Rechtsökonomie im KI-Zeitalter

Generative KI verschiebt die Kostenstruktur von Kanzleien und Rechtsbteilungen

Simon Jakob, M.A. HSG in Law & Economics, Senior Legal Counsel NZZ

Résumé: L’IA générative transforme le droit, dans un premier temps, moins par la création de nouvelles normes juridiques que par une modification de sa structure de coûts. Elle effectue les étapes clés du travail juridique plus rapidement et avec une meilleure qualité qu’auparavant, réduisant ainsi considérablement leurs coûts.
L’IA ne transforme donc pas «le droit» en produit de base, mais bien une part croissante des étapes de production qui permettent de rechercher, d’analyser, de formuler et de mobiliser le droit. Cette évolution concerne l’ensemble du corps des avocats de la même manière – tant les cabinets d’avocats que les services juridiques d’entreprise. La réduction des coûts peut élargir l’accès à l’information juridique et la mobilisation du droit. Mais elle peut aussi générer des requêtes supplémentaires et de qualité plus hétérogène, dont le traitement par les tribunaux continue de mobiliser des ressources humaines considérables. Aucun choc quantitatif causal lié à l’IA n’a encore été démontré en Suisse; toutefois, en 2026, le Tribunal fédéral s’est déjà expressément penché sur un recours généré en grande partie par l’IA.

Zusammenfassung: Generative KI verändert das Recht zunächst weniger durch neue Rechtsnormen als durch eine Verschiebung

seiner Kostenstruktur. Sie erledigt zentrale juristische Arbeitsschritte schneller und in höherer Qualität als bisher und senkt damit deren Kosten erheblich.
KI macht damit nicht «das Recht» zur Commodity, wohl aber einen wachsenden Teil jener Produktionsschritte, mit denen Recht recherchiert, analysiert, formuliert und mobilisiert wird. Diese Entwicklung betrifft die gesamte Anwaltschaft gleichermassen – die Kanzlei ebenso wie die Rechtsabteilung. Die Kostensenkung kann Rechtsinformation und Rechtsmobilisierung verbreitern. Sie kann aber zugleich zusätzliche und qualitativ heterogenere Eingaben erzeugen, deren Verarbeitung bei Gerichten weiterhin erhebliche menschliche Ressourcen bindet. Ein kausaler KI-bedingter Mengenschock ist für die Schweiz bislang nicht nachgewiesen; das Bundesgericht hat sich 2026 aber bereits ausdrücklich mit einer weitgehend KI-generierten Beschwerde befasst.

Dieses Werk ist lizenziert unter einer Creative Commons Namensnennung – Weitergabe unter gleichen Bedingungen 4.0 International Lizenz.

I. Einleitung: Der Preisfilter des Rechts verliert an Bedeutung

1

Im Juli 2026 erreichte die Diskussion über KI-generierte Rechtsschriften das Bundesgericht in bemerkenswerter Klarheit. Im Verfahren 4A_231/2026 räumte der Beschwerdeführer ein, seine Beschwerde unter Einsatz künstlicher Intelligenz erstellt zu haben, ausdrücklich um Kosten einzusparen. Das Bundesgericht äusserte deutliche Zweifel daran, ob eine

derartige Beschwerde überhaupt zulässig sei, liess diese Grundsatzfrage aber offen. Entscheidend waren schliesslich herkömmliche Mängel der Beschwerdeführung: Der Eingabe wurden unter anderem Erwägungen der Vorinstanz zugrunde gelegt, die dort nicht vorhanden waren; Tatsachen wurden ohne hinreichende Aktenbezüge ergänzt und die tatsächlichen Entscheidgründe der Vorinstanz teilweise verfehlt. Das Gericht behandelte damit einen technologisch neuen Entstehungsprozess letztlich mit weitgehend vertrauten prozessrechtlichen Kategorien (BGer 4A_231/2026 vom 3. Juli 2026).

