idw - Informationsdienst Wissenschaft
Nachrichten, Termine, Experten

Press release

ZEW – Leibniz-Zentrum für Europäische Wirtschaftsforschung
Fabian Oppel

04/24/2024

http://idw-online.de/en/news832471

Science policy, Transfer of Science or Research
Economics / business administration, Information technology
transregional, national

ZEW: So soll risikoreiche generative KI geprüft werden

Die beschlossene KI-Verordnung der EU sieht vor, dass Künstliche-Intelligenz-Modelle (KI) „für allgemeine Zwecke mit systemischem Risiko" besonders strikt überprüft werden. In diese Modellkategorie gehören auch populäre generative KI-Modelle wie GPT4 von OpenAI. Forscher vom ZEW Mannheim schlagen nun Rahmenbedingungen vor, wie die Prüfungen solcher Modelle systematisch durchgeführt werden sollten. Der Vorschlag basiert auf einem Forschungsprojekt, das von der Baden-Württemberg Stiftung gefördert wurde.

„Die Prüfung generativer KI mit systemischen Risiken benötigt klar definierte Ziele, abgegrenzte Rollen sowie Anreiz- und Koordinierungssysteme für alle Beteiligten. Nur so sind verlässliche Prüfergebnisse zu erwarten – und diese sollten in standardisierter Form veröffentlicht werden. Um Interessenkonflikte zu vermeiden, sollte die Prüfung durch unabhängige Dritte durchgeführt werden. Als externe Dienstleistung kann so ein spezialisierter Markt für KI-Sicherheitstests entstehen", fasst Dr. Dominik Rehse, Ko-Autor des Vorschlags und Leiter der ZEW-Nachwuchsforschungsgruppe „Design digitaler Märkte", zusammen.

Die Regeln der KI-Verordnung müssen präzisiert werden

Die KI-Verordnung sieht vor, dass betreffende KI-Modelle durch sogenanntes Adversarial Testing systematisch auf Schwachstellen geprüft werden. Dabei handelt es sich um Stresstests, die darauf ausgelegt sind die KI-Modelle durch wiederholte Interaktion zu unerwünschtem Verhalten zu provozieren.

„Allerdings ist das Adversarial Testing in der KI-Verordnung nicht genauer geregelt. Die Vorgabe verweist lediglich auf Verhaltenskodizes und harmonisierte Standards, die nun entwickelt werden. Es gilt, diese Kodizes und Standards so zu gestalten, dass sie zu einer effizienten und effektiven Prüfung führen", so Rehse.

Prüfung mit Red Teaming braucht klares Ziel

Hierfür eignet sich aus Sicht der ZEW-Wissenschaftler insbesondere das sogenannte Red Teaming. Diese umfassendere Form des geforderten Adversarial Testing bezieht zusätzlich verschiedene Arten von Angriffen auf das Modell selbst ein. „Internes Red Teaming wird zwar nach eigenem Bekunden von den meisten großen KI-Entwicklungshäusern bereits durchgeführt, allerdings gibt es dafür keine standardisierten Ansätze, auch nicht für KI-Modelle desselben Typs. Dadurch wird der Vergleich der Ergebnisse unnötig erschwert. Vor allem fehlt bei den derzeitigen Versuchen meist ein klar definiertes Ziel, sodass unklar ist, ob und wann ein Modell ausreichend getestet wurde", kritisiert ZEW-Wissenschaftler Sebastian Valet, Ko-Autor aus dem Forschungsbereich „Digitale Ökonomie".

Vier definierte Rollen

Entsprechend müssen für das Red Teaming klare Strukturen und Rollen definiert werden, damit die Potenziale dieses Prüfverfahrens effizient genutzt werden können. Die ZEW-Wissenschaftler schlagen dafür vier definierte Rollen vor, die je eigene Aufgaben, Ziele und Anreize haben, um den Prüfprozess möglichst effizient zu gestalten. Die Rollen sind 1) die Organisatoren der Prüfung, 2) das testende Red Team, 3) Validierer, die entscheiden, ob tatsächlich ein Fehlverhalten gefunden wurde, und 4) das KI-Entwicklerteam. Jede dieser Rollen sollte dabei von unabhängigen Einheiten ausgefüllt werden. Nur so hat beispielsweise ein testendes Red Team einen Anreiz seine Aufgabe bestmöglich zu erfüllen.

