Editorial

https://doi.org/10.1515/abitech-2023-0013

Clickworker auf der ganzen Welt erstellen für Firmen Datensets mit Texten, Audio- und Bildaufnahmen, die diese als Trainingsdaten für KI-Algorithmen verkaufen. Für die meisten von uns ist es vermutlich unvorstellbar, dass man sein eigenes Foto mit verschiedenen Gesichtsausdrücken aufnimmt und verkauft. Im Gegenteil, viele Menschen versuchen, möglichst keine hochwertigen Fotos von sich selbst ins öffentliche Internet gelangen zu lassen. Beim Austesten von ChatGPT konnten viele von uns schon feststellen, dass die Software überhaupt in der Lage ist zu erkennen, wenn man sie nach einer Person befragt. Häufig kennt sie diese Person aber wegen fehlendem Trainingsmaterial nicht und ist auch darauf programmiert, diesen Umstand zu benennen. Das ist insofern überraschend, als auf eigentlich jede andere Frage eine Antwort gegeben wird, wenn auch nicht immer eine für die Fragenden nachvollziehbare.

Dass die KI so zurückhaltend mit Aussagen über Entitäten ist, die sie als mögliche Personen identifiziert hat, während sie über stärker generische Begriffe oder breite Themenstellungen gern endlos schwafelt, ist zunächst einmal beruhigend. Im Sinne von Datenschutz und überprüfbarer Information erscheint es erfreulich, dass Aussagen über konkrete Personen nicht so ohne weiteres aus dem Trainingsmaterial geschlussfolgert werden wie bspw. allgemeinere Fragestellungen wie die Zukunft von Gedächtnisinstitutionen.

Diffusionsmodelle zur Generierung von Bildern nach Vorgabe sind im Unterschied zu ChatGPT, das erst vor kurzem die Lizenzkosten für die reguläre Nutzung bekanntgegeben hat, schon häufig im Einsatz und haben auch bereits kommerzielle Relevanz, wenn auch in scheinbar unproblematischen Feldern wie dem Webdesign oder anderen Gestaltungsbereichen. Sie fügen eindeutigen Bildern aus Trainingsmaterial schrittweise „Rauschen“ hinzu und können dieses aus den Bildern anschließend wieder entfernen. Um ein gänzlich neues Bild zu erzeugen, wird der Software reines Rauschen vorgegeben, in dem diese aufgrund des verwendeten Modells etwas „erkennen“, d. h. ein Bild produzieren kann, das einen Sinn ergibt. Durch eine Texteingabe erhält das Modell zusätzliche Informationen, welchen „Sinn“ das neue Bild ergeben soll.

Dass Diffusionsmodelle jedoch unter bestimmten Bedingungen auch ihr Trainingsmaterial wieder freigeben, was auf keinen Fall erwünscht ist, wurde unlängst von Forschenden nachgewiesen.[1] Beim betrachteten Modell von Stable Diffusion ist diese „Erinnerungsrate“ bezogen auf den untersuchten Datensatz allerdings nur sehr gering gewesen. Trotzdem bleibt die unerwünschte Reproduzierbarkeit eine Herausforderung für die Datenqualität in Trainingssätzen, die für die Forschung unverzichtbar sind. Es ist eine offene Frage, ob die mit entsprechend hohem Aufwand mögliche Reproduzierbarkeit wirklich gegen die Anwendung von Diffusionsmodellen etwa im Bereich sensibler Medizindaten spricht, oder ob die Schutzziele auch auf anderem Wege sichergestellt werden können.

Konstanze Söllner

1 https://arstechnica.com/information-technology/2023/02/researchers-extract-training-images-from-stable-diffusion-but-its-difficult/. Zuletzt geprüft am 07.03.2023.