JURISTISCHE SPRACHMODELLE ZWISCHEN TRANSPARENZ UND DATENSICHERHEIT

Rolf H. Weber

Rolf H. Weber, Professor, Rechtsanwalt, Universität Zürich, Rechtswissenschaftliche Fakultät
Rämistrasse 74/38, 8001 Zürich, CH
rolf.weber@ius.uzh.ch; https://www.ius.uzh.ch/de/staff/professorships/alphabetical/weberr.html

Keywords: ***Datensicherheit, Datenqualität, Formen von Risiken/Angriffen, Nachvollziehbarkeit, Transparenz***

Abstract: *Juristische Sprachmodelle kommen vermehrt und in umfangreicherem Ausmass bei der Erbringung von Rechtsdienstleistungen zum Einsatz. Oft erfahren die Transparenz und die Datensicherheit aber nur eine ungenügende Beachtung. Diese beiden Ziele sind zudem nicht deckungsgleich. Der Beitrag analysiert die eintretenden Spannungsfelder und schlägt Vorkehren zur Risikominimierung beim Einsatz von rechtlichen Sprachmodellen vor, etwa zur Gewährleistung der Datenqualität oder zum Schutz vor Angriffen (z.B. Datenschutzverletzungen)*

1. Einleitung

Die juristischen Sprachmodelle bzw. allgemein die Large Language Models (LLM) haben jüngst erheblich an Bedeutung gewonnen. Solche Modelle sind in der Lage, die natürliche Sprache durch Künstliche Intelligenz (KI) zu verarbeiten, und gestützt darauf sachbezogene Texte zu entwickeln.[1] Dank des Trainings mit grossen Datenmengen lassen sich wiederkehrende Muster erkennen, z.B. im Falle oft aufeinanderfolgender Worte in einem bestimmten Kontext.[2]

Die technischen Innovationen bzw. gar «Umwälzungen» führen auch zu neuen rechtlichen Herausforderungen. Abgesehen vom zentralen Thema des Urheberrechtsschutzes, das einer gesonderten Behandlung bedürfte,[3] wird nachfolgend aus der Vielzahl der rechtlichen Fragestellungen insbesondere auf die Transparenz und die Datensicherheit sowie deren Spannungsverhältnis eingegangen. Bewusst sind die Überlegungen allgemein formuliert und nicht auf die heute am meisten eingesetzten Programme (z.B. ChatGPT,[4] Google Bard, usw.) konzentriert.

2. Wesen, Charakteristiken und Risiken juristischer Sprachmodelle

2.1. Wesen und Charakteristiken

Juristische Sprachmodelle zeichnen sich dadurch aus, dass sie die natürliche Sprache in schriftlicher Form auf der Basis von maschinellem Lernen verarbeiten und auch als Text präsentieren.[5] Weil juristische Arbeit oft sehr sprachorientiert ist, sind rechtliche LLM ein geeignetes Unterstützungsinstrument.[6] Entsprechende

1 Gruetzemacher, 2022, passim.
2 Vokinger et al., 2022, 2.
3 Dazu grundlegend Straub, 2023, Rz 2 ff.
4 Eine umfangreiche Sammlung von Materialien (inkl. Podcasts) zur Funktionsweise von und zu rechtlichen Fragestellungen mit Bezug auf ChatGPT lassen sich unter https://www2.weblaw.ch/home/academy/chatgpt-archiv (accessed on 17 October 2023) finden.
5 Bundesamt für Sicherheit in der Informationstechnik, 2023, 5; Barenkamp, 2023, 1.
6 Selman et al., 2023, 289 ff.

Sprachmodelle lassen sich vielfältig in allen Bereichen einsetzen, in denen eine (teil-)automatisierte Textverarbeitung und/oder Textproduktion stattfindet. Beispiele sind:[7]

- *Textgenerierung*: Verfassen von Entwürfen für Dokumente, Verfassen von Texten in einem bestimmten Schreibstil, Werkzeuge zur Textfortführung oder -vervollständigung;
- *Textbearbeitung*: Rechtschreib- und Grammatikprüfung;
- *Textverarbeitung*: Wort- und Textklassifikation, Markierung von Begriffen im Text, Textzusammenfassung, Frage-Antwort-Systeme, Übersetzung;
- *Programmcode*: Werkzeuge zur Unterstützung beim Programmieren, Erzeugen von Programmcode, Umprogrammierung und Übersetzung eines Programmes in andere Programmiersprachen.

