Laurenz Strassemeyer
Schriftleitung
Datenschutz-Berater

Datenschutz im Hanse-Stil: (K)ein Personenbezug in LLMs

Sehr geehrte Leserinnen und Leser,

Mitte Juli sorgte der Hamburgische Beauftragte für Datenschutz und Informationsfreiheit für Aufsehen: Er veröffentlichte das „Diskussionspapier: Large Language Models und personenbezogene Daten" (https://ogy.de/HBfDI-LLM-PB). Das Paper wird seither nicht nur in Deutschland, sondern in der gesamten europäischen Datenschutz-Community heiß diskutiert. Die Hamburger Behörde schuf die Grundlage dafür einmal mehr, indem sie ihr Paper direkt in Deutsch und Englisch bereitstellte. Das inkludiert von Beginn an auch die internationale Fachwelt und wird sowohl dem Geltungsbereich der DSGVO gerecht als auch der Wichtigkeit des Themas.

Schon die Entstehung des Papers wirkt von außen so, als handle es sich um ein Dokument, das ernsthaft den Diskurs anregen will. Vertreter der Hamburger Behörde waren in den letzten Monaten auf zahlreichen Konferenzen präsent, teilten ihre Thesen mit und stellten sich der öffentlichen Diskussion. Die Grundthese war zwar spätestens seit Anfang Mai 2024 bekannt, als Thomas Fuchs sie erstmals öffentlich vorstellte. Aber die Behörde trat so noch in der Entstehung in einen kritischen Austausch mit Akteuren aus Wissenschaft, Unternehmenspraxis und dem öffentlichen Sektor. Ich finde dieses Vorgehen mutig und anerkennenswert. Genau so sollte eine rechtliche Debatte zu grundlegenden Themen geführt werden: inklusiv und unter Berücksichtigung verschiedener Perspektiven.

Die im Diskussionspapier auf elf Seiten begründeten Kernthesen sorgen nun aber erneut für Aufsehen. Die Behörde kommt dabei insbesondere zu folgenden Schlüssen:

- Die bloße Speicherung eines LLMs stellt keine Verarbeitung dar, weil im LLM selbst keine personenbezogenen Daten gespeichert sind.
- In der Folge können sich Betroffenenrechte nicht auf das LLM selbst beziehen, sondern nur auf die in einem KI-System vor- und nachgelagerte Verarbeitungen. Also etwa das Erheben und Aufbereiten des Trainingsdatensatzes sowie die Ein- und Ausgabedaten oder deren Nutzung.

Die auf die Veröffentlichung dieser Thesen folgende Diskussion wurde auf LinkedIn teils sachlich, teils recht polemisch geführt. Unabhängig von der rechtlichen Position wird Letzteres dem Paper der Behörde nicht gerecht. Auch wenn ich nicht meine Rolle darin sehe, die Behörde zu verteidigen – das kann sie selbst gut genug – ist das Diskussionspapier unabhängig vom Ergebnis als beachtliche Grundlage für eine faktenbasierte Diskussion zu loben.

Über die Jahre habe ich viel zu Künstlicher Intelligenz und LLMs gelesen. Oft setzt die Literatur umfangreiches Grundlagenwissen voraus, um rechtliche Schlussfolgerungen zu bilden. Manchmal wird das Thema auch für die juristische Fragestellung überkomplex dargestellt oder es werden Themen vermischt, die nicht zusammengehören. Dies erschwert oft fachlich fundierte Diskussionen. Der Hamburger Behörde ist es jedoch gelungen, die technologischen Hintergründe stark zu abstrahieren und sehr eingängig zu erklären. Das Paper ist deshalb sofort auch Leuten ohne technisches Vorwissen als Einführung zu empfehlen. Es bietet einen verständlichen Einblick in die Funktionsweise von LLMs (nicht KI-Systemen im Gesamten, was ganz bewusst gewählt ist). Natürlich gibt es auch kleine Schwächen. Aber das ist unumgänglich. Während die Tokenisierung anschaulich dargestellt ist, sind der weitere Trainingsvorgang und die Repräsentation des Wissens im LLM vielleicht für meinen Geschmack etwas zu sehr runtergebrochen worden. Aber ganz ehrlich, für den Diskurs in der juristischen Fachwelt dürfte es vielleicht genau richtig sein.

Und zu guter Letzt halte ich die Thesen inhaltlich für wichtig (noch dazu für überwiegend zutreffend). Die Hanseaten zeigen, dass auch Aufsichtsbehörden nicht zwangsläufig pauschale Datenschutz-Extrempositionen einnehmen müssen. Stattdessen wird das LLM als das „Gehirn" der Sprachverarbeitung im KI-System isoliert betrachtet, in seine Bestandteile zerlegt und die bisherige Rechtsprechung des EuGH zum Personenbezug kritisch angewendet. Dabei ist es richtig und wichtig, wenn die Behörde betont, dass die Rechtsprechung beim Personenbezug bisher auf eindeutig zuordenbare „Identifier" wie IP-Adresse, TC-String oder VIN abstellt. Solche Identifier fehlen im LLM. Daneben gibt es auch die eine oder andere witzige Art der Behörde, mit Fehlannahmen aufzuräumen. Beispielsweise wird in einer Fußnote, hanseatisch dezent und zugleich sehr klar, die unpassende Analogie zwischen LLM und Verschlüsselung ausgeräumt.

Aber ich für meinen Teil möchte an dieser Stelle inhaltlich gar nicht in die Diskussion einsteigen, sondern Sie vielmehr dazu anregen, das Paper zu lesen. Unabhängig davon, ob Sie den Thesen zustimmen oder die Gegenposition vertreten: Die Hamburger Behörde hat allein durch die Darstellung ihrer Schlussfolgerungen Wichtiges geleistet. Jetzt kann die Gegenseite ihre Argumente mit fundierten Analysen untermauern. In solchen grundlegenden Themenbereichen sollten wir uns glücklich schätzen, wenn verschiedene Positionen – insbesondere auch von Aufsichtsbehörden – veröffentlicht werden.

Ich wünsche Ihnen mit diesem Heft eine ähnlich spannende Lektüre und möchte noch kurz darauf hinweisen: Die DSK 2024 steht bald an. Kommen Sie unbedingt vorbei für Diskussionen hierzu und zu anderen Themen (https://ogy.de/DSK-2024).

Ihr

Laurenz Strassemeyer