2

Auch aus der kantonalen Praxis kommen entsprechende Signale: So wird in Berichten von NZZ und swissinfo geschildert, dass Parteien vermehrt KI einsetzten und selbst bei Verfahren mit geringem Streitwert plötzlich umfangreichere Rechtsschriften eingereicht würden. Solche Beobachtungen sind praktisch bedeutsam, dürfen aber nicht mit einem Kausalnachweis verwechselt werden. Weder für Basel noch für die Schweiz insgesamt ist bislang empirisch gezeigt, dass generative KI die Zahl der Verfahren erhöht hat, vgl. dazu Beiträge von NZZ und swissinfo.

3

Der tragfähige Ausgangspunkt der Analyse liegt deshalb nicht bei einer behaupteten «KI-Klageflut», sondern eine Stufe früher: bei den Kosten juristischer Arbeit selbst. Generative KI erledigt zahlreiche juristische Arbeitsschritte schneller und in besserer Qualität als bisher (vgl. dazu Beispiele unten Rn. 17 f.). Ökonomisch ist das bedeutsamer als die häufig gestellte Frage, ob ein Sprachmodell «besser als ein Anwalt» sei. Märkte verändern sich nicht erst, wenn eine Maschine die höchste menschliche Qualität erreicht. Es genügt, wenn eine bislang arbeitsintensive Zwischenstufe für einen bestimmten Risikobereich in brauchbarer Qualität zu wesentlich tieferen Grenzkosten hergestellt werden kann – erst recht, wenn die Qualität dabei sogar steigt.

4

Ein materiell begründeter Anspruch kann etwa für seinen Inhaber einen negativen erwarteten Durchsetzungswert aufweisen, wenn Informations-, Beratungs-, Prozess- und Vollstreckungskosten seinen erwarteten realisierbaren Nutzen übersteigen. Sinkt ein Teil dieser Kosten, verändert sich das materielle Recht nicht. Es verändert sich aber die Menge jener Rechtspositionen, deren Prüfung und Mobilisierung wirtschaftlich sinnvoll wird. Die klassische Prozessökonomik zeigt zugleich, dass private Prozessanreize und gesellschaftlich optimale Prozessaktivität auseinanderfallen können, weil Parteien weder sämtliche gesellschaftlichen Kosten noch sämtliche positiven Wirkungen eines Verfahrens miteinberechnen (SHAVELL, 2003).

5

Die These dieses Beitrags ist deshalb bewusst asymmetrisch: KI senkt zunächst private Produktionskosten. Die öffentliche Verarbeitung derselben Rechtsmobilisierung bleibt dagegen wesentlich stärker an knappe menschliche Zeit gebunden. Gerade der Erfolg von Legal AI kann damit einen neuen institutionellen Flaschenhals erzeugen: die Verarbeitungskapazität der Justiz.

II. Juristische Qualität hat einen Preis – und Verifikation ebenfalls

6

Rechtsberatung folgt häufig einem impliziten Ideal maximaler Absicherung: mehr Recherche, mehr Präjudizien, mehr Reviews, mehr Eventualbegründungen. Aus rechtsökonomischer Perspektive ist maximale Genauigkeit jedoch nicht mit optimaler Genauigkeit identisch. KAPLOW analysiert Genauigkeit in gerichtlichen Verfahren als ein Gut, das Nutzen stiftet, aber Ressourcen kostet. KAPLOW und SHAVELL zeigen zudem, dass

private Parteien Anreize besitzen können, mehr Ressourcen in zusätzliche Genauigkeit zu investieren, als gesellschaftlich optimal wäre. Der belastbare Gedanke hinter dem populären 80/20-Prinzip (wonach oft 80 % des Ergebnisses mit 20 % des Aufwands erreicht werden, während die verbleibenden 20 % Präzision überproportional viel Aufwand verschlingen) ist deshalb nicht eine bestimmte Prozentzahl, sondern ein allgemeinerer: Zusätzliche Absicherung ist dann nicht mehr effizient, wenn ihr erwarteter zusätzlicher Nutzen unter den dadurch ausgelösten zusätzlichen Kosten liegt (KAPLOW, 1994; KAPLOW/SHAVELL, 1996).