Die konkrete Ausgestaltung der einzelnen juristischen Sprachmodelle hängt von den beabsichtigten Einsatzformen und den angestrebten Innovationen ab. Der Einsatz der Sprachmodelle im Rechtswesen vermag insbesondere die Erledigung von standardisierten Aufgaben zu erleichtern.[8] Denkbar sind vollautomatisierte Prozesse, die direkt zur Rechtsprognose bzw. zur Entscheidung führen, oder teilautomatisierte Prozesse, die nur Entscheidungsgrundlagen bzw. Empfehlungen generieren.

2.2. Risiken

Juristische Sprachmodelle sind nicht zu unterschätzenden Risiken ausgesetzt. Large Language Models generieren in Regel einen sprachlich fehlerfreien und meist auch inhaltlich überzeugenden Text. Diese Tatsache darf aber nicht den Eindruck eines menschenähnlichen Leistungsvermögens erwecken und ein unbegrenztes Vertrauen in die Aussagen begründen. Als allgemeine Risiken sind, ungeachtet der durch mangelnde Transparenz oder ungenügende Datensicherheit hervorgerufenen Herausforderungen, folgende Problembereiche zu beachten:[9]

- *Fehlende Faktizität und Reproduzierbarkeit*: Die Generierung von Text beruht auf stochastischen Korrelationen, d.h. das «Wissen» wird aus (bereits gesehenen) Texten abgeleitet. Bezüge zur realen Welt existieren für das Modell nicht; deshalb kann es bei für Menschen verständlichen Sachverhalten zu inkorrekten Aussagen kommen oder ggf. fallen Aussagen gestützt auf dieselbe Eingabe aufgrund des wahrscheinlichkeitsbasierten Ansatzes unterschiedlich aus.
- *Fehlende Aktualität*: Fehlt ein Zugriff auf Live-Internetdaten, d.h. auf Informationen über aktuelle Ereignisse, produziert das Sprachmodell aufgrund von in der Vergangenheit geprägten Texten; die Ergebnisse sind diesfalls nicht aktuell.
- *Fehlerhafte Reaktion auf spezifische Eingaben*: Fehler vermögen aufzutreten, wenn die Sprachmodelle gewisse Eingaben verarbeiten, die so stark von den Texten in den Trainingsdaten abweichen, dass sie sich nicht mehr korrekt als Text bzw. Word-Dokument verarbeiten lassen (im Schrifttum zum Teil «Hallucination»[10] genannt). Abweichungen können zufällig entstehen oder absichtlich als Täuschungsakt integriert werden.
- *Anfälligkeit für «versteckte» Eingaben mit manipulativer Absicht*: Gefahrengeneigt sind Modelle, die als Input ungeprüfte Dokumente Dritter enthalten, oder Situationen, in denen Angreifende für Nutzende unbemerkt gewisse Eingaben in das Sprachmodell einzubringen vermögen, insbesondere wenn der Betrieb den Zugriff auf Live-Daten aus dem Internet erlaubt; ein solcher Angriff kann z.B. ein Chat-Tool betreffen, das eine Person beim Surfen im Internet unterstützt.

7 Bundesamt für Sicherheit in der Informationstechnik, 2023, 7.
8 Vgl. auch Barenkamp, 2023, 2; zu den Besonderheiten der Automatisierung von Rechtsprechung vgl. Gless, 2023, 434 ff.
9 Bundesamt für Sicherheit in der Informationstechnik, 2023, 10 f.
10 Für Einzelheiten vgl. Zhang et al., 2023, 2 ff.

– *Vertraulichkeit der eingegebenen Daten*: Bei der Nutzung von externen Schnittstellen fliessen alle Eingaben, die im Sprachmodell verwendet werden, zunächst an den Betreiber des Modells ab. Die Nutzung externer Schnittstellen sollte deshalb bei der Verarbeitung von sensiblen und vertraulichen Informationen mit besonderen Schutzvorkehren verknüpft sein.

Eine grosse Herausforderung stellt die oft fehlende Datenhoheit bei der Verwendung von juristischen Sprachmodellen dar; überdies ist eine meist grosse Abhängigkeit vom Hersteller und Betreiber des Modells (insbesondere in technischer Hinsicht) problematisch.