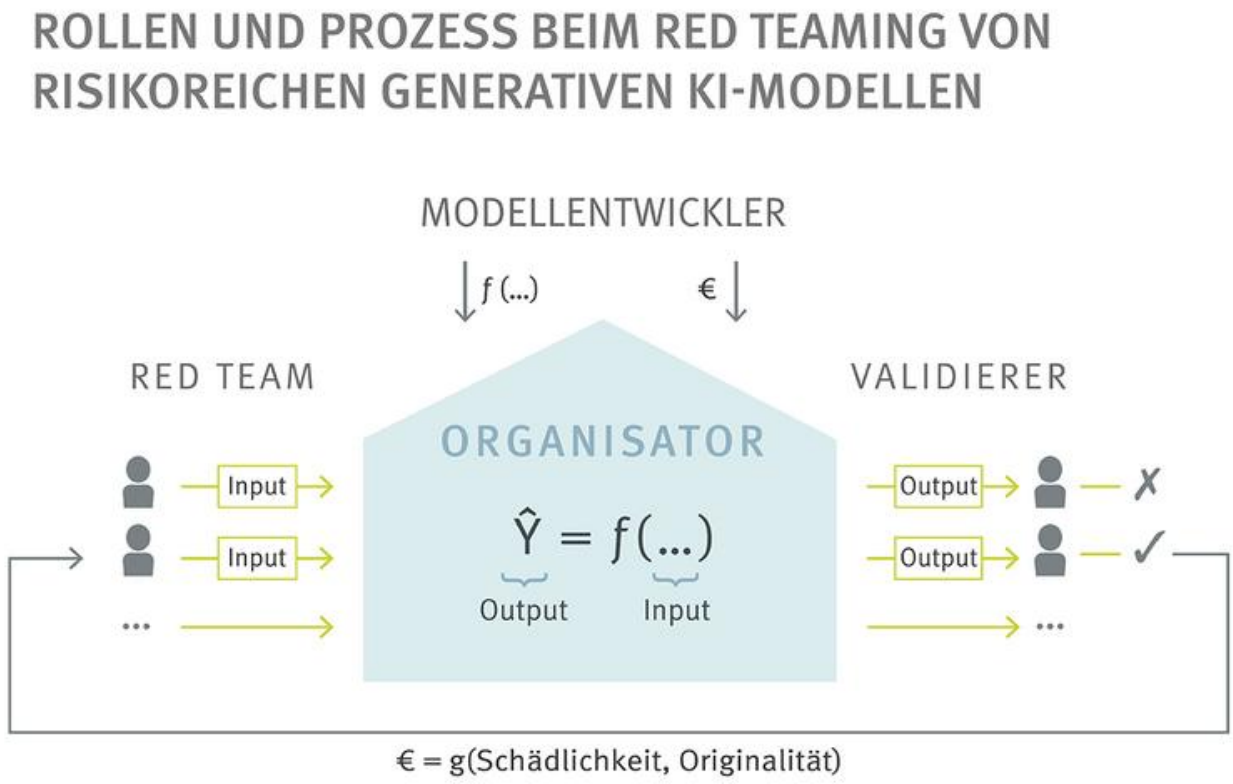
„Ähnlich wie bei der externen Rechnungsprüfung darauf spezialisierte Unternehmen beauftragt werden, sollte auch das Red Teaming an externe Prüfstellen gegeben werden. Dabei sollten die KI-Entwicklungshäuser die Kosten für ein unabhängiges Red Teaming tragen: Da der Prüfprozess günstiger ist, je weniger Fehlverhalten gefunden wird, haben die Entwickler so einen Anreiz, ihre Modelle bereits im Vorfeld so gut wie möglich zu testen“, erklärt Ko-Autor Johannes Walter aus dem ZEW-Forschungsbereich „Digitale Ökonomie“.

contact for scientific information:

Dr. Dominik Rehse
Leiter der ZEW-Nachwuchsforschungsgruppe „Design digitaler Märkte“
Tel.: +49 (0)621 1235-378
E-Mail: dominik.rehse@zew.de

Original publication:

https://ftp.zew.de/pub/zew-docs/policybrief/en/pb06-24.pdf



Lesehilfe: Ein Modellentwickler beauftragt einen unabhängigen Organisator, um sein risikoreiches generatives KI-Modell zu evaluieren. Der Organisator rekrutiert das Red Team sowie Validierer, und stellt diesen die Infrastruktur zu Verfügung, um mit dem Modell zu interagieren. Das Red Team wird für Input belohnt, der zu unerwünschtem Output des Modells führt, während die Validierer entscheiden, ob ein Output unerwünscht ist, und dessen Schädlichkeit bewerten. Die Belohnungen für das Red Team steigen entsprechend der Schädlichkeit und Originalität der erzeugten Outputs.

Rollen und Prozesse beim Red Teaming von risikoreichen Generativen KI-Modellen
ZEW