7

Für KI-gestützte Rechtsarbeit muss dieses Modell allerdings um einen Punkt ergänzt werden. Billige Erstproduktion ist nicht dasselbe wie ein billiges verifiziertes Ergebnis. Eine einfache Heuristik lautet:

$$q^* \in \arg\min_q \{ C_{prod}(q) + C_{verif}(q) + \mathrm{E}[L \mid q] \}$$

8

So liest man die Formel in Worten: Die beste Sorgfaltsstufe q* ist jene, bei der Herstellungskosten, Kontrollkosten und verbleibendes Fehlerrisiko zusammen am geringsten sind. Es ist die Abwägung, die jede Anwältin täglich intuitiv trifft: Wie gründlich muss ich hier arbeiten?

9

Die Bausteine im Einzelnen:

Baustein	**Bedeutung in einfachen Worten**
q	Das gewählte *Sorgfaltsniveau* – ein Regler zwischen «kurz überfliegen» und «jede Fundstelle im Original nachlesen».

Baustein	Bedeutung in einfachen Worten
q^*	Der Stern markiert das *optimale* Sorgfaltsniveau, nach dem die Formel sucht.
$C_{prod}(q)$	*Herstellungskosten* (K = Kosten) für Recherche, Entwurf und Formulierung. «(q)» bedeutet nur: Sie hängen vom Sorgfaltsniveau ab – mehr Sorgfalt kostet mehr.
$C_{verif}(q)$	*Kontrollkosten*: Fundstellen prüfen, Aussagen mit den Akten abgleichen, gegenlesen. Auch sie steigen mit dem Sorgfaltsniveau.
L	Der *Verlust* (englisch *loss*), wenn trotzdem ein Fehler passiert: verlorener Prozess, Haftung, verpasste Frist, Reputationsschaden.
$E[L \mid q]$	Der *erwartete Fehlerverlust*. E steht für Erwartungswert: Schadenshöhe × Eintrittswahrscheinlichkeit. Der Strich «|» wird «bei» gelesen. Beispiel: 2 % Fehlerrisiko × CHF 100’000 Schaden = CHF 2’000. Mehr Sorgfalt senkt diesen Wert.
$\{ \dots \}$	Fasst die drei Posten zu *einer Gesamtrechnung* zusammen.
$\arg\min_q$	«Wähle jenes q, bei dem die Gesamtrechnung *am kleinsten* ist.» Gesucht ist also die Sorgfaltsstufe, nicht der Frankenbetrag.
$\in$	«ist eines von» – es kann mehrere gleich gute Stufen geben.

10

Für jede Sorgfaltsstufe werden die drei Posten addiert. Sie bewegen sich gegenläufig: Mehr Sorgfalt verteuert Herstellung und Kontrolle, senkt aber das Fehlerrisiko. Wer zu wenig investiert, bezahlt über Fehler; wer zu viel investiert, über Aufwand. Optimal ist die Stufe mit der tiefsten Summe. Ein vereinfachtes Zahlenbeispiel (Beträge in CHF, frei gewählt; ein Fehler würde einen Schaden von CHF 100’000 verursachen):

Sorgfaltsstufe (Fehlerrisiko)	**C_{prod} ohne KI → mit KI**	**C_{verif}**	**E[L | q]**	**Summe ohne KI**	**Summe mit KI**
□ Kursorisch (10 %)	1’000 → 200	500	10’000	11’500	10’700
□ Sorgfältig (2 %)	3’000 → 600	1’000	2’000	**6’000**	3’600
□ Umfassend (0,5 %)	9’000 → 1’000	1’800	500	11’300	**3’300**

11

Ohne KI ist Stufe □ optimal; die umfassende Prüfung □ wäre sicherer, aber zu teuer. KI senkt vor allem die Herstellungskosten – und verschiebt das Optimum zu Stufe □. Die bessere Arbeit wird erschwinglich, sofern ein Teil der Ersparnis in Kontrolle reinvestiert wird. Die Verifikation wird damit zum grössten verbleibenden Arbeitsposten (im Beispiel CHF 1’800 von CHF 2’800).