3. Transparenz und Nachvollziehbarkeit

3.1. Transparenz

Ein grundlegendes Problem ist die mangelnde Transparenz in der Funktionsweise von juristischen Sprachmodellen. Intransparenz vermag hinsichtlich des Einsatzes von Algorithmen selbst zu bestehen, aber auch mit Blick auf die Verwendung von Daten betroffener Personen. Akzentuiert wird das Risiko der Intransparenz bei selbstlernenden Algorithmen, bei denen je nach angewendeter Methode des Machine Learning angesichts des gegenwärtigen Standes der Technik u.U. selbst der Systembetreiber den Lösungsweg nicht mehr nachvollziehen kann.[11]

Die primär- und verfassungsrechtlichen Regeln und Vorgaben zum Zugang zu Dokumenten können für mehr Transparenz sorgen, doch nicht einzelne Vorgänge erklären. Im Falle der Verwendung von Personendaten erlangen die Datenschutzbestimmungen Relevanz. «Automatisierte Entscheidungen» sind, wenn das Erfordernis der Vollautomatisierung erfüllt ist, grundsätzlich verboten (Art. 22 DSGVO);[12] ob bzw. inwieweit tatsächlich Personendaten verarbeitet werden, hängt von den gegebenen Umständen ab.

Konkrete Vorgaben sind im EU-Sekundärrecht vorgesehen; so verlangt Art. 5 der Verordnung (EU) 2019/1150 von den Anbietern, die wichtigsten ein Ranking bestimmenden Hauptparameter, die Gründe für deren relative Gewichtung gegenüber anderen Parametern und die potentielle Beeinflussung des Ranking durch Entgeltsleistungen transparent zu machen.[13]

Selbst wenn Transparenzerfordernisse von Bedeutung sind, um die betroffenen Personen von den mit Sprachmodellen verbundenen Risiken zu schützen, ist deren Wirkung nicht zu überschätzen. Empirische Untersuchungen zu den vorhandenen daten- und verbraucherschutzrechtlichen Informationspflichten lassen daran zweifeln, ob Transparenzerfordernisse stets einen Mehrwert für die betroffenen Personen generieren und ihren angedachten Zweck tatsächlich erreichen.[14] Die erhaltenen Informationen sind oft nicht ausreichend verständlich und überfordern angesichts des Überflusses an Informationen die Empfänger (sog. «Information Overload»).[15] Diese allgemeine Problematik verstärkt sich angesichts der hohen Komplexität selbstlernender System- und Sprachmodelle, nicht zuletzt auch angesichts der Tatsache, dass eine eigentliche Erklärung der Verarbeitungsschritte bei Machine Learning-Verfahren bisweilen technisch nur schwer zu bewerkstelligen ist.[16]

Schliesslich ist nicht zu übersehen, dass selbst eine ausreichende Transparenz bei automatisierten Entscheidungen die menschliche Nachkontrolle nicht vollständig zu ersetzen vermag.

---

[11] Weber/Henseler, 2020, 30; Barenkamp, 2023, 3.

[12] In der Schweiz geht Art. 19 DSG weniger weit als Art. 22 DSGVO (vgl. dazu Weber/Henseler, 2020, 35).

[13] Verordnung (EU) 2019/1150 zur Förderung von Fairness und Transparenz für gewerbliche Nutzer von Online-Vermittlungsdiensten, ABl 2019 L 186/57 vom 11. Juli 2019.

[14] Weber/Henseler, 2020, 38.

[15] Weber, 2023, 67 ff. m.w. H.

[16] Weber/Henseler, 2020, 39.

3.2. Nachvollziehbarkeit

Die Arbeitsweise von juristischen Sprachmodellen muss erklärbar und interpretierbar sein;[17] die von der Datenbearbeitung betroffenen Personen sind in die Lage zu versetzen, die algorithmischen Vorgänge nachvollziehen zu können. Von grosser Bedeutung ist die Erklärbarkeit insbesondere in den Bereichen der Gesundheit, der Arbeitswelt, der Finanzen und der Justiz.

Erklärbarkeit bedingt, abhängig vom Wissenshorizont des Betroffenen und der Bedeutung der Vorgänge, ausreichende Verständlichkeit; die zugrundeliegende Logik muss begreiflich sein und die notwendigen Informationen enthalten, um einen ausreichenden Grad an Nachvollziehbarkeit zu erreichen.[18] Dieses Ziel ist in einer hoch technologisierten Umgebung nur schwer zu erreichen, doch sind die Rahmenbedingungen durch die entsprechend konkrete Ausgestaltung der Sprachmodelle zu schaffen, damit die Betroffenen mit vernünftigem Aufwand die Vorgänge zu verstehen in der Lage sind.[19]

Erklärbarkeit und Nachvollziehbarkeit haben einen Bezug zur Schaffung von Transparenz innerhalb des Kreises der Betroffenen von eingesetzten LLM; die Vertraulichkeit bzw. der Geheimnisschutz betrifft «nur» die Drittbeziehungen.

