SAV-Wegleitung für den Umgang mit künstlicher Intelligenz

1. Einleitung

Künstliche Intelligenz ("KI") ist derzeit in aller Munde und soll gemäss Herstellern von Software in vielen Anwendungen stecken. Auch die Anwaltschaft nutzt vermehrt KI-Systeme, insbesondere im Zusammenhang mit Übersetzungssoftware oder mit der Analyse von grossen Datenmengen bei internen Untersuchungen für die Klientschaft oder bei Due Diligence Prozessen. Weiter kommt aber auch die sogenannte generative KI, d.h. eine KI, die Inhalte wie Bilder oder Texte selbst generiert, immer öfter zum Einsatz (so u.a. zwecks Zusammenfassens, Korrektur oder Verbesserung von Texten).

Die nachfolgenden Guidelines fokussieren auf generative KI. Die Empfehlungen gelten aber für alle Anwendungen, in denen KI zum Einsatz kommt.

Wie alle neuen Technologien eröffnen auch KI-Systeme neue Möglichkeiten, insbesondere Effizienzgewinne, bergen aber auch Risiken. Die vorliegenden Guidelines sollen als Leitfaden dienen, um den Einsatz von KI in der Anwaltspraxis auf verantwortungsvolle Weise umzusetzen. Insbesondere sollten Kanzleien für den Gebrauch von KI-Systemen eine interne Weisung erlassen und die Regeln diesbezüglich festlegen.

KI-Systeme durchlaufen derzeit eine rasante Entwicklung. Die vorliegenden Ausführungen können also sehr schnell überholt sein. Es wird denn auch darauf verzichtet, spezifische KI-Anwendungen hervorzuheben.

2. Risiken im Umgang und sichere Nutzung

2.1. Anwaltsgeheimnis, Datenschutz und andere Geheimhaltungsverpflichtungen

Bei der Auswahl und Nutzung jeglicher Software und damit auch von KI-Anwendungen ist vorgängig zu klären, was mit den Input Daten geschieht, insbesondere wer auf die Daten Zugriff hat und wo diese (zwischen-) gespeichert werden. Zu beachten sind nicht nur das anwaltliche Berufsgeheimnis, sondern auch das Datenschutzgesetz ("DSG") sowie allfällige weitere Geheimhaltungsverpflichtungen. Mitglieder des schweizerischen Anwaltsverbandes müssen zudem auch die Standesregeln berücksichtigen.

Grundsätzlich gibt es folgende Möglichkeiten:

1. Die Software wird im kanzleieigenen Netzwerk installiert und betrieben (sog. *On-Premise* Lösung), und es ist sichergestellt, dass keinerlei Daten dieses interne Netzwerk verlassen oder ausserhalb der Infrastruktur der Anwaltskanzlei gespeichert werden.
2. Einhaltung der Regeln über das Outsourcing, wenn Anwendungen über einen Provider bezogen und allenfalls auf dessen Infrastruktur genutzt werden. Diesbezüglich kann auf die SAV-Wegleitung für IT-Outsourcing und Cloud-Computing verwiesen werden (abrufbar unter https://digital.sav-fsa.ch/digitale-kanzlei-nutzung-von-clouddiensten).
3. Weiter besteht die Möglichkeit der Einholung einer Einverständnis- und Verzichtserklärung der aufgeklärten Klientschaft in Bezug auf Berufsgeheimnis und DSG.

Ansonsten dürfen vertrauliche Informationen, Unternehmensgeheimnisse (Strategien, Finanzdaten etc.), Informationen oder persönliche Daten von Mitarbeitern, Klienten oder Geschäftspartnern oder anderen Personen (egal in welchem Format, d.h. auch einschliesslich Fotos, Videos, usw.) oder Inhalte, die durch Rechte des geistigen Eigentums, insbesondere Urheberrechte, geschützt sind, nicht in KI-Systeme eingegeben werden.

2.2. (unabhängige) Überprüfung der Resultate

KI-Systeme sind weder allwissend noch arbeiten sie immer perfekt, und die Ergebnisse von KI-Systeme können falsch, unzureichend oder unvollständig sein. Es ist demzufolge von grösster Wichtigkeit, die Resultate, den sogenannten "*Output*", unabhängig und kritisch zu überprüfen und allenfalls zu korrigieren bzw. zu ergänzen.

Dabei ist es wichtig zu wissen, dass eine KI nicht dazu in der Lage ist, die von ihr generierten Resultate zu überprüfen. Man kann also zum Beispiel nicht einfach ein KI-System fragen, ob der gelieferte Output der Wahrheit entspricht, da KI-Systeme dazu nicht in der Lage sind.

