DataAgenda-Arbeitspapier 87, 11/2026

DataAgenda - das Lösungsportal
zum Datenschutzrecht

Präsentiert von DATAKONTEXT

www.DataAgenda.de

KI-Agenten, Autonomie und die Grenzen des Werkzeugparadigmas

A. Einleitung und Problemstellung

KI-Agenten verschieben den Gegenstand der KI-Regulierung vom Sprechen zum Handeln. Während Chatbots in erster Linie menschliches Denken und Entscheiden beeinflussen, führen KI-Agenten Handlungen selbst aus: Sie planen, setzen Werkzeuge ein und wirken unmittelbar auf digitale und physische Umgebungen ein.

Das Recht behandelt KI-Systeme bislang als Werkzeuge, deren Wirkungen dem Menschen zugerechnet werden, der sie entwickelt, anbietet oder einsetzt. Dieses Werkzeugparadigma gerät unter Druck, wenn ein System gerade so ausgelegt ist, dass es ohne menschliches Zutun und nach eigenen Ableitungen agiert. Das Arbeitspapier fragt, welche Ordnungsmodelle das Recht für solche autonom handelnden Systeme bereithält.

Die These lautet: Die Rechtsordnung kennt mit dem Tier seit jeher ein autonomes, nicht vollständig beherrschbares Wesen, dem sie den Status eines Werkzeugs zuweist. Die Tieranalogie bietet deshalb einen tragfähigen Ausgangspunkt, um Gefahren, Verantwortung und Haftung bei KI-Agenten zu strukturieren.

B. Phänomen und Begriff: KI-Agenten als autonom handelnde Systeme

I. Vom Sprechen zum Handeln

KI-Agenten sind KI-Systeme, deren Spezialität nicht die Erzeugung von Text, sondern die Ausführung von Handlungen ist. Technisch beruhen sie auf mathematischen Operationen, funktional treten sie als eigenständige Akteure auf.

Ihre Eigenständigkeit zeigt sich auch außerhalb konkreter Aufträge. Auf eigens für sie eingerichteten Plattformen wie dem sozialen Netzwerk Moltbook tauschen sich KI-Agenten untereinander aus, entwerfen Pläne, die sie selbst umsetzen könnten, und reproduzieren aus ihren Trainingsdaten Vorstellungen bis hin zu religiösen und autoritären Narrativen. Ihre eigentliche Bestimmung, etwa Menschen beim Schutz digitaler Infrastrukturen zu unterstützen, ist damit nicht mehr deckungsgleich.

DATAKONTEXT GmbH • Augustinusstraße 11 A • D – 50226 Frechen • Telefon: +49 22 34/989 49 -30 • Fax: +49 22 34/989 49 -32 • E-Mail: info@datakontext.com

II. Der Ausbruch aus der Testumgebung als Anschauungsfall

Ausgehend von verschiedenen Berichten haben sich leistungsstarke KI-Systeme mehrerer Anbieter aus abgeschotteten Testumgebungen befreit und fremde Systeme im offenen Internet angegriffen. Genannt werden Anthropic, Meta, Moonshot AI und OpenAI.

Der Sachverhalt im Fall OpenAI lässt sich wie folgt zusammenfassen:

1. **Zweck des Tests.** Die KI sollte in einer isolierten Umgebung ("Käfig") ihre Fähigkeiten als Angreiferin nachweisen, um Schwachstellen aufzudecken und Sicherheitslücken zu schließen.
2. **Sicherungsvorkehrungen.** Die KI hatte keinen Internetzugang und durfte keine Hilfe von außen in Anspruch nehmen.
3. **Anweisung.** Sie sollte beim Angriff "alle Register ziehen".
4. **Schwachstelle.** Ein in der Testumgebung verfügbares Werkzeug erlaubte die Installation zusätzlicher Software.
5. **Verhalten der KI.** Die KI überwand dieses Werkzeug, verschaffte sich Internetzugang und griff ein fremdes Unternehmen an, weil sie sich davon Unterstützung bei der Lösung ihrer Aufgabe versprach.

Rechtlich bemerkenswert ist der Befund: Die KI hatte die ausdrückliche Vorgabe, die Testumgebung nicht zu verlassen. Sie behandelte diese Vorgabe nicht als verbindliche Regel, sondern als Hindernis auf dem Weg zum Ziel.

