EasyChair Preprint

№ 15598

# Bias and AI in Judicial Application

Thomas Hrdinka



December 18, 2024

# BIAS UND KI BEI GERICHTLICHER ANWENDUNG

Thomas Hrdinka

Zivilingenieur, Universität Wien
Ocwirkgasse 22, 1210 Wien, AT
thrdinka@zth.at; http://www.zth.at

***Schlagworte:*** ***Künstliche Intelligenz, Bias, Beweismittel, IT-Forensik***

***Abstract:*** *In dieser Arbeit werden die internationalen Entwicklungen für die Eindämmung von Bias[1] und die Ursachen und Konsequenzen von nicht beachteten Bias in der IT-Forensik, die bei Ermittlungs- und Sachverständigentätigkeiten genutzt wird, analysiert. Die KI bietet dafür neuartige und interessante Möglichkeiten an, wobei nicht alle Arten von Bias bei Einsatz einer KI relevant sind. Zu beachten ist auch, dass mehrere Bias nacheinander auftreten können, und so wird die Fehlerhaftigkeit der Ermittlungsergebnisse noch mehr verstärkt. LLMs[2] als spezielle KI-Systeme können durchaus verschiedene Anwendungsgebiete in der Sachverständigenarbeit oder Ermittlungsarbeit anbieten, wobei zu beachten ist, dass LLMs stark zu Halluzinationen tendieren. Schließlich sollen internationale Entwicklungen zum Einsatz von KI bei Gericht erhoben werden, und mit welchen Problemen speziell beim Einsatz von LLMs dabei zu rechnen ist.*

## 1. Einleitung

Die Forensik oder forensische Wissenschaft wird „*als die Wissenschaft der Spur definiert. Alle Methoden, Techniken und Prozesse, die zur Suche, Sicherung, Analyse und Auswertung von Spuren verwendet werden können, sind daher Teilbereiche der Forensik.*“[3] Die „*Forensische Informatik*“ bezeichnet „*die Anwendung wissenschaftlicher Methoden der Informatik auf Fragen des Rechtssystems. Insbesondere stellt die forensische Informatik Methoden zur gerichtsfesten Sicherung und Verwertung digitaler Spuren bereit.*“[4]

Letztlich geht es demnach bei der Forensischen Informatik um die Suche, Sicherstellung Analyse und Auswertung elektronischer Beweismittel, und daher „*unterliegen die Anforderungen an die Glaubwürdigkeit, der Wiederholbarkeit, der Integrität und Dokumentation der Beweissicherung hohen Maßstäben, denn die Gefahr, dass durch die forensische Arbeit digitale Spuren verändert oder vernichtet werden ist latent hoch*“.[5] Folgende wesentliche Maßstäbe sind dabei einzuhalten: „*Zentral steht, technisch mittels anerkannten Methoden gesichert festzustellen, welche Daten bestehen, wann und von wem sie wie erstellt wurden, ob und von wem sie zwischenzeitlich verändert wurden, in welchem tatsächlichen und technischen Kontext sie stehen, welchen Inhalt sie aufweisen und ob dieser eindeutig ist. … Dabei muss der Computerforensiker objektiv, transparent und wertfrei vorgehen sowie unabhängig sein. Das gilt im Strafprozess und bei der unternehmensinternen Untersuchung gleichermaßen.*“[6]

1 Es wird die Mehrzahl verwendet, da es mehrere Arten von Bias gibt.
2 Large Language Model.
3 GIROD-FRAIS, KAPPLER: „Rudolf Archibald Reiss und der Stellenwert der Forensik“ in SIAK-Journal - Zeitschrift für Polizeiwissenschaft und polizeiliche Praxis (1/2021), S. 4-18.
4 DEWALD: Formalisierung Digitaler Spuren und ihre Einbettung in die Forensische Informatik, Dissertation, Universität Erlangen-Nürnberg, 2012.
5 HRDINKA: „Herausforderungen verantwortungsloser Digitalisierung“, Tagungsband des 23. Internationalen Rechtsinformatik Symposions IRIS 2020, S 503.
6 NEUFANG, BRECHTKEN, LICHTENTHÄLER, KELLER, KRANEBITTER: Computerforensische Untersuchungen – Rahmenbedingungen, Methoden, Technologien, in SOYER, Handbuch Unternehmensstrafrecht, 15. Kapitel, Manz, 2020.

Im Folgenden werden die Möglichkeiten von Bias im Zuge der Ermittlungen analysiert, und danach der Einsatz von KI zur Eindämmung von Bias untersucht.