4. Datensicherheit

4.1. Chancen und Nutzen für die IT-Sicherheit

Juristische Sprachmodelle können zur Verbesserung der IT-Sicherheit beitragen. Als Chancen der Nutzung von juristischen Sprachmodellen lassen sich nennen:[20]

- *Unterstützung bei der Detektion unerwünschter Inhalte*: Sind juristische Sprachmodelle für Textklassifikationsaufgaben geeignet, ergeben sich auch Anwendungsmöglichkeiten im Bereich der Detektion von Spam-Mails oder unerwünschten Inhalten.
- *Unterstützung bei der Textverarbeitung*: Sprachmodelle im Bereich der Textanalyse und -strukturierung, sind geeignet, grössere Mengen von Text zu verarbeiten und bei der Berichterstellung zu Sicherheitsvorfällen unterstützend mitzuwirken.
- *Unterstützung bei der Erstellung und Analyse von Programmcode*:[21] Sprachmodelle lassen sich dazu einsetzen, die vorhandenen Programme auf Sicherheitslücken zu untersuchen sowie Wege zur Ausnutzung dieser Schwächen für Angriffe oder zur Codeverbesserung vorzuschlagen.
- *Unterstützung bei der Analyse von Datenverkehr*:[22] Nach zusätzlichen Trainings vermögen Sprachmodelle auch bei Aufgaben zu unterstützen, die nicht natürlich-sprachigem Text im engeren Sinne entsprechen, aber doch die Detektion von bösartigem Netzwerk-Verkehr erleichtern.

Chancen und Nutzen hängen aber von der Benutzerfreundlichkeit der Anwendungen der verschiedenen Modelle ab; regelbasierte Systeme (z.B. Experten-Entscheidungsbäume) können dabei hilfreich sein.[23] Die Schnittstellen der einzelnen Modelle müssen so ausgestaltet sein, dass sich Brücken zwischen den einzelnen Modellen bilden lassen.[24]

---

17 Danilevsky et al., 2022, passim.
18 Weber, 2022, B 12.
19 Bundesamt für Sicherheit in der Informationstechnik, 2023, 8.
20 Bundesamt für Sicherheit in der Informationstechnik, 2023, 9; vgl. auch Zellers et al., 2020, passim.
21 Bubeck et al., 2023, 1 ff.
22 Almodovar et al., 2022, 1 ff.
23 Vgl. Gless, 2023, 440 f.
24 Das Kriterium der Nachvollziehbarkeit ist in vielen Bereichen, die von neuen Technologien geprägt sind, zu einem Problem geworden. Wer versteht die Code-basierte Ausgestaltung eines Smart Contract im Zeitpunkt seines Abschlusses? Wer vermag die mit Künstlicher Intelligenz durchgeführten Datenanalysen nachzuvollziehen? Diesem Problem wird die Rechtswissenschaft vermehrt Beachtung schenken müssen.

4.2. Risiken und Missbrauch bei der Nutzung

Wie erwähnt sind juristische Sprachmodelle verschiedenen Risiken und Missbrauchsgefahren ausgesetzt. Abgesehen von den bereits erläuterten Risken, die von juristischen Sprachmodellen verursacht werden, ist in der Realität nicht auszuschliessen, dass die Produktion von Text für böswillige Zwecke missbraucht wird. Als Beispiele lassen sich nennen:[25]

– *Social Engineering*: Cyber-Angriffe, bei denen Kriminelle versuchen, ihre Opfer zu veranlassen, Daten preiszugeben, Schutzmassnahmen zu umgehen oder selbstständig einen Schadcode zu installieren, fallen unter den Begriff des Social Engeneering.
– *Erhöhte Risiken wegen Malware*: Neuere Sprachmodelle besitzen immer ausgereiftere Code-Generierungsfähigkeiten, die es den Kriminellen mit geringen technischen Fähigkeiten ermöglichen, einen Schadcode ohne viel Hintergrundwissen zu erzeugen; polymorphe Schadsoftware gibt es zwar bereits seit einiger Zeit, doch ist heute die Generierung von Malware «leichter».
– *Hoax (Falschmeldung)*: Angesichts der grossen Zahl an Daten, die sich nicht vollständig überprüfen lassen, verbleiben regelmässig Texte mit fragwürdigem Inhalt (z.B. Desinformation, Propaganda) in der Trainingsmenge; sie tragen zu einer unerwünschten Struktur des Modell bei, die eine Neigung zu potentiell kritischen Inhalten zeigt.[26]