12

Juristisches Beispiel – Option A (Prozess): Für eine Rechtsschrift in einem Mietrechtsstreit erstellt ein KI-System in wenigen Minuten einen gegliederten Entwurf mit fünfzehn zitierten Bundesgerichtsentscheiden. Die Produktionskosten (C_{prod}) sinken von Tagen auf Stunden. Die Verifikationskosten (C_{verif}) sinken nicht im selben Mass: Für jede Fundstelle ist zu prüfen, ob der Entscheid existiert, ob die Erwägung sagt, was

ihr zugeschrieben wird, und ob sie noch aktuell ist; jede Sachverhaltsbehauptung ist mit den Akten abzugleichen. Wie weit diese Kontrolle reicht, bestimmt der erwartete Fehlerverlust (E[L | q]): Bei einer Nebenkostenabrechnung über einige hundert Franken kann die Prüfung der tragenden Präjudizien genügen; droht einem Unternehmen der Verlust seiner Geschäftsräume, ist jede Fundstelle im Original nachzulesen.

Juristisches Beispiel – Option B (Transaktion): Bei einem Unternehmenskauf sind im Datenraum 600 Verträge auf Change-of-Control-Klauseln, Exklusivitäten und Kündigungsrechte zu prüfen. Ein KI-System erstellt in wenigen Stunden einen Red-Flag-Report, für den ein Team früher Wochen brauchte; die Produktionskosten (C_{prod}) brechen ein. Die Verifikation bleibt eine eigenständige Aufgabe – heikel ist weniger, was das System markiert, als was es übersieht. Ihre Intensität lässt sich nach dem erwarteten Fehlerverlust (E[L | q]) staffeln: Die zwanzig wichtigsten Kundenverträge und alle markierten Auffälligkeiten liest ein Mensch vollständig, bei den übrigen Standardverträgen genügt eine Stichprobe. Denn eine übersehene Change-of-Control-Klausel im wichtigsten Kundenvertrag kann den Kaufpreis entwerten, im Wartungsvertrag für die Kaffeemaschinen kaum.

Juristisches Beispiel – Option C (Interne Routineanfrage / Low-Exposure): Für eine interne Orientierung bittet eine Mitarbeiterin um eine erste Übersicht zu den generellen Fristen bei der Rückgabe von Firmenmaterialien nach einer Kündigung. Das KI-System erstellt aus den internen Richtlinien eine kurze Zusammenfassung. Da die Auskunft rein informativ dient, keine externen Rechtsfolgen auslöst und bei Unklarheiten jederzeit nachgefragt werden kann, tendiert der erwartete Fehlerverlust (E[L | q]) gegen null. Auf eine manuelle Verifikation (C_{verif}) wird vollständig verzichtet, weil selbst eine lückenlose Kontrolle teurer wäre als das verbleibende Risiko.

13

«Good enough law» bedeutet damit weder «möglichst billig» noch «ein bisschen ungenau». Gemeint ist eine risikoadäquate Kombination von Produktion und Kontrolle, mit der die

erforderliche Qualitäts- und Sicherheitsstufe bei möglichst geringem Gesamtaufwand erreicht wird. Bei einer alltäglichen, unkomplizierten Vertragsfrage mit überschaubarem Risiko kann ein schneller KI-gestützter Erstcheck mit gezielter menschlicher Verifikation vernünftig sein – oder wie in Option C gezeigt, bei vernachlässigbarem Schadenspotenzial gar komplett auf eine Verifikation verzichtet werden. Bei strafrechtlicher Verantwortlichkeit, irreversiblen Rechtsverlusten, existenziellen regulatorischen Risiken oder sehr hohen Schadenspotenzialen kann dagegen schon eine geringe verbleibende Fehlerwahrscheinlichkeit einen wesentlich höheren Kontrollaufwand rechtfertigen.

14

Daraus folgt die zentrale Einsicht dieses Abschnitts: Verifikation verschwindet nicht komplett; sie wird zum eigenständigen Produktionsschritt. Gute Legal-AI-Systeme senken nicht nur die Kosten der Textproduktion, sondern idealerweise auch die Kosten der Kontrolle, indem sie Quellen zugänglich machen, Aussagen auf Fundstellen zurückführen und Unsicherheit sichtbar machen. Ein System, das einen Entwurf in zehn Sekunden erstellt, aber zwei Stunden Quellenprüfung auslöst, ist ökonomisch weniger revolutionär als ein System, dessen Ergebnis in zehn Minuten zuverlässig verifiziert werden kann.