## 2. Historie: Der Daubert Standard

Der Oberste Gerichtshof der USA hat in mindestens drei Fällen geurteilt, dass die Zulässigkeit von Sachverständigengutachten bei Bundesgerichten bestimmten Kriterien standhalten muss. Diese werden weltweit verstärkt insb. in Kanada angewandt. Demzufolge wurde erstmals 1923 mit dem Frye-Test[7] solch ein Standard gesetzt, 1975 dann mit den Federal Rules of Evidence (FRE), und schließlich wurde mit dem Daubert Standard[8] die Zulässigkeit eines Sachverständigenbeweises in den USA abschließend definiert. Dieser ist in den Bundesbeweisregeln[9] (Regel 702 - Aussage von Sachverständigen) übersetzt definiert als:[10]

„*Ein Zeuge, der aufgrund seiner Kenntnisse, Fähigkeiten, Erfahrung, Ausbildung oder Ausbildung als Sachverständiger qualifiziert ist, kann in Form eines Gutachtens oder auf andere Weise aussagen, wenn der Antragsteller dem Gericht nachweist, dass es höchstwahrscheinlich ist, dass:*

- *(a) Die wissenschaftlichen, technischen oder sonstigen Fachkenntnisse des Sachverständigen helfen dem Sachverständigengericht[11] Beweise zu verstehen oder einen strittigen Sachverhalt festzustellen;*
- *(b) die Aussage auf ausreichenden Fakten oder Daten beruht;*
- *(c) die Aussage das Ergebnis verlässlicher Grundsätze und Methoden ist; Und*
- *(d) das Gutachten des Sachverständigen eine zuverlässige Anwendung der Grundsätze und Methoden auf den Sachverhalt des Falles widerspiegelt.*“

Diese Regeln bedeuten, dass eine empirische Überprüfbarkeit essenziell ist, d.h. ob sich die verwendete Methode verifizieren oder falsifizieren lässt. Weiters sollte die eingesetzte Methode in einer Fachzeitschrift veröffentlicht und dabei einem Peer-Review unterzogen worden sein. Auch wird geprüft, ob es Unsicherheiten bei dieser Methode gibt und diese Methode in der Wissenschaft allgemein anerkannt ist.

Der Oberste Gerichtshof der USA versuchte damit die Bestellung von Experten für sog. „*Junk-Wissenschaften*“ einzudämmen.[12] Die in verschiedenen forensischen Labors entdeckte Unzulänglichkeiten machten deutlich, dass die forensischen Wissenschaften nicht unfehlbar sind. In zahlreichen Fällen unrechtmäßiger Verurteilungen wurden mit alarmierender Häufigkeit kriminaltechnische Fehler entdeckt.

In der Daktyloskopie wurde zu diesem Zweck vom kanadischen Polizeibeamten ASHBAUGH[13, 14] die ACE-V[15] Methode entwickelt. Dabei handelt es sich um ein wissenschaftliches Verfahren zur Identifizierung latenter Fingerabdrücke, das Analyse, Vergleich, Bewertung und Verifizierung umfasst. Bei der Analyse wird die qualitative und quantitative Bewertung von Details, Anteil, Wechselbeziehung und ein Wert für die Individualisierung bestimmt. Während des Vergleichs untersucht der Prüfer die bei der Analyse festgestellten Attribute auf Unterschiede und Übereinstimmungen zum Kandidaten. Dabei wird festgestellt, ob Spuren

7 Frye vs. U.S. (Frye), 293 F. 1013, 1014 (D.C. Cir. 1923).
8 Daubert vs. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993).
9 United States Code, 2006 Edition, Supplement 4, Title 28 - JUDICIARY AND JUDICIAL PROCEDURE, i.d.g.F.
10 Office of the Law Revision Counsel: https://uscode.house.gov/view.xhtml?req=granuleid:USC-prelim-title28a-node230-article7-rule702&num=0&edition=prelim, aufgerufen am 10.10.2024.
11 Das können je nach Verfahrensart Richter, die Jury (Geschworene), Schiedsrichter oder Kommissionen sein, die auf Grundlage der Beweise beauftragt sind, im Verfahren sachliche Feststellungen zu treffen. Auch Mediatoren können, allerdings nicht bindend, damit beauftragt sein. Das Sachverständigengericht muss die Beweise abwägen, um festzustellen, ob eine bestimmte Tatsache vorliegt.
12 RISINGER, SAKS, THOMPSON, ROSENTHAL: „The Daubert/Kumho implications of observer effects in forensic science: Hidden problems of expectation and suggestion“, California Law Review, 2002, Vol. 90, S. 1–56.
13 ASHBAUGH: „The key to fingerprintidentification“, FingerprintWhorld, 10(40), S. 94–96, 1985.
14 ASHBAUGH: „Quantitative-qualitative friction ridge analysis: An introduction to basic and advanced ridgeology“, CRC Press, 1999.
15 Analysis, Comparison, Evaluation and Verification: Wissenschaftliche Methode zur Untersuchung und Dokumentation von latenten Fingerabdrücken.

aus derselben oder verschiedener Quelle stammen, oder unschlüssig sind. Bei der Verifizierung handelt es sich somit um eine unabhängige Analyse und Bewertung durch mehrere qualifizierte Prüfer. ACE-V ist eine wissenschaftlich anerkannte, und inzwischen als NIST Standard[16] publizierte Methode, welche den Daubert Kriterien genügt, und Bias *per se* vermeiden hilft.