Fehler oder fehlerhafte Outputs von KI-Anwendungen können insbesondere folgenden Ursprung haben:

- Halluzinationen, d.h. die KI "erfindet" den Output.
- Falsche oder fehlende Informationen, weil dem KI-System die Datengrundlage fehlt. Ein KI-System kann beispielsweise auf einer in der Vergangenheit liegender Datengrundlage trainiert werden und demzufolge nach diesem Datum liegende Ereignisse nicht kennen. KI-Systeme neigen zudem desto eher zu Halluzinationen, je weniger Quellmaterial für eine bestimmte Fragestellung in das Training eingeflossen ist.
- Sycophancy: ein KI-Modell passt seine Antworten so an, dass sie mit der Sichtweise des Benutzers übereinstimmen, selbst wenn diese Sichtweise objektiv verzerrend ist.

Zu beachten ist auch, dass KI-Systeme voreingenommen sein können. Solche sogenannte Biases können aufgrund des Datensatzes, der für das Training des KI-Systems verwendet wurde, die Art und Weise des Trainings an sich sowie Modellierungsentscheide der Programmierer entstehen. Das ist primär bei der Verwendung von KI-Systemen zur Datenanalyse ein Thema, muss aber auch bei generativen KI-Systemen im Hinterkopf behalten werden.

2.3. Haftung

Als Auftragnehmer haftet der Rechtsanwalt bzw. die Rechtsanwältin für eine allfällige Schlechterfüllung seines oder ihres Auftrags. Sie können sich nicht darauf berufen, die KI habe einen Fehler gemacht.

Beim Einsatz von KI-Systemen in Absprache mit der Klientschaft (z.B. beim Durchforsten grosser Datenmengen) empfiehlt es sich, im Vorfeld die Verantwortung für das Resultat anzusprechen und im Rahmen des rechtlich Zulässigen eine Haftungsbeschränkung zu vereinbaren.

2.4. Urheberrecht

Aktuell kontrovers diskutiert ist die Frage, ob die Nutzung von urheberrechtlich geschütztem Material, insbesondere zum Training von Large Language Models ("LLMs") oder von Bild- bzw. Videogeneratoren. Umstritten ist, ob solche Trainingsdaten ohne Zustimmung von Rechteinhaberinnen deren Urheberrechte verletzt bzw. ob die Nutzung zu Trainingszwecken überhaupt eine urheberechtsrelevante Nutzung darstellt. Diese Fragen steht allerdings weniger bei der reinen Nutzung von KI-Anwendungen im Vordergrund; sie können aber relevant sein, wenn Kanzleien ihre eigenen LLMs trainieren oder bestehende LLMs ergänzen.

Ohne entsprechendes Nutzungsrecht können Inputdaten für das Training von KI-Systemen oder zur Generierung von Outputs allenfalls Urheberrechte Dritter verletzen.

Der Output gilt, sofern es sich um eine reine KI-Schöpfung handelt, grundsätzlich nicht als geistige Schöpfung im Sinne des Urheberrechts. Wird ein KI-System nur für die Ideensammlung, für einen

ersten Entwurf o.ä. eingesetzt, kann in der Folge dennoch eine geistige Schöpfung mit individuellem Charakter entstehen, und es kann trotz Nutzung von KI-Systemen ein urheberrechtlich geschütztes Werk entstehen. Nicht gänzlich ausgeschlossen ist es auch, dass ein bestimmter KI-generierter Output Urheberrechte Dritter verletzt, indem der Output einem urheberrechtlich geschützten Werk zu stark ähnlich bleibt. Dies ist aber sehr unwahrscheinlich.

2.5. Hinweispflicht

Einzelne Anbieter haben in ihren Nutzungsbedingungen Bestimmungen, welche die Nutzer verpflichten, den Einsatz von KI offenzulegen. Vor dem Einsatz einer KI-Anwendung sollten daher, nicht zuletzt aus diesem Grund, die Nutzungsbedingungen des betreffenden Anbieters geprüft werden.

Abgesehen davon könnte eine Hinweispflicht allenfalls auch dann bestehen, wenn der Klient als Auftraggeber die höchstpersönliche Ausführung des Auftrags durch die Anwältin oder den Anwalt verlangt oder zu Recht erwartet.

3. Regulierung von KI-Systemen

3.1. Schweiz

In der Schweiz hat der Bundesrat im November 2023 das UVEK beauftragt, bis Ende 2024 mögliche Ansätze zur Regulierung von KI aufzuzeigen. Diese Auslegeordnung ist nun am 12. Februar 2025 erschienen (siehe dazu https://www.bakom.admin.ch/bakom/de/home/digital-und-internet/strategie-digitale-schweiz/ki_leitlinien.html). Im Bericht werden mögliche Regulierungsansätze diskutiert und folgende Analysen vorgenommen:

- Länderanalyse: Untersuchung der KI-Regulierungen in 20 Ländern.
- Rechtliche Basisanalyse: Prüfung der Vereinbarkeit mit dem bestehenden Schweizer Recht.
- Sektorielle Analyse: Erhebung bestehender und geplanter Regulierungen in verschiedenen Wirtschaftsbereichen.
- Wirtschafts- und europapolitische Einschätzung: Bewertung der Auswirkungen auf die Schweizer Wirtschaft und das Abkommen zur gegenseitigen Anerkennung von Konformitätsbewertungen (MRA) mit der EU.