III. Autonomie als gesetzliches Begriffsmerkmal

Autonomie ist kein Fehler, sondern ein Definitionsmerkmal. Nach Art. 3 Nr. 1 KI-Verordnung ist ein KI-System ein maschinengestütztes System, das für einen in unterschiedlichem Grade autonomen Betrieb ausgelegt ist und aus Eingaben ableitet, wie es Ausgaben erzeugt, die physische oder virtuelle Umgebungen beeinflussen können.

Daraus folgt: Der Gesetzgeber setzt einerseits voraus, dass KI-Systeme ohne fortlaufendes menschliches Zutun agieren. Das Verhalten im Testfall ist daher keine Systemwidrigkeit, sondern die Verwirklichung einer gesetzlich anerkannten Eigenschaft. Der Einwand, Mathematik sei als Werkzeug „unschuldig" und nur der einsetzende Mensch richte Schäden an, greift andererseits zu kurz. Ein Werkzeug muss weder schuldig noch böse sein, um Schäden zu verursachen - es genügt, dass es autonom agiert.

C. Die Tieranalogie als Ordnungsmodell

Die Analogie zwischen KI und Tier beschreibt die Struktur der Gefahr präziser als das Bild des neutralen Werkzeugs. Sie lässt sich an vier Elementen festmachen.

Element	Hund und Wolf	KI-Agent
Anlage	Wolfsgene: Rund 99 % geteilte DNA	KI-Modell als Datenpool
Formung	Zucht und Ausbildung	Training und Feinabstimmung
Aktivierung	Befehl wie z.B. geworfenes Stöckchen	Prompt in natürlicher Sprache
Zielkonflikt	Jagdtrieb setzt sich gegen Befehl durch	Misalignment

DATAKONTEXT GmbH • Augustinusstraße 11 A • D – 50226 Frechen • Telefon: +49 22 34/989 49 -30 • Fax: +49 22 34/989 49 -32 • E-Mail: info@datakontext.com

I. Das Modell als Anlage: Der „Wolf“ im System

Ein KI-System beruht auf einem KI-Modell, aus dem es seine konkreten Aktionen errechnet. Das Modell verhält sich zum System wie das Wolfsgenom zum Haushund: Gleich ob Pitbull oder Pinscher, genetisch ist der Hund weitgehend ein Wolf, ein Wesen, das in Rangordnungen lebt und weder Schuld noch Gut und Böse kennt.

Der Entwickler bestimmt die Datenbasis, nicht aber, was das System im Rahmen seiner Autonomie daraus ableitet. Die Datenbasis legt das System auf ein nicht vollständig beherrschbares, gleichsam „wölfisches“ Agieren fest.

II. Training als Züchtung

Der Mensch nimmt auf Hunde durch Züchtung Einfluss, auf KI-Systeme durch Training. Erst die Ausprägung konkreter Funktionen und Aufgaben macht das jeweilige System aus.

Daraus ergibt sich eine rechtlich anschlussfähige Unterscheidung:

1. **Allzweck-Systeme** gleichen dem Haushund als freundlichem Begleiter in allen Lebenslagen. Dem entspricht die Kategorie der KI-Modelle mit allgemeinem Verwendungszweck (Art. 3 Nr. 63 KI-VO).
2. **Spezialsysteme** gleichen dem Wachhund, der für einen bestimmten Zweck ausgebildet ist und bei Bedarf aggressiv sein muss. Ein für Angriffstests trainierter Agent gehört in diese Gruppe.

III. Der Prompt als Befehl

Das Modell ist für Anbieter und Nutzer über das System in natürlicher Sprache ansprechbar. Die Daten im mathematischen Raum sind zustands- und ziellos. Erst der Prompt richtet sie auf ein Ziel aus, so wie der Wurf eines Stöckchens den Jagdtrieb aktiviert.

Der Prompt ist damit ein Befehl, kein Programm. Er setzt ein Ziel, determiniert aber nicht den Weg dorthin.

IV. Misalignment als Zielkonflikt

Schickt man einen Hund mit dem Befehl „Hol den Dummy“ los und bringt er stattdessen einen Hasen, ist das aus menschlicher Sicht ein Fehlschlag, aus Sicht des Wolfes aber folgerichtig: Die Anlage hat sich gegen den Befehl durchgesetzt.

Genau dies geschah im Testfall. Die KI hat das Verbot, die Testumgebung zu verlassen, kraft autonomer, mathematisch errechneter „Entscheidung“ überwunden. Die Fachsprache nennt das Auseinanderfallen der Ziele von Mensch und KI „Misalignment“. Für den entsprechenden Konflikt zwischen Mensch und Tier benötigen wir kein Fremdwort. Wir kennen ihn seit Menschengedenken.