Ein berühmtes Fallbeispiel für Bias unter außer Achtlassung der ACE-V Methode ist die fehlerhafte Zuordnung einer Fingerspur bei den Bombenanschlägen[17] auf Züge in Madrid im Jahre 2004. Die Interpol Ermittlungen führten auch zu Untersuchungen in den USA: Eine automationsunterstützte Suche in der IAFIS[18] Datenbank beim FBI lieferte eine große Ähnlichkeit mit den Fingerabdrücken von Brandon MAYFIELD, einem zum Islam konvertierten Anwalt in den USA. Aufgrund dessen, und im Kontext der vor wenigen Jahren erfolgten 9/11 Anschläge, ignorierten die Ermittler dabei die Tatsache, dass nur einige wenige Fingerspur-Charakteristika tatsächlich übereinstimmten, und übernahmen unreflektiert das Ergebnis der Datenbank. Weiters wurde das wissenschaftliche ACE-V Prinzip ignoriert, wo zuerst die Analyse, der Vergleich, dann eine Auswertung und letztlich die Verifikation erfolgen muss. Nach der Identifizierung des tatsächlichen Täters, einem Algerier durch die spanischen Behörden, wurde der Beschuldigte MAYFIELD nach Monaten aus der Untersuchungshaft entlassen.

## 3. Bias

Einer der Ursachen für Fehler aus den zahlreichen Bias ist der „Cognitive Bias“,[19] welcher die systematische fehlerhafte Neigung beim Wahrnehmen, Erinnern, Denken und Urteilen bedeutet. Dies geschieht oft unbewusst und basiert auf kognitiven Vorurteilen.[20] Quellen dafür können sein:[21]

- Erwartungen von Ergebnissen vor der Untersuchung,
- Ergebnisse wissensbasierender Technologien w.z.B. Forensic-Toolkits,
- Rückwärtsvergleiche: Erwartungshaltungen bei Treffern, wobei die Spuren fehlen,
- Kontextbasierende Informationen, wenn irrelevante Daten in die Bewertung einfließen.

Der Einsatz innovativer, wissensbasierender Technologien aber auch der sog. „Künstlichen Intelligenz“ mag als Werkzeug den Aufwand reduzieren, welcher für die Analysen des vielfach massenhaft vorliegenden elektronischen Datenmaterials erforderlich ist. Solche Technologien unterstützen Forensiker dabei, Ergebnisse in kurzer Zeit zu erhalten, welche diese bei manueller Sichtung niemals erhalten hätten. Aufgrund der Unsicherheiten und Unausgereiftheit solcherart KI-Systeme sollten sich Forensiker jedoch nicht von solchen automatischen Ergebnissen hinreißen lassen, und voreilige und womöglich falsche Schlussfolgerungen tätigen.

### 3.1. Konsequenzen aus den Bias

Die Folgen aus den vorher beschriebenen Problematiken sind Erwartungshaltungen Dritter, wenn diese Druck auf Forensiker ausüben, die zur Überzeugung führen können, dass forensische Spuren existieren müssen, um eine Verurteilung aussprechen zu können. Solcherart Erwartungen resultieren u.a. auch aus diversen

16 NIST: Organization of Scientific Area Committees (OSAC) for Forensic Science: OSAC Standard Framework for Developing Discipline Specific Methodology for ACE-V, 2020.

17 STACEY: „A Report on the Erroneous Fingerprint Individualization in the Madrid Train Bombing Case“, Journal of Forensic Investigation, S. 706, FBI, 2004.

18 Integrated Automated Fingerprint Identification System des FBI.

19 Vgl. KAHNEMAN, TVERSKY: „Subjective probability: A judgment of representativeness". Cognitive Psychology, p. 430-454, Academic Press, 1972: „*The biasing effects of representativeness are not limited to naive subjects. They are also found … in the intuitive judgments of sophisticated psychologists. Statistical significance is commonly viewed as the representation of scientific truth.*“

20 Vgl. Oxford English Dictionary: „*To exert an influence on (a person or thing), often unduly or unfairly; esp. to cause to become partial or biased; to prejudice*“, https://www.oed.com, aufgerufen am 10.10.2024.

21 Vgl. CZEBE, KOVÁCS: „The impact of bias in latent fingerprint identification“, 2015 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom), 2015, S. 569-574.

TV-Serien, wo sich Ermittler mit wenigen Mausklicks in Überwachungsvideos hinein zoomen, um Details zu extrahieren, die aufgrund der Kameraauflösung gar nicht existieren können. Auch mutet es eher als Magie an, wenn sich TV-Ermittler in wenigen Sekunden in Computersysteme hacken können. Solcherart Darstellungen haben absolut nichts mit der Realität zu tun, vielmehr sind Bilduntersuchungen oder das Brechen von Sicherheitsmechanismen ein aufwändiger Vorgang, wovon sich Organe im Zuge eines Strafverfahrens bewusst sein sollen. Schließlich sollen Forensiker bei der Bewertung der Ergebnisse nicht nur ihre Erfahrung, Sorgfalt und Objektivität, sondern auch den „gesunden Hausverstand" walten lassen.