15

Die eigentliche Transformation lautet deshalb nicht: KI ersetzt menschliche Qualität. Sie lautet: KI verschiebt den Ort, an dem menschliche Aufmerksamkeit ihren höchsten Wert besitzt.

III. Nicht das Recht wird zur Commodity,

sondern Teile seiner Produktion

16

Richard Susskinds Begriff der Commoditisierung juristischer Dienstleistungen bleibt zur Beschreibung dieser Entwicklung nützlich, sofern er eng verstanden wird. Die technische Veränderung macht weder materielles Recht noch richterliche Legitimation, Verantwortung, Verhandlungsmacht oder strategisches Urteil austauschbar. Sie betrifft zunächst jene Tätigkeiten, die bisher als untrennbarer Teil einer teuren professionellen Gesamtleistung erbracht wurden – ob als abrechenbare Stunden in der Kanzlei oder als Arbeitszeit in der Rechtsabteilung.

17

Recherche, erste Issue-Listen, Dokumentenvergleiche, Chronologien, Standardklauseln, Zusammenfassungen und erste Entwürfe lassen sich zunehmend entbündeln und systematisieren. Generative KI beschleunigt diesen Prozess, weil nicht mehr nur Dokumentenverwaltung und Suche, sondern Teile der eigentlichen sprachlichen und analytischen Produktion schneller, besser und zu sehr niedrigen zusätzlichen Kosten erfolgen können.

18

Diese Entwicklung trifft die gesamte Anwaltschaft gleichermassen. Der Associate in der Grosskanzlei, die Einzelanwältin und der Legal Counsel im Unternehmen erledigen im Kern dieselben Arbeitsschritte – und genau diese Schritte werden automatisierbar. Die entscheidende Verschiebung verläuft deshalb nicht zwischen Rechtsabteilung und Kanzlei, etwa im Sinne eines umfassenden Insourcings, sondern quer durch alle juristischen Organisationen: Es kommt zu einer Entbündelung der juristischen Wertschöpfung. Sachverhaltsaufbereitung, erste Recherche, Dokumentenanalyse und Entwürfe werden überall zunehmend maschinell vorbereitet. Und überall verbleiben dieselben knappen Leistungen beim Menschen: schwierige Wertungsfragen, Spezialwissen,

Prozessstrategie, Verhandlungsführung, unabhängige Zweitmeinungen und persönliche Verantwortungsübernahme. Wer juristisch berät, wird künftig weniger für das Erstellen von Texten entlohnt als für die Beurteilung, welche Ausführungen tragfähig sind.

19

Daraus folgt nicht, dass das klassische Kanzleimodell bereits zusammenbricht oder Rechtsabteilungen externe Beratung schlicht ersetzen. Vertretbar ist eine vorsichtigere, aber weiter reichende Aussage:

20

GenAI erzeugt strukturellen Anpassungsdruck auf die pyramidenförmige Produktions- und Abrechnungslogik juristischer Arbeit – in Kanzleien wie in Rechtsabteilungen –, ohne dass daraus zwingend ein Rückgang der Nachfrage nach anwaltlicher Leistung folgen muss.

21

Disruption muss nicht zuerst als Umsatzeinbruch sichtbar werden. Sie zeigt sich vorher im Kleinen: im Verhältnis zwischen Partnern und Associates, in der Frage, wie sich Stundensätze für Arbeit rechtfertigen lassen, die eine Maschine in Minuten vorbereitet, in der Stellenplanung von Rechtsabteilungen, in Make-or-buy-Entscheidungen und in den Ausbildungswegen junger Juristinnen und Juristen, die ihr Handwerk bisher gerade an jenen Routineaufgaben gelernt haben, die nun wegfallen.

IV. Die «1-Klick-Klage»: Rechtsmobilisierung wird billiger, nicht kostenlos

22

Die gleiche Kostenverschiebung, die Anwältinnen und Anwälte in Kanzleien und Rechtsabteilungen bei Verträgen und Rechtsanalysen nutzen, wirkt auch auf die gerichtliche Rechtsmobilisierung. Dabei ist Präzision wichtig. Die Grenzkosten eines weiteren Absatzes juristisch klingenden Textes können technisch gegen null tendieren. Die Grenzkosten einer verifizierten und prozessual wirksamen Eingabe tun dies nicht. Sachverhaltsbeschaffung, Beweismittel, Fristen, formelle Anforderungen, Gerichtskosten, Gegenparteirisiken, Vollstreckung und Kontrolle verschwinden nicht durch ein Sprachmodell.

