Liebe Mitglieder des Netzwerkes der ehemaligen SSAB, liebe Interessierte

Der AI Act der EU, das weltweit erste staatenübergreifende Gesetzeswerk zum Einsatz von künstlicher Intelligenz, wird stufenweise gemäss einem Zeitplan umgesetzt. Aus den Gesetzesvorgaben müssen vorerst konkrete, mess- und überprüfbare technische Anforderungen abgeleitet werden. Das ist nicht einfach.

ETHZ entwickelt mit COMPL-AI ein Umsetzungsmodell
Gemeinsam mit dem bulgarischen KI-Institut INSAIT, bei dessen Gründung 2022 ETHZ und EPFL mitwirkten, hat das ETH Spin-off LatticeFlow AI mit COMPL-AI ein Modell vorgelegt, mit dem sich messen lässt, ob eine untersuchte KI-Anwendung den Anforderungen des AI Act der EU entspricht. COMPL-AI ist das weltweit erste Tool zur Evaluierung regulatorischer Vorgaben.

Mit COMPL-AI wurden zwölf prominente grosse Sprachmodelle – darunter ChatGPT von OpenAI, Llama von Meta, Claude von Anthropic – untersucht. Defizite wurden bei allen festgestellt, insbesondere bei den Kriterien Transparenz, Sicherheit, Datenschutz, Fairness und auch bei anderen Anforderungen des AI Act der EU (s. ETH News vom 21.10.2024). Die Studie wurde auf der Plattform arXiv der Cornell University veröffentlicht.

Die EU-Kommission begrüsste das als Open Source zur Verfügung gestellte Tool als ersten Schritt zur Umsetzung. COMPL-AI steht auch auf GitHub zum Download bereit. Prof. Martin Vechev, ETHZ, Mitgründer und wissenschaftlicher Direktor von INSAIT, ermutigte andere Forschungsgruppen und Praktiker, das Open-Source-Tool weiterzuentwickeln. Die mit COMPL-AI erarbeitete Methodik sei auch auf andere vergleichbare Gesetzgebungen anpassbar. Das Tool ermögliche es nun allen Anbietern, ihre KI-Modelle selber zu überprüfen und Massnahmen zu treffen, damit sie künftigen Gesetzesanforderungen entsprechen (s. Information auf der Homepage von LatticeFlow AI).

Vertrauen in KI – ein wichtiges Thema an ETHZ und EPFL
Eine verlässliche, vertrauenswürdige KI-Anwendung entsteht im Zusammenwirken von Technik und ethischen Prinzipien und Werten. Die Voraussetzungen für eine vertrauenswürdige KI sind an den beiden Eidgenössischen Technischen Hochschulen, ETHZ und EPFL, denn auch ein sehr wichtiges, disziplinübergreifendes Thema. Erforderlich seien nicht nur entsprechende Prinzipien. Eine KI muss auch technisch überprüfbar sein, dies gilt insbesondere für heikle Anwendungsgebiete, z.B. Diagnosen in der Medizin.

Im Rahmen der Swiss AI Initiative und dem im Oktober 2024 gegründeten Swiss National AI Institute (SNAI) entwickeln zurzeit über 650 Forschende der ETHZ, EPFL und zehn weiterer Schweizer Hochschulen ein grosses Schweizer Sprachmodell als Open-Source-Basis für Weiterentwicklungen von Unternehmen und Start-ups. Nicht nur Quellcodes, auch Trainingsdaten und wichtige Parameter werden frei zugänglich sein (s. ETH News vom 31.3.2025).

Gemäss einer ETH News vom 28.3.2025 hat der Mathematiker Juan Gamella Mini-Labors entwickelt, in denen sich Algorithmen und KI-Modelle in einer Testumgebung mit echten Messdaten überprüfen lassen. Oftmals funktionieren KI-Anwendungen unter realen Bedingungen nicht wunschgemäss und müssen nachgebessert werden. Wie Flugzeugbauer, die ein am Computer entworfenes Flugzeug als Miniaturmodell zuerst im Windkanal prüfen, können Forschende ihre KI-Anwendungen in den Mini-Labors mit den Tests sicherer machen. Zudem könnten die Mini-Labors auch genutzt werden, um Zusammenhänge von Ursachen und Wirkungen, also Kausalzusammenhänge, zu erforschen und in Algorithmen abzubilden.

Forschende der EPFL – so in ihrer News vom 10.4.2025 – haben ein bahnbrechendes Tool entwickelt, dank dem die Aussagen einer KI robuster und damit vertrauenswürdiger werden, auch wenn die Trainingsdaten, wie dies meistens der Fall ist, viele falsche oder irritierende Angaben enthalten. Beim Tool handelt es sich um das Ergebnis jahrelanger Forschungsarbeit, beschrieben in einer von Prof. Rachid Guerraoui mit anderen Autoren herausgegebenen Publikation. Die Forschenden sind überzeugt, dass die Schweiz mit dieser Innovation für eine sicherere KI eine Vorreiterrolle wahrnehmen könne.

Mit freundlichen Grüssen
Für das Netzwerk der ehemaligen SSAB: Hanna Muralt Müller 29.4.2025

Falls Sie diese E-Mail nicht mehr erhalten möchten, melden Sie sich bitte bei mir: info@muralt-mueller.ch.