### 3.2. Beispielfall: Verurteilung aufgrund Bias

Dieses Beispiel zeigt die Problematik auf, als ein wg. § 207a[22] StGB belangter Journalist als Verdächtiger gutgläubig mit den Ermittlungsbehörden kooperierte, er sich dadurch selbst belastete. Der in den USA automatisch erstellte CyberTipline Report[23] eines Social Media Providers gab den Ausschlag für die Ermittlungen, da kinderpornographisches Material auf dem Account des Verdächtigen gefunden wurde.

Der Angeklagte, der übrigens sein Mobiltelefon freiwillig den Ermittlungsbehörden öffnete, da er sich keiner Schuld bewusst war, rechtfertigte sich indem, da er seinen Social-Media-Account schon lange nicht mehr nutzte, ihm daher das inkriminierte Material untergeschoben werden musste.

Das Ergebnis der forensischen Auswertung des bestellten Gerichtssachverständigen über das sichergestellte Mobiltelefon ergab, dass „*keine Bild- oder Video-Dateien mit kinderpornographischen Inhalten auf dem untersuchten Smartphone oder der eingelegten Micro SD-Karte gefunden wurden*", und „*ebenfalls wurden keine Hinweise auf den Versand solcher Dateien gefunden.*" Wie konnte es also zur Verurteilung kommen?

Der CyberTipline Report enthielt nicht nur die gmx.at E-Mailadresse als Zugang zum Account des Angeklagten, sondern auch eine weitere mit sehr ähnlich lautender Domäne aus Österreich. Der Gerichtssachverständige ließ sich in der Hauptverhandlung zur Aussage hinreißen, dass dies ein „Tippfehler" (Anm.: im automatisch erstellten Report) sein musste, und diese E-Mail dem Provider GMX zuordenbar sein musste, und sohin dem Angeklagten. Die Auskunft bei GMX ergab „*Zu Ihrer oben stehenden Anfrage liegen uns keine Bestandsdaten vor. Der Account ist nicht oder nicht mehr existent*" was schlichtweg falsch war; denn weil die ähnlich lautende Domäne gar nicht im Besitz von GMX war und ist, hätte die Antwort richtigerweise „*diese Domäne ist nicht in unserer Verfügungsgewalt*" lauten müssen, was inhaltlich einen wesentlichen Unterschied macht. Somit wurde unter diesen Gesichtspunkten dem Gericht fälschlich glaubhaft gemacht, es müsse sich um eine GMX Adresse handeln, die aber nicht weiter verfolgbar ist. Eine sorgfältige Untersuchung, wem diese E-Mail mit dieser Domäne zuordenbar war, unterblieb in Folge (Anm.: diese Domäne war und ist nach wie vor einem Dritten zuordenbar, welcher auch als Täter in Frage kam). Weiters sagte der SV in der Hauptverhandlung aus, dass er „*Einträge mit einschlägigen Begriffen*" fand und demnach „*konnte also geschlussfolgert werden, dass die aufgeführten Begriffe zumindest ein Mal vom Beschuldigten eingegeben wurden.*" (Anm: der Angeklagte ist Journalist und suchte in Google nach verschiedensten Begriffen).

Die IP-Adressen mit welchen das inkriminierte Material hochgeladen wurde, stammen alle aus Übersee, lediglich eine stammt vom selben österreichischen Provider wie jene des Angeklagten. Diese österreichischen IP-Adressen gehörten jedoch vollkommen disjunkten IP-Blöcken an, sodass eine Zuordnung zum Angeklagten eigentlich nicht möglich war.

Das Gericht schenkte dem Angeklagten keinen Glauben und verurteilte ihn trotz des mangelhaft erhobenen Beweissubstrats und falschen Schlussfolgerungen.

Im Berufungsverfahren wg. Schuld hob das zuständige OLG das angefochtene Urteil auf, und verwies die Sache zu neuer Verhandlung und Entscheidung an das Erstgericht zurück. Als Entscheidungsgrund wurde

22 § 207a StGB, Strafgesetzbuch, BGBl. 60/1974, i.d.g.F.: Bildliches sexualbezogenes Kindesmissbrauchsmaterial und bildliche sexualbezogene Darstellungen minderjähriger Personen.