23

Trotzdem verschiebt sich die Schwelle, ab der sich Rechtsmobilisierung für eine Partei lohnt. Heuristisch lässt sich der private erwartete Durchsetzungswert so darstellen:

$$EV_P = p \cdot R - C_P - (1 - p) \cdot C_D$$

24

So liest man die Formel in Worten: Netto erwarten darf eine Partei, was sie voraussichtlich erhält, abzüglich ihrer eigenen Kosten und abzüglich dessen, was sie bei Unterliegen voraussichtlich an Gericht und Gegenpartei zahlt. Es ist die Prozessrisikoanalyse vor Klageeinleitung – in Kurzschrift:

Baustein	Bedeutung in einfachen Worten
EV_P	Der *private Erwartungswert des Klägers* (englisch *expected value of plaintiff*): Lohnt sich das Verfahren für die Partei, die es führt?
p	*Obsiegenswahrscheinlichkeit* des Klägers ($0 \leq p \leq 1$). Die Gegenwahrscheinlichkeit zu unterliegen beträgt $(1 - p)$.

Baustein	Bedeutung in einfachen Worten
R	Der *realisierbare Streitwert / Anspruch* (englisch *recovery*): was im Erfolgsfall tatsächlich realisiert werden kann.
C_P	*Eigene Prozesskosten des Klägers* (P = plaintiff) für Information, Beratung und Vertretung. Diesen Posten senkt KI am stärksten.
C_D	*Kostenrisiko der Gegenseite / Gerichtskosten* (D = defendant), die dem Kläger bei Unterliegen (1 – p) auferlegt werden.
Minuszeichen	Ertrag minus Kosten. *Positives* Ergebnis: Die Klage lohnt sich. *Negatives* Ergebnis: Verzicht ist wirtschaftlich rational, auch wenn der Anspruch materiell besteht.

25

Juristisches Beispiel: Eine Person hat einen Anspruch R = CHF 2'000 und schätzt ihre Realisierungschance auf p = 0,8. Die Rechnung in drei Schritten:
Schritt 1 – erwarteter Ertrag: p × R = 0,8 × CHF 2'000 = CHF 1'600.
Schritt 2 – eigene Kosten: Die rechtliche Aufbereitung C_P kostet CHF 1'400.
Schritt 3 – Kostenrisiko bei Unterliegen: Bei Unterliegen drohen C_D = CHF 3'000 Gerichtskosten und Parteientschädigung; bei 20 % Verlustrisiko (1 – p = 0,2) beträgt das erwartete Risiko (1 – p) × C_D = CHF 600. Ergebnis: ***EV_P* = 1'600 – 1'400 – 600 = –400 Franken**.

26

Die Durchsetzung ist wirtschaftlich negativ. Sinkt die rechtliche Aufbereitung C_P durch KI – sei es in der digitalen Selbsthilfe, sei es, weil die beauftragte Anwältin das Mandat mit weniger Stunden bearbeitet – auf CHF 600, ändert sich nur der Posten “eigene Kosten” (600 statt 1’400): ***EV_P* = 1’600 – 600 – 600 = +400 Franken.** Derselbe unveränderte materielle Anspruch wird wirtschaftlich mobilisierbar. Das Gesetz hat

sich nicht geändert; geändert hat sich der Preis seines Gebrauchs.

27

SHAVELLS Prozessökonomik zeigt, weshalb daraus nicht ohne Weiteres auf einen gesellschaftlichen Wohlfahrtsgewinn geschlossen werden kann. Eine klagende Partei berücksichtigt typischerweise nicht alle Bearbeitungskosten, die ihr Verfahren bei Gerichten und Gegenparteien auslöst. Umgekehrt kann sie positive Wirkungen wie Abschreckung, Rechtsklärung oder die Durchsetzung systematisch unterdurchgesetzter Rechte ebenfalls nicht vollständig miteinberechnen (SHAVELL, 2003).