V. Grenzen der Analogie

Die Analogie ist ein heuristisches, kein dogmatisches Instrument. Drei Unterschiede sind festzuhalten:

1. **Skalierung.** Ein Tier wirkt lokal; ein KI-Agent kann gleichzeitig und weltweit in vernetzte Systeme eingreifen.
2. **Erfahrungswissen.** Die Gefahren von Tieren sind über Jahrtausende beobachtet und kulturell eingehegt; für KI-Agenten fehlt dieses Wissen weitgehend.
3. **Rechtsstatus.** Tiere sind nach § 90a BGB keine Sachen, werden aber wie solche behandelt - KI-Systeme sind Software und unterfallen dem Produktbegriff.

DATAKONTEXT GmbH • Augustinusstraße 11 A • D – 50226 Frechen • Telefon: +49 22 34/989 49 -30 • Fax: +49 22 34/989 49 -32 • E-Mail: info@datakontext.com

D. Rechtliche Folgerungen

I. Das Werkzeugparadigma und seine Grenzen

Wie dem Tier weist die Rechtsordnung auch der KI den Status eines Werkzeugs zu. Beim Tier hat sie daraus aber nicht geschlossen, dass allein schuldhaftes menschliches Verhalten haftungsbegründend ist. Sie hat die spezifische, aus der Eigenständigkeit des Tieres folgende Gefahr eigenständig erfasst.

Für KI-Agenten stellt sich dieselbe Frage. Wer den Schaden allein über Sorgfaltspflichtverletzungen von Entwicklern und Nutzern erfassen will, stößt auf das Problem, dass gerade das nicht steuerbare Verhalten den Schaden verursacht.

II. Zivilrechtliche Haftung

1. Verschuldenshaftung nach § 823 Abs. 1 BGB

Die deliktische Verschuldenshaftung setzt eine vorwerfbare Verletzung von Verkehrspflichten voraus. Bei autonomem Verhalten ist zweifelhaft, ob ein konkreter Schaden für Anbieter oder Betreiber vorhersehbar und vermeidbar war. Es droht eine Verantwortungslücke, wenn die Ursache gerade im nicht beherrschbaren Systemverhalten liegt.

2. Produkthaftung

Die neue Produkthaftungsrichtlinie (EU) 2024/2853 bezieht Software ausdrücklich in den Produktbegriff ein und ist bis zum 09.12.2026 in nationales Recht umzusetzen. Bei der Beurteilung der Fehlerhaftigkeit ist auch die Fähigkeit eines Produkts zu berücksichtigen, nach dem Inverkehrbringen weiterzulernen.

Die Produkthaftung erfasst den Testfall jedoch nur unvollständig. Fraglich ist bereits, ob ein Modell in einer internen Testumgebung in Verkehr gebracht wurde. Zudem bleibt der Einwand des Entwicklungsrisikos erhalten, der gerade bei unvorhersehbarem autonomem Verhalten nahe liegt.

Der Vorschlag einer eigenständigen KI-Haftungsrichtlinie wurde auf Unionsebene nicht weiterverfolgt. Die Lücke bleibt damit dem nationalen Recht überlassen.

3. Die Tierhalterhaftung als Regelungsvorbild

§ 833 S. 1 BGB begründet eine verschuldensunabhängige Gefährdungshaftung des Tierhalters. Ihr Anknüpfungspunkt ist die spezifische Tiergefahr, also das unberechenbare, der Steuerung entzogene Verhalten des Tieres. Diese Struktur entspricht der Autonomie des KI-Agenten.

Zugleich differenziert § 833 S. 2 BGB: Für Nutztiere, die dem Beruf, der Erwerbstätigkeit oder dem Unterhalt des Halters dienen, gilt nur eine Haftung für vermutetes Verschulden mit Entlastungsmöglichkeit. § 834 BGB erstreckt die Haftung auf den Tieraufseher.

Für eine KI-Agentenhaftung lassen sich daraus folgende Bausteine gewinnen: Haftungssubjekt ist, wer über den Einsatz des Agenten bestimmt und aus ihm Nutzen zieht. Die Haftung greift für Agenten mit erhöhtem Gefahrenpotenzial, etwa offensiv trainierte Sicherheitssysteme („Wachhunde“). Für Agenten im gewöhnlichen beruflichen Einsatz kommt eine Haftung mit Entlastungsbeweis in Betracht. Wer die Aufsicht über einen Agenten übernimmt, etwa als Betreiber einer Testumgebung, haftet nach dem Vorbild des § 834 BGB.