Gleichentags hat der Bundesrat entschieden:

- Die KI-Konvention des Europarats ins Schweizer Recht zu übernehmen (siehe dazu nachstehend Ziff. 3.3). Diesbezüglich sind primär zusätzliche Massnahmen insbesondere in Bereichen wie Transparenz, Haftung und Risikoabschätzung zu treffen.
- Notwendige Gesetzesanpassungen möglichst sektorbezogen vorzunehmen. Eine allgemeine, sektorübergreifende Regulierung wird sich auf zentrale, grundrechtsrelevante Bereiche, wie beispielsweise den Datenschutz beschränken. Folglich wird der AI-Act der EU in der Schweiz nicht umgesetzt oder nachvollzogen; zu beachten ist aber, dass dieser aufgrund des Auswirkungsprinzips trotzdem Einfluss auf Schweizer Anbieter und Anwender haben wird (siehe dazu Ziff. 3.2).
- Zusätzlich zu den Gesetzesänderungen auch rechtlich nicht verbindliche Massnahmen zur Umsetzung der Konvention zu erarbeiten. Zu diesen können Selbstdeklarationsvereinbarungen oder Branchenlösungen gehören.

Der Bundesrat hat für die Umsetzung folgende Ziele definiert:

- Stärkung des Innovationsstandorts Schweiz
- Wahrung des Grundrechtsschutzes (einschliesslich Wirtschaftsfreiheit)
- Förderung des Vertrauens der Bevölkerung in KI

Die nächsten Schritte werden die folgenden sein: Das EJPD wird mit dem UVEK und dem EDA bis Ende 2026 eine Vernehmlassungsvorlage erstellen, die die notwendigen gesetzlichen Massnahmen zur Umsetzung der KI-Konvention des Europarates festlegt. Ebenfalls bis Ende 2026 wird das UVEK zusammen mit dem EJPD, dem EDA und dem WBF einen Plan für die weiteren Massnahmen von rechtlich nicht verbindlicher Natur erarbeiten. Dieser soll insbesondere auch die Vereinbarkeit des Schweizer Ansatzes mit jenen der wichtigsten Handelspartner berücksichtigen. Bundesinterne und -externe Anspruchsgruppen sollen in die Arbeiten einbezogen werden.

3.2. EU

Am 12. Juli 2024 hat die EU, eine der ersten umfassenden Regulierungen zum Thema KI im Amtsblatt publiziert ("KI-Gesetz" oder "AI-Act"). Der AI-Act entfaltet wie folgt Wirkung:

- 1. August 2024: Inkrafttreten des AI-Acts.
- 2. Februar 2025: Verbot gewisser Praktiken bezüglich KI-Systemen mit inakzeptablem Risiko.
- 2. August 2025: Anwendbarkeit der Regeln zu Mehrzweck KI-Systeme und des Sanktionssystems.
- 2. August 2026: Restliche Bestimmungen des AI-Acts treten in Kraft (mit Ausnahme der Nachstehenden).
- 2. August 2027: Restliche Bestimmungen bezüglich der Klassifizierung von KI-Systemen welche nicht in Anhang 3 des AI-Acts als Hochrisiko-KI-Systemen und die diesbezüglichen Bestimmungen kommen zur Anwendung.

Der AI-Act hat sowohl in Bezug auf die betroffenen Personen als auch auf die territoriale Reichweite einen breiten Anwendungsbereich. Es gilt nicht nur für Entwickler von KI-Systemen und Personen, die diese auf den Markt bringen oder in Betrieb nehmen (im Gesetz als „Anbieter" bezeichnet), sondern auch für alle Personen, die im Rahmen ihrer beruflichen Tätigkeit ein KI-System nutzen (als „Anwender" bezeichnet). Der AI-Act erfasst Anbieter und Anwender innerhalb der EU, aber in gewissem Umfang auch solche ausserhalb der EU, die auf den EU-Markt zugreifen. Wer als Anbieter oder Anwender qualifiziert, tut gut daran zu prüfen, ob der AI-Act auf sie Anwendung findet.