Um die Sicherheit juristischer Sprachmodelle zu gewährleisten, sind Massnahmen gegen mögliche Risiko- oder Missbrauchsfälle rechtzeitig einzurichten.[27] Gegenmassnahmen können sowohl technischer als auch organisatorischer Art sein; Ziel ist die Absicherung der Authentizität von Texten und Nachrichten sowie der Nachweis, dass bestimmte Texte oder Informationen tatsächlich von einer bestimmten Person, Personengruppe oder Institution stammen.

5. Entwicklung transparenter und sicherer Sprachmodelle im Rechtsbereich

Der immer umfangreichere Einsatz von juristischen Sprachmodellen zwingt zu Anstrengungen, den beiden Anliegen der Transparenz und der Datensicherheit eine grössere Beachtung zu schenken. Nur wenn die Betroffenen davon ausgehen können, dass die algorithmischen Abläufe und Vorgehensweisen nachvollziehbar sind und die eingegebenen Daten ein hohes Mass an Datensicherheit aufweisen, ist es vertretbar, sich im beruflichen Alltag auf Large Language Models abzustützen.

5.1. Überbrückung unterschiedlicher Interessenlagen

(i) *Transparenz* und Nachvollziehbarkeit bedeuten, dass klare Verfahrensgrundsätze mit Blick auf den Einsatz von rechtlichen Sprachmodellen implementiert werden. Transparent ist ein Modell, wenn es Klarheit schafft.[28] Zu umschreiben sind die typischen Tätigkeitsbereiche von LLM im Rechtswesen (z.B. Verfassen von Dokumenten, Beratungstätigkeit, Literaturrecherche, Administration);[29] innerhalb der einzelnen Tätigkeitsgebiete lassen sich Ausmass und Art des Einsatzes von Sprachmodellen festlegen. Über die anwaltlichen Qualitätskriterien an die Leistungserbringung hinaus bedürfen insbesondere die Vorgaben des Datenschutzes und des Berufsgeheimnisses einer besonderen Beachtung.[30]

Im Rechtsbereich stellen sich spezifische Herausforderungen mit Blick auf die Validierung von Personendaten (insbes. Datenrichtigkeit) sowie auf die Rückverfolgung der Gewichtung von Parametern innerhalb des Modells; zudem sind Methoden zu nutzen, die eine Angabe von Rechtsquellen ermöglichen. Ausgeschlossen

[25] Bundesamt für Sicherheit in der Informationstechnik, 2023, 12 f.
[26] Vgl. auch Weidinger et al., 2022, 1 ff.
[27] Bundesamt für Sicherheit in der Informationstechnik, 2023, 13 f.
[28] Grundlegend zu den Transparenzanforderungen Weber, 2009, 121 ff.
[29] Lips et al., 2023, 323 ff.; vgl. auch Selman et al. 2023, 291.
[30] Eingehender dazu Seiler/Griesinger, 2023, Rz 7 ff.

sein sollte eine Nutzung von rechtlichen Sprachmodellen, welche auf die Prognose von Informationen abzielen, die sich nicht auf anderem Weg verifizieren lassen.

Der Transparenz zu dienen vermögen auch neue Verfahren zur Authentisierung durch technische Prozesse,[31] welche die Urheberschaft einer Informationsübermittlung kryptografisch nachweisen. Wenn solche technische Prozesse in nachvollziehbarer Weise die Identifikation der personellen Quellen ermöglichen, lässt sich verhindern, dass Akteure, die ggf. einen negativen Einfluss auf die Dienstleistungserbringung ausüben, in der Intransparenz «verschwinden».

(ii) Aus der Perspektive der *Datensicherheit* sind zwei Schutzstrategien auf der Modellebene denkbar: Einerseits lassen sich die Nutzungsmöglichkeiten allgemein einschränken, andererseits können Massnahmen zur Unterbindung potentiell schädlicher Auswirkungen auf die Sprachmodelle in Betracht gezogen werden. Zudem sind Massnahmen zur Verringerung des Angriffsrisikos durch Verbesserung der Sensibilisierung und Aufklärung der Nutzenden zu treffen.