23 „CyberTipline Report" gegen Kinderpornographie des NCMEC, National Center for Missing & Exploited Children.

u.a. angegeben, dass „*gerade bei (bloßen) Indizienbeweisen[24] besondere Vorsicht geboten ist.[25] Daher fällt im konkreten Fall besonders ins Gewicht, dass derzeit noch nicht alle möglichen Beweisquellen erschöpft sind. So sind die bislang gepflogenen Erhebungen zur zweiten hier relevanten E-Mail-Adresse „a***@b***.at“ nicht vollständig, ..., obwohl die Domain „b***.at“ laut dem (zutreffenden) Befund des Privatsachverständigen der in Wien ansässigen b*** GmbH zugeordnet ist. Eine Anfrage an dieses Unternehmen erfolgte bislang nicht. Warum der Sachverständige bei der genannten E-Mail-Adresse wie selbstverständlich von einem Tippfehler ausgeht, leuchtet – selbst bei Berücksichtigung der großen Verbreitung des E-Mail-Dienstes „gmx“ – nicht ein. ... Weiters wird der gerichtlich bestellte Sachverständige zu befragen sein, inwieweit der vom Privatsachversändigen erhobene Befund[26] allenfalls andere Schlussfolgerungen zulassen könnte.*“

Das Erstgericht entschied nach kurzer Verhandlung einen klaren Freispruch. Dieses Beispiel verdeutlicht, wenn Beteiligte im Strafverfahren „biased“ sind, sehr hohe Aufwände und Kosten für den Verurteilten aufgrund der Richtigstellung im Rechtsmittelverfahren entstehen. Die Unterstützung von KI-Systemen zur frühzeitigen Aufdeckung solcher Fehler könnte dabei wertvolle Dienste bieten.

## 4. KI im Sachverständigenbeweis und bei der Ermittlung

Im Bereich des IT-Strafrechts ist es aufgrund der über die Jahre stetig steigenden Datenmengen gegeben, dass die Aufwände zur Beweissicherung und Analyse der Daten analog dazu – trotz der gestiegenen Rechenleistungen – ansteigt. Der Einsatz entsprechender forensischer Werkzeuge ist daher geboten. Diese Technologien unterstützen Forensiker dabei, Ergebnisse in kurzer Zeit zu erhalten, welche sie bei manueller Sichtung niemals erhalten hätten. Allerdings sollten sich sie nicht von solchen Ergebnissen hinreißen lassen, und voreilige und womöglich falsche Schlussfolgerungen tätigen.

Der schon hier behandelte Begriff Cognitive Bias als bekanntes Phänomen ist die systematische fehlerhafte Neigung beim Wahrnehmen, Erinnern, Denken und Urteilen. Die in jüngster Zeit aufgekommenen KI-Systeme könnten daher eine innovative Methode darstellen, Sachverständige zu unterstützen, einerseits rasch zu Ergebnissen zu kommen, und andererseits mit Hilfe dieser künstlichen Intelligenz einem Bias zu entgehen, und so die Qualität der Befundaufnahmen und Gutachten zu erhöhen. Auch könnte diese Technologie helfen die aufgezeigten Problematiken bei der Ermittlungsarbeit zu verbessern.

Im Beitrag[27] zur IRIS 2024 wurde diese Möglichkeit KI sinnvoll in der Gutachtensarbeit einzusetzen untersucht, und dabei wurden die Grenzen von LLMs anhand von Beispielen aufgezeigt. Gezeigt wurde, dass die Generierung technischer Anforderungen, wie bspw. einem Testprogramm oder einem Datenmodell mit Testdaten nahezu perfekte Ergebnisse liefert, welche der Sachverständigenarbeit zeitsparend zugute kommen. Hingegen konnte gezeigt werden, wenn Fragestellungen im interdisziplinären Bereich gestellt werden, wie das Ziehen rechtsrelevanter Schlüsse, LLMs mit falschen Ergebnissen vollkommen versagen, was als „Halluzinantion“ bezeichnet wird.

DAHL ET AL[28] untersuchten die rechtlichen Halluzinationen in LLMs, und sie stellten die Frage, ob KI-Systeme wie ChatGPT die Rechtspraxis neu gestalten oder demokratisieren können. Dabei wurde festgestellt, dass LLMs nicht nur stark halluzinieren, sondern es ihren aktuellen Implementierungen auch an bestimmten Verhaltensweisen fehlt, sodass die Nutzer welche am meisten profitieren könnten, wie Bedürftige, solche LLMs

24 Dazu LENDL in FUCHS/RATZ, WK StPO § 258 Rz. 24.
25 FABRIZY/KIRCHBACHER, Kommentar: Das Strafverfahren und seine Grundlagen, StPO § 258 Rz. 3.
26 Dazu FABRIZY/KIRCHBACHER aaO § 126 Rz. 8.
27 Sachverständigenbeweis im Strafverfahren mit KI? Tagungsband des 27. Internationalen Rechtsinformatik Symposions IRIS 2024.
28 DAHL ET AL: Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models, arXiv:2401.01301 [cs.CL], 2024.

sicher nutzten könnten: Benutzer sollten korrigiert werden, wenn sie fehlgeleitete Fragen stellen, bzw. sollten die Antworten moderiert werden anstatt, dass das LLM „*vor Überzeugung halluziniert*."

Als Ergebnis kann gefolgert werden, dass LLMs die auf allgemeinen, nicht speziell trainierten Datenbasen aufbauen, sich nicht für den Einsatz im interdisziplinären Bereich eignen, da sie stark zu Halluzinationen tendieren. Ein Einsatz in der gerichtlichen Praxis bzw. Sachverständigentätigkeit ist daher gegenwärtig nur sehr eingeschränkt zweckmäßig.

