bereiche, die in der KI-Forschung unter dem Begriff Vertrauenswürdige KI behandelt werden, sollten bei der Entwicklung und dem Betrieb von KI in Betracht gezogen werden. Für die Zuverlässigkeit von KI könnten insbesondere neurosymbolische Systeme, also die Kombination von datengetriebenen Bausteinen mit formalisiertem Weltwissen, ein großer Gewinn sein“ erklärt Krüger.

Im Bericht wird darauf hingewiesen, dass die Wahrung der Privatsphäre über den gesamten Lebenszyklus eines KI-Systems hinweg sicherzustellen ist. Risikovermeidung bei General-Purpose AI kann nicht erst beim fertigen KI-Modell ansetzen, sondern muss bereits bei der Entwicklung mit einem Training beginnen, das, zum Beispiel durch die Steigerung der Datenqualität, robustere Ergebnisse erzeugt. Zusätzlich soll das KI-System im laufenden Betrieb überwacht werden (Monitoring). Hierfür kann die Erklärbarkeit eines Systems weiterhelfen.

Das DFKI widmet sich den noch vielen offenen Fragen bereits in mehreren Initiativen.

CERTAIN, das „Centre for European Research in Trusted AI“ arbeitet auf Garantien für sichere KI hin und erforscht die verschiedenen Aspekte vertrauenswürdiger KI, darunter auch Human Oversight: Forschende fragen sich, wann ein Mensch in ein laufendes System eingreifen kann und haben sich die technischen (Erklärbarkeit) und die nicht-technischen Voraussetzungen dafür angesehen[1].

In Mission KI steht die Zertifizierung von KI im Mittelpunkt, dabei wird besonders auf den Mittelstand eingegangen.

Wissenschaftler am Standort Darmstadt untersuchen wie KI-Systeme, in die General-Purpose AI eingebaut ist, gestaltet werden müssen, damit die Ausgaben weniger Bias und unerwünschte Konzepte enthalten[2]. Sie stellten zum Beispiel fest, dass sich die Antworten eines Sprachmodells auf sicherheitskritische Fragen je nach Sprache gravierend unterscheiden[3]. Von Interesse ist auch, dass KI-Agenten mit ihren Aktionen keinen Schaden anrichten.

„Teil der deutschen und europäischen KI-Strategie muss es sein, stärker in sicherheitsrelevante Technologien zu investieren, auf Innovationsförderung als Lösung zu setzen und nicht nur mithilfe von Regulatorik Risiken eindämmen zu wollen“, so Krüger.

Der Bericht ist eine Vorarbeit für den AI Action Summit, der in zwei Wochen, am 10. und 11. Februar 2025, in Paris stattfindet.

---
[1] https://dl.acm.org/doi/10.1145/3630106.3659051
[2] https://arxiv.org/abs/2211.05105
[3] https://arxiv.org/abs/2412.15035

Über das DFKI:
Das Deutsche Forschungszentrum für Künstliche Intelligenz GmbH (DFKI) wurde 1988 als gemeinnützige Public-Private-Partnership (PPP) gegründet. Es unterhält Standorte in Kaiserslautern, Saarbrücken, Bremen, Niedersachsen und Darmstadt, Labore in Berlin und Lübeck, sowie eine Außenstelle in Trier. Das DFKI verbindet wissenschaftliche Spitzenleistung und wirtschaftsnahe Wertschöpfung mit gesellschaftlicher Wertschätzung. Das DFKI forscht seit über 35 Jahren an KI für den Menschen und orientiert sich an gesellschaftlicher Relevanz und wissenschaftlicher Exzellenz in den entscheidenden zukunftsorientierten Forschungs- und Anwendungsgebieten der Künstlichen Intelligenz.

Material:
Unter https://cloud.dfki.de/owncloud/index.php/s/girRD85wRaMwkBW finden Sie Bildmaterial und den Report. Bitte beachten Sie bei der Verwendung der Fotos das Copyright. Für die Grafik ist dies „DFKI“.

DFKI-Pressekontakt:
Jennifer Oberhofer
Communications & Media
Tel.: +49 541 386050 7088
E-Mail: jennifer.oberhofer@dfki.de

Der AI Safety Report ist das Ergebnis internationaler Kollaboration.
DFKI

DFKI CEO Prof. Antonio Krüger ist Teil des internationalen Expertengremiums, das an der Erstellung des AI Safety Reports beteiligt war.
DFKI, Jürgen Mai