Kira – Zentrum für KI-Risiken & -Auswirkungen gUG (haftungsbeschränkt) | Prenzlauer Allee 186 | 10405 Berlin

28.11.2023

– Brief an die Bundesregierung –

Der europäische AI Act braucht Grundlagenmodell-Regulierung

- Dieser Brief ging zuerst am 20.11. an den Bundeskanzler sowie die Bundesminister für Wirtschaft und Klimaschutz sowie Digitales und Verkehr -

Sehr geehrte Damen und Herren,

wir sind eine Gruppe deutscher und internationaler Expertinnen und Experten für Künstliche Intelligenz (KI) sowie Führungskräfte in Wirtschaft, Zivilgesellschaft und Wissenschaft. Unser Fachwissen ist vielfältig und relevant: Unter uns sind die beiden weltweit am häufigsten zitierten KI-Forscher und Turing-Preis-Gewinner („Nobelpreis für Computerwissenschaften“); ein Autor des weltweit am meisten genutzten KI-Lehrbuchs; Fachleute, die die Bundesregierung in KI-Fragen und verwandten Themen beraten haben; sowie Gründer erfolgreicher KI-Unternehmen.

Der AI Act der Europäischen Union steht kurz vor der Fertigstellung und wird derzeit im Trilog verhandelt. Unserer gemeinsamen Einschätzung nach könnte der AI Act zu einer bahnbrechenden Gesetzgebung werden, die die Zukunft von KI nicht nur in der EU, sondern weltweit prägt. Wir begrüßen, dass sich die EU und ihre Mitgliedsstaaten dieses wichtigen Themas auf ernsthafte und angemessene Weise annehmen und dabei eine globale Führungsrolle einnehmen.

Das Potenzial des Gesetzes ist unserer Meinung nach nun aber in Gefahr: Ein zentrales Element des AI Act, nämlich verbindliche Regeln für Grundlagenmodelle, stößt bei einigen Mitgliedstaaten auf Widerstand. Unserem Verständnis nach ist die deutsche Bundesregierung Teil dieser Opposition.

Wir glauben, dass die Regulierung von Grundlagenmodellen im AI Act für ein florierendes und sicheres KI-Ökosystem entscheidend ist. Wir raten dringend davon ab, bei Grundlagenmodellen lediglich auf ein System der Selbstregulierung zu setzen.

Viele der weltweit angesehensten KI-Fachleute haben zuletzt vor den vielfältigen Risiken fortgeschrittener KI gewarnt, darunter Risiken für die öffentliche Sicherheit wie KI-generierte Desinformation und Manipulation, KI-gestützte Cyberangriffe oder KI-generierte Pathogene. Solche Risiken gehen primär von den leistungsfähigsten Grundlagenmodellen aus; diese Erkenntnis schlägt sich auch in der jüngsten Executive Order zu KI des Weißen Hauses und in der historische Bletchley-Erklärung, die diesen Monat von 28 Ländern und der EU unterzeichnet wurde, nieder. Diese Risiken für die öffentliche Sicherheit sind Grundlagenmodellen *inhärent*. Deshalb sollten sie auf Ebene der Grundlagenmodelle adressiert werden. Die vom

EU-Parlament und der spanischen Ratspräsidentschaft vorgesehenen Regelungen für Grundlagenmodelle – z.B. hinsichtlich (Cyber-)Sicherheit, Risikobewertungs- und Minderungs-systemen, Red-Teamings vor der Veröffentlichung und Audits nach der Veröffentlichung – sind daher für ein florierendes und sicheres KI-Ökosystem in der EU unerlässlich, weil sie solche inhärenten Risiken adressieren.

Solche verbindlichen Regeln sind sowohl aus ökonomischen als auch aus Sicherheitsgründen wichtig. Ökonomisch betrachtet ist die Sicherheit von Grundlagenmodellen eine notwendige Voraussetzung für Tausende von KMUs und andere nachgelagerte Anwender, die diese Grundlagenmodelle für ihre innovativen Produkte nutzen möchten. Sie können sich Haftungsrisiken und exzessive Compliance-Kosten, die sich aus einem potenziell unsicheren, ihrem Produkt zugrunde liegenden Grundlagenmodell ergeben, nicht leisten. Grundlagenmodelle aus dem AI Act auszunehmen, würde Innovation deshalb erheblich hemmen.