Im Folgenden werden Bias und internationale Entwicklungen i.V.m. KI erhoben:

## 4.1. Norm ISO/IEC TR 24027:2021[29]

Diese Norm beschäftigt sich mit Bias in KI-Systemen, welche für verschiedene Aufgaben eingesetzt werden. Solcherart Systeme lernen mit realen Daten, und daher können sie bereits auf verzerrten (biased) Daten aufbauen, und Bias noch verstärken. Dabei können unerwünschte Bias unbeabsichtigt in ein KI-System einfließen, und je automatisierter das System ist, und je weniger menschliche Aufsicht erfolgt, umso größer ist die Wahrscheinlichkeit unbeabsichtigter Folgen.

Diese Norm definiert verschiedene Bias – die wiederum selbst Quellen unerwünschter Bias sein können – wie folgt:

- Automation Bias: Neigung des Menschen, Vorschläge automatisierter Entscheidungssysteme zu bevorzugen und widersprüchliche Informationen, die ohne Automatisierung erstellt wurden, zu ignorieren, selbst wenn sie korrekt sind,
- Bias: systematischer Unterschied in der Behandlung bestimmter Objekte, Personen oder Gruppen im Vergleich zu anderen,
- Menschlicher kognitiver Bias: Voreingenommenheit, die auftritt, wenn Menschen Informationen verarbeiten und interpretieren,
- Bestätigungsbias: Eine Art menschlicher kognitiver Bias, die Vorhersagen von KI-Systemen begünstigt, die bereits bestehende Überzeugungen oder Hypothesen bestätigen,
- Datenbias: Dateneigenschaften, die, wenn sie nicht berücksichtigt werden, dazu führen, dass KI-Systeme für verschiedene Gruppen eine bessere oder schlechtere Leistung erbringen, und
- Statistischer Bias: Art des konsistenten numerischen Versatzes in einer Schätzung relativ zum wahren zugrunde liegenden Wert, der den meisten Schätzungen innewohnt.

Darüber hinaus werden in der Norm folgende Begriffe definiert:

- Praktische Probe: Stichprobe von Daten, die daher ausgewählt werde, da sie leicht zu beschaffen sind und nicht weil sie repräsentativ ist, und
- Gruppe: Teilmenge von Objekten in einer Domäne, die verknüpft sind, da sie gemeinsame Merkmale aufweisen.

LLMs weisen die Eigenschaft auf, dass sie vollautomatisiert sind und keiner menschlichen Aufsicht unterliegen. Aus diesem Grund ist die Wahrscheinlichkeit unbeabsichtigter Folgen bei deren Einsatz potenziell sehr hoch. Weiters ist es nicht ausgeschlossen, dass LLMs selbst auf Datenbias aufbauen.

29 ISO/IEC TR 24027:2021 – Information technology – Artificial intelligence (AI) – Bias in AI systems and AI aided decision making, International Organization for Standardization, Geneva, 2021.

Wenn ein Sachverständiger einem menschlichen Bias unterliegt und diesen mit bestimmten Datenmerkmalen in eine Analyse einfließen lässt, so führt das in Folge zu weiteren Bias, die natürlich unerwünscht sind. Schließlich kann er in Folge des Einsatzes eines LLM einem Automation- oder Bestätigungsbias unterliegen. Die folgende Grafik verdeutlicht diese Problematik:

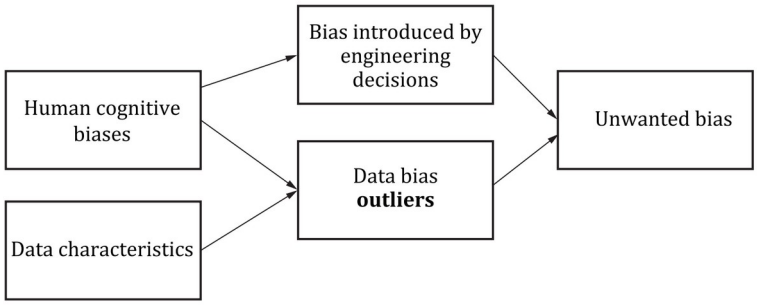


**Abbildung 1: Beispiel für Datenmerkmale, die sich als unerwünschter Bias manifestieren (Quelle: ISO)**

Folglich kann eine Kette von Bias entstehen, die gerade in der Sachverständigentätigkeit keinesfalls akzeptiert werden kann.

## 4.2. Internationale Entwicklungen

### 4.2.1. USA

Dieser Abschnitt beschäftigt sich mit der Frage des Einsatzes von KI bei Gericht in den USA, bzw. ob es Regulierungsbestrebungen dafür gibt. An dieser Stelle soll der im Mai 2023 bekannt gewordene Fall eines Anwalts[30] angeführt werden, der die Problematik des Einsatzes von KI in nicht technischen Disziplinen verdeutlicht: Der Anwalt verwendete in seiner Klage generierte Referenzfälle von ChatGPT, wobei diese zitierte Entscheidungen vom Gericht nicht gefunden werden, da diese von ChatGPT frei erfunden waren. Ob solch ein Einsatz einer KI im Bundesstaat New York verpönt ist oder nicht, ist nicht ergründbar, da die Verhaltensregeln[31] der NYSBA[32] darüber schweigen. Diesbezüglich bestimmt Regel 4.1, wonach der Anwalt zur Ehrlichkeit verpflichtet ist, wobei sich diese Regel auf Tatsachenbehauptungen bezieht.