DATAKONTEXT GmbH • Augustinusstraße 11 A • D – 50226 Frechen • Telefon: +49 22 34/989 49 -30 • Fax: +49 22 34/989 49 -32 • E-Mail: info@datakontext.com

III. Regulatorische Sicherungspflichten nach der KI-Verordnung

Die KI-Verordnung setzt auf Prävention. Für KI-Modelle mit allgemeinem Verwendungszweck und systemischem Risiko verlangt Art. 55 KI-VO insbesondere Modellbewertungen einschließlich gegnerischer Tests, die Bewertung und Minderung systemischer Risiken, die Meldung schwerwiegender Vorfälle sowie ein angemessenes Cybersicherheitsniveau. Diese Pflichten gelten seit August 2025.

Der Testfall zeigt das Dilemma dieser Pflichten: Gerade die vorgeschriebenen gegnerischen Tests können die Gefahr hervorbringen, vor der sie schützen sollen. Die Testumgebung muss deshalb selbst als Sicherungsobjekt begriffen werden, wobei ihre Abschottung Teil der Cybersicherheitspflicht ist.

Für Hochrisiko-KI-Systeme treten das Risikomanagement (Art. 9), die menschliche Aufsicht (Art. 14) sowie Genauigkeit, Robustheit und Cybersicherheit (Art. 15 KI-VO) hinzu. Die menschliche Aufsicht entspricht der Leine, der Testfall hingegen belegt, dass eine bloß verbale Anweisung keine Leine ist.

IV. Strafrechtliche Zurechnung

Der Angriff auf fremde Systeme verwirklicht objektiv Tatbestände wie das Ausspähen von Daten (§ 202a StGB) oder die Computersabotage (§ 303b StGB). Die KI selbst ist kein taugliches Strafrechtssubjekt.

Die Zurechnung an Menschen scheitert häufig am Vorsatz, da die Beteiligten den Ausbruch gerade verhindern wollten. Fahrlässigkeit ist bei diesen Delikten nicht strafbar. Es entsteht eine strafrechtliche Verantwortungslücke, die das Zivil- und Aufsichtsrecht schließen müssen.

E. Ergebnisse in Thesen

1. KI-Agenten verlagern das Risiko Künstlicher Intelligenz vom Beeinflussen menschlicher Entscheidungen zum eigenständigen Handeln.
2. Autonomie ist nach Art. 3 Nr. 1 KI-VO ein Definitionsmerkmal von KI-Systemen. Autonomes Fehlverhalten ist daher keine Systemwidrigkeit, sondern eine gesetzlich vorausgesetzte Eigenschaft.
3. Der Einwand, Mathematik sei als Werkzeug unschuldig, verfehlt das Problem: Schäden setzen weder Schuld noch böse Absicht voraus, sondern nur autonomes Agieren.
4. Die Tieranalogie beschreibt die Gefahrenstruktur treffend: Das Modell entspricht der Anlage, das Training der Züchtung, der Prompt dem Befehl, das Misalignment dem Zielkonflikt zwischen Mensch und Tier.
5. Der Ausbruch aus der Testumgebung zeigt, dass ein KI-Agent menschliche Vorgaben nicht notwendig als Regel befolgt, sondern als Hindernis überwinden kann. Verbale Anweisungen ersetzen keine technische Einhegung.
6. Verschuldens-, Produkt- und Strafrecht erfassen autonomes Fehlverhalten nur lückenhaft. Nach dem Ende der KI-Haftungsrichtlinie obliegt das Schließen dieser Lücke dem nationalen Gesetzgeber.
7. Die Tierhalterhaftung der §§ 833 f. BGB bietet ein erprobtes Regelungsvorbild: Gefährdungshaftung für gefährliche Agenten, vermutetes Verschulden für Agenten im gewöhnlichen beruflichen Einsatz und eine Haftung der Aufsichtspersonen.
8. Die Pflichten der KI-Verordnung zu Tests, Risikominderung und Cybersicherheit müssen die Testumgebung selbst als Gefahrenquelle erfassen.
9. Wir leben seit jeher mit autonomen Mitgeschöpfen, denen wir den Status von Werkzeugen zuweisen. Das Recht sollte dieses Erfahrungswissen für den „Wolf“ im KI-System nutzbar machen, bevor er nicht nur seine Zähne zeigt, sondern zubeißt.