Zwischenzeitlich bestehen auch verschiedene komplementäre Ansätze zur Detektion automatisch generierter Texte; durch die Detektion erhalten Nutzende die Fähigkeit, Texte als maschinengeschrieben zu erkennen und somit ggf. ihre Authentizität und die Richtigkeit der enthaltenen Daten anzuzweifeln. Automatisierte Detektionswerkzeuge mit Bezug auf maschinengenerierte Texte nutzen in der Regel statische Eigenschaften der Texte aus und verwenden Parameter eines Modells, um einen Score zu berechnen, der als Indiz für maschinengeneriete Texte dient.

5.2. Konkrete Vorgaben für transparente und sichere Sprachmodelle

Mit Blick auf die Entwicklung transparenter und sicherer juristischer Sprachmodelle können verschiedene Risikovermeidungs- und Risikominderungsmassnahmen relevant sein:[32]

Datenqualität bei der Auswahl von Trainingsdaten: Die konkrete und transparente Auswahl der Trainingsdaten ist wichtig, um ein hohes Qualitätsniveau beim juristischen Sprachmodell zu erreichen. Die verwendeten Trainingsdaten sind Grundlage für die späteren Textgenerierungs-Modelle.[33] Zur Qualität gehört auch die Beachtung der rechtlichen Vorgaben; im Vordergrund steht dabei die Berücksichtigung der urheberrechtlichen Rahmenbedingungen.[34]

Zur Datenqualität gehört weiter die Vermeidung von (versteckten) Diskriminierungen, die sich durch die Verwendung der Algorithmen perpetuieren könnten.[35] Zwar lassen sich Diskriminierungen in der Praxis nie ganz ausschliessen, z.B. wenn geographische oder personenbezogene Merkmale verwendet werden; zumindest ist aber zu versuchen, entsprechende negative Effekte zu minimieren.

Privacy Attacks: Angesichts der grossen Menge an Trainingsdaten ist es oft schwierig, durch transparente technische Massnahmen sicherzustellen, dass die Personendaten lediglich für die datenschutzrechtlich unbedenklichen eingeschränkten Zwecke verwendet werden.[36] Möglichkeiten zur Verminderung der Anfälligkeit für Privacy Attacks sind beispielsweise[37] die manuelle Auswahl oder automatische Filterung bzw. Anonymisierung von Daten, die Entfernung von Dopplungen aus den Trainingsdaten zwecks Herabsetzung der Wahrscheinlichkeit einer möglichen Rekonstruktion,[38] die Anwendung von Mechanismen, welche die sog. Differential Privacy garantieren,[39] die Einschränkung des Zugriffs auf das Modell (enge Zahl an Nutzenden) sowie die Vornahme von zusätzlichen Trainings bzw. eines besonderen Training für sensible Daten.[40]

---

31 Für eine Problemanalyse zum Finanzmarktbereich vgl. Boškić/Hepp, 2023, 214 ff.
32 Bundesamt für Sicherheit in der Informationstechnik, 2023, 15 ff.
33 Vgl. auch Rosenthal, 2023, Rz 26 ff.
34 Straub, 2023, Rz 2 ff; aus US-amerikanischer Sicht vgl. Leujeune, 2023, 142 ff.
35 Weber/Henseler, 2020, 31 f.
36 Zur Herstellung eines datenschutzrechtlich sicheren Modells vgl. Rosenthal, 2023, Rz 10 ff.
37 Für einen allgemeinen Überblick vgl. Bundesamt für Sicherheit in der Informationstechnik, 2023, 16.
38 Carlini et al., 2021, 1 ff,
39 Desfontaines/Pejó, 2020, passim.
40 Aus informationstechnologischer Sicht dazu Staab et al., 2023, passim.

Adversarial Attacks: Angriffe können darauf abzielen, Texte absichtlich nur leicht zu verändern bzw. intransparent zu gestalten, um die Wahrnehmung der Veränderung unwahrscheinlich zu machen. Als besonders anfällig dafür erweisen sich Klassifikationen. Gegenmassnahmen sind das Training des Modells mit realen oder möglichst realistischen Daten, die Vorverarbeitungen des möglicherweise adversarialen Textes,[41] automatische Rechtschreibkorrekturen, Modellverbesserungen durch Trainings mit manipulierten/veränderten Texten, die Einbindung einer externen Wissensbasis oder das Clustering von Word-Embeddings.