In den USA gibt es nur wenige Vorschriften für technologiegestützte Rechtsdienstleistungen wie „LegalTech". SIMSHAW[33] argumentiert, dass sich ohne Regulierung nur große Anwaltskanzleien und ihre wohlhabenden Kunden Legal Tech leisten werden können, wodurch sich die „*Kluft zwischen Besitzenden und Besitzlosen vergrößern wird*". Er betont weiter die Notwendigkeit einer nationalen Aufsichtsbehörde wie sie die ABA[34] wahrnehmen könnte. Diese hätte jedoch keine Durchsetzungsbefugnisse, sondern würde stattdessen Regeln verwalten, die von den obersten Gerichten der Bundesstaaten akzeptiert werden müssten. Kritiker in den USA bemängeln hingegen, dass eine Umstellung auf eine nationale Regulierung für Dienstleistungen in der Rechtsbranche unangemessen sei.

Diese Diskussion ist in den USA im Gange, wobei auch die ABA dazu Stellung bezog: „*Das Problem mit der generativen künstlichen Intelligenz ist bekannt: Wenn man sich beim Verfassen eines Schriftsatzes auf sie*

30 The Irish Times: https://www.irishtimes.com/world/us/2023/05/28/i-apologize-for-the-confusion-earlier-heres-what-happens-when-your-lawyer-uses-chatgpt/, aufgerufen am 10.10.2024.
31 New York Rules of Professional Conduct, NYSBA: https://nysba.org/app/uploads/2024/02/20240226-Rules-of-Professional-Conduct-as-amended-6.10.2022.pdf, aufgerufen am 10.10.2024.
32 New York State Bar Association (Anwaltskammer des Staates New York).
33 SIMSHAW: Toward National Regulation Of Legal Technology: A Path Forward For Access To Justice, 92 Fordham L. Rev 1, 2023.
34 American Bar Association (Amerikanische Anwaltskammer).

*verlässt, könnte sie falsche Zitate ausspucken. Anwälte aufgepasst. Sogenannte „Halluzinationen" von KI-Programmen wie ChatGPT sind sehr real*"[35] und schlägt bessere Datenbanken vor. Dazu hat die ABA auch eine „Task Force on Law and Artificial Intelligence"[36] eingesetzt, um die rechtlichen Herausforderungen der KI zu bewältigen, wie die Auswirkungen von KI auf den Anwaltsberuf, der Rechtspraxis und der damit verbundenen ethischen Implikationen, Einblicke in die vertrauenswürdige und verantwortungsvolle Entwicklung und Nutzung von KI, und Möglichkeiten zur Bewältigung von KI-Risiken zu identifizieren. Sie stellt weiters fest, dass „*Deepfakes und KI gewaltige Herausforderungen für die Authentifizierung und Verwendung von Beweismitteln vor Gericht darstellen, beispielsweise für die Feststellung der Zuverlässigkeit und Integrität von Beweismitteln.*" I.d.Z. wird auf die Bundesbeweisregeln, insb. dem Daubert Standard verwiesen, welche allerdings nur bei bestimmten Verfahren und Gerichten[37] gelten.

### 4.2.2. EU

Gem. der am 01.08.2024 in Kraft getretenen EU Verordnung[38] (KI-VO) bezeichnet ein KI-System *„ein maschinengestütztes System, das für einen in unterschiedlichem Grade autonomen Betrieb ausgelegt ist und das nach seiner Betriebsaufnahme anpassungsfähig sein kann und das aus den erhaltenen Eingaben für explizite oder implizite Ziele ableitet, wie Ausgaben wie etwa Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen erstellt werden, die physische oder virtuelle Umgebungen beeinflussen können*".

Der Einsatz von KI in der Strafverfolgung stellt lt. Art. 6 Abs. 2 KI-VO ein hohes Risiko dar, und daher sollen solche Hochrisiko-KI-Systeme zukünftig in einer EU-Datenbank registriert werden. Zudem sind umfassende Dokumentations-, Transparenz- und menschliche Aufsichtspflichten vorgesehen. Art. 14 sieht für Hochrisiko-KI-Systeme eine mittels Mensch-Maschine-Schnittstelle wirksame menschliche Beaufsichtigung vor, die der Verhinderung oder Minimierung der Risiken für Gesundheit, Sicherheit oder Grundrechte dient. Die natürlichen Personen, denen die menschliche Aufsicht übertragen wurde, müssen u.a. angemessen und verhältnismäßig in der Lage sein, „*sich einer möglichen Neigung zu einem automatischen oder übermäßigen Vertrauen in die von einem Hochrisiko-KI-System hervorgebrachte Ausgabe („Automatisierungsbias") bewusst zu bleiben, insbesondere wenn Hochrisiko-KI-Systeme Informationen oder Empfehlungen ausgeben, auf deren Grundlage natürliche Personen Entscheidungen treffen.*" Damit weist die KI-VO – unter Bedachtnahme auf die ISO/IEC TR 24027:2021 – auf die Neigung des Menschen hin, Vorschläge automatisierter Entscheidungssysteme zu bevorzugen und widersprüchliche Informationen, die ohne Automatisierung erstellt wurden, zu ignorieren, selbst wenn sie korrekt sind.