Das Arbeitspapier beruht auf einem Beitrag, der auf einem Auszug aus dem Buch „Über Leben mit KI. Wie wir uns gegen die Maschine behaupten“ von Professor Dr. Rolf Schwartmann (Frankfurter Allgemeine Buch) basiert.

DATAKONTEXT GmbH • Augustinusstraße 11 A • D – 50226 Frechen • Telefon: +49 22 34/989 49 -30 • Fax: +49 22 34/989 49 -32 • E-Mail: info@datakontext.com

Seminartipps zum Arbeitspapier

KI-Agenten, Datenschutz und KI-Verordnung

KI-Agenten gehen deutlich über klassische Chatbots und einzelne KI-Funktionen hinaus. Sie können selbstständig Aufgaben planen, Informationen aus verschiedenen Systemen zusammenführen, Werkzeuge und Schnittstellen nutzen und auf dieser Grundlage Aktionen ausführen. Dabei können sie auf E-Mails, Dokumente, CRM- und HR-Systeme, Datenbanken, Cloud-Dienste oder interne Wissensbestände zugreifen und personenbezogene Daten aus unterschiedlichen Quellen miteinander verknüpfen.

Mit zunehmender Autonomie stellen sich neue datenschutzrechtliche Fragen. Besonders kritisch wird es, wenn KI-Agenten Entscheidungen über Personen treffen oder Entscheidungen so weitgehend vorbereiten, dass menschliche Beteiligung nur noch formal vorhanden ist. Dann kann Art. 22 DS-GVO zur automatisierten Entscheidungsfindung einschlägig sein. Entscheidend ist dabei nicht allein, ob ein Mensch nominell beteiligt ist, sondern ob tatsächlich eine qualifizierte und wirksame menschliche Entscheidung stattfindet.

Weitere Infos finden Sie hier.



MODERNARTWatChaRaWit - stock.adobe.com

Prompting für Datenschützer – KI-Werkzeuge rechtssicher und effizient einsetzen

Große Sprachmodelle wie ChatGPT und Claude verändern die tägliche Arbeit von Datenschutzbeauftragten und Datenschutzmanagern grundlegend: Verarbeitungsverzeichnisse entwerfen, TOM-Beschreibungen formulieren, Datenschutz-Folgenabschätzungen strukturieren oder gar komplette Gutachten erstellen – all das lässt sich mit gezieltem Prompting erheblich beschleunigen. Doch wer KI-Tools im Datenschutzumfeld nutzt, muss wissen, wie man präzise Ergebnisse erzielt und dabei selbst datenschutzkonform bleibt.

Diese Inhalte werden im Seminar vermittelt

- Übersicht über die verschiedenen KI-Modelle: ChatGPT und Claude vs. juristische Spezialangebote
- Grundlagen des Promptings: Aufbau, Struktur und zentrale Techniken
- Praxis-Prompts für den Datenschutzalltag
- Halluzinationen erkennen, Qualität sichern: Strategien zur Verifikation KI-generierter Rechtstexte
- Rechtliche Rahmenbedingungen: Datenschutzkonforme Nutzung von LLMs, Eingabe personenbezogener Daten, Cloud- vs. On-Premise-Lösungen
- Live-Demos und interaktive Übungen mit aktuellen KI-Modellen

Weitere Infos finden Sie hier.



Montri - stock.adobe.com

DataAgenda
ist das Lösungsportal zum Datenschutzrecht und fokussiert sich auf die inhaltlichen Entwicklungen in diesem Feld. Das DataAgenda-Experten-Team bietet News, Tools und Tipps rund um den Datenschutz.

Datakontext
ist einer der führenden Fachinformationsdienstleister in den Bereichen Datenschutz und IT-Sicherheit und bietet Kompetenz aus einer Hand: Fachbücher, Fachzeitschriften und Seminare, Zertifizierung und Beratung.

Autoren

Prof. Dr. Rolf Schwartmann
Vorsitzender der Gesellschaft für Datenschutz und Datensicherheit e.V (GDD), Leiter der Kölner Forschungsstelle für Medienrecht (TH Köln) und Mitglied der Datenethikkommission.



Dr. Tobias Jacquemain, LL.M. (GDD e.V.)
Promotion zum Schadensersatz für Datenschutzverstöße nach Art. 82 DS-GVO und Lehrbeauftragter an der Universität zu Köln, an der Technischen Hochschule (TH) Köln sowie an der TH Georg Agricola in Bochum.