Poisoning Attacks: Sofern die Daten aus öffentlichen Quellen stammen, sind sie einer Vielzahl von Angriffsmöglichkeiten ausgesetzt; Institutionen weisen oft einen offenen Zugang und einen sicherheitstechnisch nicht immer ausreichenden Schutz auf.[42] Möglichkeiten zur Verminderung der Anfälligkeit für Poisening-Angriffe sind die Verwendung vertrauenswürdiger Quellen als Trainingsdaten, der Einsatz von geschultem und vertrauenswürdigem Personal, die intensive Analyse von Bewertungen (Voreintritt von Rückwirkungen auf das Modell) und die Beschränkung der Auswirkungen des Einsatzes auf ein kontrollierbares Feld. Alle getroffenen Massnahmen sollten weiter zu einem hohen Grad an Transparenz führen;[43] damit lässt sich eine Annäherung an die angesprochene Kohärenz zwischen Transparenz und Datensicherheit erreichen.

6. Ausblick

Gesamthaft betrachtet steht somit ein beachtliches Arsenal an Vorgaben zur Verfügung, das sich einsetzen lässt, um geeignete Rahmenbedingungen für transparente und sichere juristische Sprachmodelle im Rechtsbereich zu schaffen. Ausgangspunkt muss die transparente Festlegung der Voraussetzungen und Anwendungskriterien für die Nutzung von Sprachmodellen sein. Gestützt darauf hat eine Implementierung der Programme unter Beachtung hoher Datensicherheitsstandards zu erfolgen.

Der Transfer von Wissen auf neuartige algorithmische Anwendungen beim Einsatz von LLM bringt regelmässig gewisse Risiken mit sich. Deren Bewältigung ist indessen möglich, wenn die sachgerechten Schutzvorkehren getroffen werden, d.h. ein transparentes Vorgehen ist mit ausreichenden Datensicherheitsgarantien vereinbar. Unausweichlich ist aber im Kontext der juristischen Sprachmodelle eine verstärkte interdisziplinäre Zusammenarbeit; nur wenn Recht und Informationstechnologie in der Theorie und in der Praxis kooperieren, ist die gewünschte Harmonie von Transparenz und Datensicherheit zu erreichen.

7. Literatur

Almodovar, Crispin/Sabrina, Fariza/Karimi, Sarvnaz/Azad, Salahuddin, Can Language Models Help in System Security? Investigating Log Anomaly Detection using BERT, Proceedings of the 20th Annal Workshop of the Australasian Language Technology Association, Adelaide, December 2022, https://aclanthology.org/2022.alta-1.19/ (accessed on 17 October 2023).

Barenkamp, Marco, Einflussnahme grosser Sprachmodelle auf die moderne Arbeitswelt, Informatik Spektrum, 22. August 2023, https://doi.org/10.1007/s00287-023-01546-8 (accessed on 17 October 2023).

Boškić, Magdalena/Hepp, Sebastian, zkKYC in Decentralized Finance (DeFi), GesKR 2023, 209–225.

Bubeck, Sébastien/Chandrasekaran, Varun/Eldan, Ronen/Gehrke, Johannes/Horvitz, Eric/Kamar, Ece/Lee, Peter/Lee, Yin Tat/Li, Yuanzhi/Lundberg, Scott/Nori, Harsha/Palangi, Hamid/Ribeiro, Marco Tulio/Zhan, Yi, Sparks of Artificial General Intelligence: Early Experiments with GPT-4, April 2023, https://arxiv.org/abs/2303.12712 (accessed on 17 October 2023).

Bundesamt für Sicherheit in der Informationstechnik, Grosse KI-Sprachmodelle, Chancen und Risiken für Industrie und Behörden, Bonn 2021.

41 Wang et al., 2019, passim.

42 Bundesamt für Sicherheit in der Informationstechnik, 2023, 18.

43 Vgl. dazu auch Rosenthal, 2023, Rz 55 ff.

Carlini, Nicholas/Tramèr, Florian/Wallace, Eric/Jagielski, Matthew/Herbert-Voss, Ariel/Lee, Katherine/Roberts, Adam/Brown, Tom/Song, Dawn/Erlingsson, Úlfar/Oprea, Alina/Raffel, Colin, Extracting Training Data from Large Language Models, 30th Usenix Security Symposium, 2021, https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting (accessed on 17 October 2023).

Danilevsky, Marina/Qian, Kun/Aharonov, Ranit/Katsis, Yannis/Kawas, Ban/ Sen, Prithviraj, A Survey of the State of Explainable AI for Natural Language Processing, December 2020, https://aclanthology.org/2020.aacl-main.46/ (accessed on 17 October 2023).

Desfontaines, Damien/Pejó, Balász, Differential Privacies, Proceedings on Privacy Enhancing Technologies Symposium, December 2020, https://arxiv.org/abs/1906.01337 (accessed on 17 October 2023).