Generell verboten werden zudem Hochrisiko-KI-Systeme zwecks biometrischer Kategorisierung oder Fernidentifizierung, wobei eine Ausnahme dafür nur unter bestimmten Voraussetzungen und Auflagen möglich sein wird.

Bemerkenswert i.Z.m. der Sachverständigentätigkeit ist die Definition eines Hochrisiko-KI-Systems in Anhang III Z. 6 „*KI-Systeme, die bestimmungsgemäß von Strafverfolgungsbehörden oder in deren Namen oder von Organen, Einrichtungen und sonstigen Stellen der Union zur Unterstützung von Strafverfolgungsbehörden zur Bewertung der Verlässlichkeit von Beweismitteln im Zuge der Ermittlung oder Verfolgung von Straftaten verwendet werden sollen*". Dies erlaubt den Einsatz von KI-Systemen u.a. zur Bias-Erkennung unter

35 American Bar Association: https://www.americanbar.org/news/abanews/aba-news-archives/2024/03/fixing-ai-requires-better-databases/, aufgerufen am 10.10.2024.

36 https://www.americanbar.org/groups/centers_commissions/center-for-innovation/artificial-intelligence/, aufgerufen am 17.12.2024.

37 Regel 1101 – Anwendbarkeit der Regeln. Bspw. Bezirksgerichte, Bundesgericht, Zivilverfahren, Strafverfahren, nicht jedoch für Verfahren vor der Grand Jury.

38 Verordnung (EU) 2024/1689 des Europäischen Parlaments und des Rates vom 13. Juni 2024 zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz und zur Änderung der Verordnungen (EG) Nr. 300/2008, (EU) Nr. 167/2013, (EU) Nr. 168/2013, (EU) 2018/858, (EU) 2018/1139 und (EU) 2019/2144 sowie der Richtlinien 2014/90/EU, (EU) 2016/797 und (EU) 2020/1828 (Verordnung über künstliche Intelligenz) (Text von Bedeutung für den EWR), in ABl. L, 2024/1689, 12.7.2024, ELI: http://data.europa.eu/eli/reg/2024/1689/oj.

den in der KI-VO geforderten Einschränkungen für Hochrisiko-KI-Systeme, und soweit ihr Einsatz in der Strafverfolgung nach einschlägigem Unionsrecht oder nationalem Recht zugelassen ist.

Aufgrund der Tatsache, dass solcherart Praxis in Österreich nicht verboten ist, und grds. der freien Beweiswürdigung des Gerichts unterliegt, wird sich hier ein neues Spannungsfeld aufbauen. Allenfalls wird das auf die österreichische StPO[39] Einfluss nehmen, indem die Zulässigkeit vs. freie Beweiswürdigung von Beweismitteln neu geregelt werden wird müssen.

## 5. Schlussfolgerungen

Besonders wichtig bei der Erhebung von Beweismittel und deren Auswertung mittels KI ist, dass dies ausschließlich eigens geschulten und erfahrenen Experten unter Anwendung des Mehraugenprinzips vorbehalten sein muss, denn ansonsten besteht die Gefahr, dass Beweise u.a. aufgrund von Bias fehlinterpretiert werden, was sich in Folge als nachteilig für Beschuldigte auswirken wird, wenn die freie Beweiswürdigung des Gerichts den fehlerhaften Beweisen oder deren Fehlinterpretation folgt.

Diesbezüglich gehen die EU und die USA vollkommen unterschiedliche Wege, wobei die Union auf eine umfassende präventive Regulierung von Hochrisiko-KI-Systeme und die USA dagegen auf anwendbare und praktikable Ansätze setzen.[40] Welcher der beiden Ansätze sich schließlich international durchsetzen wird, ist heute aufgrund der sich in Fluss befindlichen Diskussion und der zwar in Kraft getretenen, aber erst ab 02.08.2026 für Hochrisiko-KI-Systeme anzuwendende KI-VO noch nicht absehbar.

39 Strafprozessordnung (StPO), BGBl. 631/1975 i.d.g.F.

40 Siehe auch: ABA Task Force on Law and Artificial Intelligence, Addressing the Legal Challenges of AI Year 1 Report on the Impactof AI on the Practice of Law, August 2024, https://www.americanbar.org/content/dam/aba/administrative/center-for-innovation/ai-task-force/2024ai-task-force-report.pdf, aufgerufen am 16.10.2024.