Der AI-Act verfolgt einen risikobasierten Ansatz, und KI-Systeme werden danach in vier Kategorien eingeteilt:

1. KI-Systeme mit inakzeptablem Risiko. Diese sind verboten. Dazu gehören insbesondere biometrische Echtzeit-Fernidentifizierungssysteme ("Facial Recognition") sofern diese nicht der Strafverfolgung dienen und unter strengen Auflagen angewandt werden (z.B. vorgängige Genehmigung des Einsatzes durch eine Justizbehörde). Weiter fallen darunter das sogenannte "Social Scoring", d.h. die Klassifizierung der Vertrauenswürdigkeit natürlicher Personen auf der Grundlage ihres sozialen Verhaltens oder bekannter oder vorhergesagter persönlicher Eigenschaften oder Persönlichkeitsmerkmale, sofern dies zu einer ungerechtfertigten oder unverhältnismässigen Schlechterstellung oder Benachteiligung dieser Personen führt.
2. KI-Systeme mit hohem Risiko sind unter strengen Auflagen (Risikomanagement, Daten Governance, technische Dokumentation, Aufzeichnungspflichten, Transparenzanforderungen, menschliche Aufsicht sowie Anforderungen an Genauigkeit, Robustheit und Cybersicherheit, etc.) erlaubt. Dazu gehören z.B. KI-Systeme im Zusammenhang mit Bewerbungen, Beförderungen oder Kündigungen von Arbeitsverhältnissen, mit der Prüfung von Ansprüchen auf Sozialhilfe, mit der Bewertung von Schülern, mit Zulassungsprüfungen zur Universität oder mit Prognosen über die Rückfälligkeit von Straftätern, etc.
3. KI-Systeme mit beschränkten Risiken sind unter limitierten Transparenzanforderungen (z.B. Information des Nutzers, dass er es mit einem KI-System zu tun hat, oder Kennzeichnung von mit künstlicher Intelligenz generierten Inhalten) sie zulässig.
4. KI-Systeme mit keinem / niedrigem Risiko, wie z.B. im Zusammenhang mit Videospielen, sind zulässig und bleiben weitgehend unreguliert.

Im AI-Act ist die Schaffung eines Europäischen Ausschusses für künstliche Intelligenz vorgesehen, und auch auf nationaler Ebene sollen Behörden geschaffen werden. Diese werden fehlbare Unternehmen auch sanktionieren können. Nicht im AI-Act enthalten sind Möglichkeiten zur individuellen Rechtsdurchsetzung.

3.3. Europarat

Neben der EU hat auch der Europarat an einer KI-Konvention gearbeitet. Am 17. Mai 2024 erliess der Ministerrat die erst in englischer und französischer Sprache verfügbare Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law ("KI-Konvention"). Der Fokus der KI-Konvention ist die Sicherstellung des Einsatzes von KI in einer Weise, die die Grundrechte, die Demokratie und die Grundsätze der Rechtsstaatlichkeit achtet. Sie steht zur Unterzeichnung und Ratifizierung den Mitgliedstaaten und zudem jenen anderer Staaten offen, die sich an ihrer Ausarbeitung beteiligt haben. Die KI-Konvention wird in Kraft treten, sobald ihr 5 Staaten, wovon mindestens 3 Mitgliedstaaten zugestimmt haben.

Vorstand SAV, 14 Juni 2024

Anhang: Glossar

AI oder Artificial Intelligence	Siehe KI oder künstliche Intelligenz
AI-Act oder KI-Gesetz	EU-Gesetz über die Regulierung von KI. (https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX:32024R1689)
Black Box Syndrome	Beschreibt die Unmöglichkeit bzw. die besondere Schwierigkeit die Herangehensweise und den Lösungsweg von KI-Systemen zu überprüfen. Meist sind diese für User nicht und selbst für die Programmierer einer KI schwer überprüfbar
Bias	Voreingenommenheit eines KI-Systems aufgrund des Trainingsdatensatzes, der Art und Weise des Trainings oder Modellierungsentscheiden der Programmierer
GPT	Generative Pre-trained Transformer
Halluzination	Angebliche (falsche) Fakten, die nicht auf realen Daten oder Ereignissen basieren, aber als solche dargestellt werden.
Input	Daten, die in ein KI-System eingespiesen werden.
KI oder künstliche Intelligenz	künstliche Intelligenz bezeichnet die Fähigkeit von Maschinen, Aufgaben auszuführen, die normalerweise menschliche Intelligenz erfordern würden. Dazu gehören Problemlösung, Lernen, Spracherkennung, Entscheidungsfindung und vieles mehr. KI-Algorithmen und -Systeme können Daten analysieren, Muster erkennen und darauf basierend Vorhersagen treffen oder Entscheidungen fällen.
LLM oder Large Language Model	Maschinelles Lernmodell, das Aufgaben im Bereich Natural Language Processing ausführen kann
NLP oder Natural Language Processing	Verarbeitung natürlicher Sprache durch ein KI-System
Output	Resultat einer generativen KI
Prompt	Input in eine generative KI mit den Instruktionen an das KI-System zur Erzeugung des Outputs