Gless, Sabine, Künstliche Intelligenz in der Gerichtsbarkeit, Zeitschrift für Schweizerisches Recht 2023 I 429–462.

Gruetzemacher, Ross, The Power of Natural Language Processing, Harvard Business Review, April 2022, https://hbr.org/2022/04/the-power-of-natural-language-processing (accessed on 17 October 2023).

Leujeune, Mathias, AI Systems and their Output under U.S. Copyright Law, CRi 5/2023, 141–148.

Lips, Markus/Şahin, Fatih/Rosenauer Philipp, Effizienzsteigerungsmöglichkeiten durch generative AI für Rechtsdienstleistungen, Anwaltsrevue 2023, 323–327.

Rosenthal, David, Datenschutz beim Einsatz generativer künstlicher Intelligenz, Weblaw Jusletter, 6. November 2023.

Seiler, Daniel W./Griesinger, Marcel, Spannungsfeld Künstliche Intelligenz (KI) und Datenschutzrecht, Weblaw Jusletter, 25. September 2023.

Selman, Sine/Burrichter, Annika/Hubli, Pascal, Die Verwendung von KI und ChatGPT im Anwaltsberuf, Anwaltsrevue 2023, 289–294.

Staab, Robin/Vero, Mark/Balunovic, Mislav/Vechev, Martin, Beyond Memorization: Violating Privacy via Interference with Large Language Models, October 2023, https://arxiv.org/pdf/2310.07298v1.pdf (accessed on 17 October 2023).

Straub, Wolfgang, Immaterialgüterrechtlicher Schutz mit KI geschaffener Werke und Erfindungen, Weblaw Jusletter, 7. August 2023.

Vokinger, Kerstin Noëlle/Schneider, David/Locher, Luca/Herrle, Cascal/Mühlematter, Urs Jakob, Analyse von Bundesgerichtsurteilen mit ChatGPT-3.5, Weblaw Jusletter, 7. August 2023.

Wang, Wenqi/Tang, Benxiao/Wang, Run/Wang, Lina/Ye, Aoshuang, A survey on Adversarial Attacks and Differences in Text, February 2019, https://www.researchgate.net/publication/331246520_A_survey_on_Adversarial_Attacks_and_Defenses_in_Text (accessed on 17 October 2023).

Weber, Rolf H., The Disclosure Dream – Towards a New Transparency Concept in EU Consumer Law, EuCML 2023, 67–70 (Weber, 2023).

Weber, Rolf H. Künstliche Intelligenz: Regulatorische Überlegungen zum «Wie» und «Was», EuZ 01/2022, B1–B18, (Weber, 2022).

Weber, Rolf H., Shaping Internet Governance: Regulatory Challenges, Zürich 2009 (Weber, 2009).

Weber, Rolf H./Henseler, Simon, Regulierung von Algorithmen in der EU und in der Schweiz, EuZ 2020, 28–42 (Weber, 2020)

Weidinger, Laura/Uesato, Jonathan/Rauh, Maribeth/Griffin, Conor/Huang, Po-Sen/Mellor, John/Glaese, Amelia/Cheng, Myra/Balle, Borja/Kasirzadeh, Atoosa/Biles, Courtney/Brown, Sasha/Kenton, Zac/Hawkins, Will/Stepleton, Tom/Birhane, Abeba/Hendricks, Lisa Anne/Rimell, Laura/Isaac, William/Haas, Julia/Legassick, Sean/Irving, Geoffrey/Gabriel, Iason, Taxonomy of Risks posed by Language Models, June 2022, https://doi.org/10.1145/3531146.3533088 (accessed on 17 October 2023).

Zhang, Yue/Li, Yafu/Cui, Leyang/Cai, Deng/Liu, Lemao/Fu, Tingchen/Huang, Xinting/Zhao, Enbo/Zhang, Yu/Chen, Yulong/Wang, Longyue/Luu, Anh Tuan/Bi, Wei/Shi, Freda/Shi, Shuming, Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models, September 2023, https://arxiv.org/pdf/2309.01219v1.pdf (accessed on 17 October 2023).

Zellers, Rowan/Holtzman, Ari/Rashkin, Hannah/Bisk, Yonatan/Farhadi, Ali/Roesner, Franziska/Choi, Yejin, Defending against Neural Fake News, Advances in Neural Information Processing Systems, December 2020, https://arxiv.org/abs/1905.12616 (accessed on 17 October 2023).