Auch aus Sicherheitsgründen ist es entscheidend, Risiken auf der Ebene der Grundlagenmodelle zu adressieren. Nur die Anbieter von Grundlagenmodellen können die ihnen inhärenten Risiken umfassend adressieren. Sie allein haben Zugang zu bzw. Wissen über die Trainingsdaten ihrer Modelle, deren Sicherheitsschranken, wahrscheinliche Schwachstellen und andere zentrale Eigenschaften. Wenn schwerwiegende Risiken von Grundlagenmodellen nicht auf der Ebene der Grundlagenmodelle mitigiert werden, werden sie überhaupt nicht mitigiert, was potenziell die Sicherheit von Millionen von Menschen gefährdet.

Wir wissen, dass einige Stimmen dafür plädieren, Risiken von Grundlagenmodellen durch ein System der Selbstregulierung zu adressieren. Wir raten dringend davon ab. Eine Selbstregulierung würde wahrscheinlich auf dramatische Weise hinter den Standards zurückbleiben, die für die Sicherheit von Grundlagenmodellen nötig sind. Da selbst ein einzelnes unsicheres Modell Risiken für die öffentliche Sicherheit bergen kann, reicht ein fragiler Konsens zur Selbstregulierung nicht aus, um die Sicherheit der EU-Bürgerinnen und -Bürger zu gewährleisten. Die Sicherheit von Grundlagenmodellen muss gesetzlich verankert werden.

Ein AI Act, der Grundlagenmodelle einschließt, wäre die weltweit erste umfassende KI-Regulierung und ein historisches Beispiel europäischer Vorreiterschaft. Falls Grundlagenmodellen nicht behandelt werden, würde ein geschwächter oder gescheiterter AI Act als historischer Misserfolg betrachtet werden.

Wir ermutigen daher die Bundesregierung nachdrücklich, ihre Führungsrolle in der Europäischen Union zu nutzen und eine umfassende Regulierung von Grundlagenmodellen im AI Act sicherzustellen.

Mit freundlichen Grüßen

Prof. em. Geoffrey Hinton, University of Toronto, Chief Scientific Adviser at the Vector Institute, 2018 Turing Award Winner

Prof. Yoshua Bengio, Université de Montréal, Founder and Scientific Director of Mila – Quebec AI Institute, 2018 Turing Award Winner

Prof. em. Gary Marcus, NYU, Founder and CEO, Geometric Intelligence (acquired by Uber)

Prof. Stuart Russell, UC Berkeley, Director of the Center for Human-Compatible Artificial Intelligence, co-author of the standard textbook “Artificial Intelligence: a Modern Approach"

Marietje Schaake, International Policy Fellow, Stanford Institute for Human-Centered Artificial Intelligence

Andreas Loy, Founder & CEO, KONUX

Prof. Holger Hoos, RWTH Aachen University & University of Leiden

Prof. em. Raja Chatila, Sorbonne University

Prof. Dr. jur. Silja Vöneky, Universität Freiburg

Prof. Karl Hans Bläsius, Hochschule Trier

Prof. Wolfgang Schröder, Julius-Maximilians-Universität Würzburg

Prof. Christoph Benzmüller, Chair for AI Systems Development, Otto-Friedrich-Universität Bamberg

Prof. Gerhard Lakemeyer, Chair, Department of Computer Science, RWTH Aachen

Prof. Otthein Herzog, Universität Bremen

Prof. Mathias Risse, Harvard University, Director of the Carr Center for Human Rights Policy

Prof. Marius Lindauer, Leibniz Universität Hannover

Prof. Katharina Morik, TU Dortmund, AI Chair (emerita)

Prof. Wil van der Aalst, RWTH Aachen

Kaltrina Shala LL.M., LL.M., Weizenbaum-Institut e.V.

Prof. Peter Struss, TU Munich

Prof. Kai-Uwe Kühnberger, University Professor for Artificial Intelligence, Osnabrück University

Prof. Dr. Claus Rollinger, Universität Osnabrück