28

Gesellschaftlich ist die Rechnung daher weiter:

$$W_{Verfahren} = B_{Rechtsdurchsetzung} + B_{Abschreckung} + B_{Orientierung} + B_{Zugang} - C_{lit} - C_S$$

29

So liest man die Formel: Sie stellt dieselbe Frage aus Sicht der Rechtsordnung. In Worten: *Ein Verfahren ist gesellschaftlich wertvoll, wenn sein gesamter Nutzen die gesamten Kosten bei Parteien und Justiz übersteigt.*

Baustein	**Bedeutung in einfachen Worten**
$W_{Verfahren}$	Die *gesellschaftliche Gesamtbilanz* des Verfahrens (W = Wohlfahrt).
$B_{…}$	*Nutzenposten* (B = englisch *benefit*), die addiert werden: verwirklichte Ansprüche (*Rechtsdurchsetzung*), Verhaltenssteuerung Dritter (*Abschreckung*), Präjudizwert für künftige Fälle (*Orientierung*) und die Erreichbarkeit von Rechtsschutz (*Zugang*).
C_{lit}	Die *Gesamtkosten der Parteien* (englisch *litigation costs*): $C_P + C_D$.

Baustein	Bedeutung in einfachen Worten
C_S	Der *Aufwand des Gerichts / Staates* (englisch *state / judicial costs*), den die Gerichtskosten nur teilweise decken.

30

Das Zusammenspiel. Die klagende Partei sieht nur einen Ausschnitt dieser Rechnung: Abschreckungs- und Orientierungsnutzen werden ihr nicht vergütet, Gerichtsaufwand und Kosten der Gegenseite trägt sie nicht voll. Private und gesellschaftliche Bilanz können deshalb in beide Richtungen auseinanderfallen. KI senkt vor allem die Prozesskosten der Kläger (C_P) – nicht aber ohne Weiteres den Aufwand des Gerichts/Staates (C_S).

31

Juristisches Beispiel: Eine Konsumentin klagt wegen einer systematisch verrechneten, für sich allein kleinen Gebühr. Ihr individueller Streitwert ist gering, ihre private Rechnung womöglich sogar negativ. Ein klärendes Urteil kann aber verhindern, dass dieselbe Gebühr bei Tausenden weiteren Verträgen erhoben wird, und so einen erheblichen Orientierungs- und Abschreckungsnutzen erzeugen ($B_{Orientierung}$ und $B_{Abschreckung}$ sind hoch). Umgekehrt kann eine offensichtlich unbegründete, automatisch auf zwanzig Seiten aufgeblähte Eingabe kaum zusätzlichen Rechtsdurchsetzungsnutzen schaffen, während sie bei Gericht und Gegenpartei reale Prüfungskosten verursacht (C_S und C_D steigen, ohne dass ein Nutzenposten wächst). Aus gesellschaftlicher Sicht sind deshalb weder «mehr Verfahren» noch «weniger Verfahren» für sich genommen ein sinnvolles Ziel.

32

Die schweizerische Entwicklung passt bemerkenswert gut zu dieser Unterscheidung. Im Fall 4A_231/2026 war die Kostenersparnis gerade das erklärte Motiv des KI-Einsatzes. Die dadurch billiger hergestellte Beschwerde scheiterte jedoch weitgehend an klassischen prozessualen Anforderungen. Das ist

kein Beweis für eine allgemeine Erfolglosigkeit von KI-Eingaben. Es illustriert aber, dass billigere Textproduktion die Kosten sorgfältiger Sachverhalts- und Quellenprüfung nicht beseitigt.

33

Für die prozessrechtliche Reaktion ist das Urteil zugleich differenzierter, als eine reine Formel «Filtere Verhalten, nicht Technologie» vermuten lässt. Das Bundesgericht hat die Möglichkeit einer technologiespezifischen Zulässigkeitsfrage ausdrücklich offengelassen. Eine abschliessende schweizerische Rechtsprechung, wonach der Entstehungsprozess einer Eingabe stets irrelevant wäre, existiert damit nicht. Tragend waren im konkreten Urteil aber weiterhin die Mängel des Outputs und seiner Begründung.

34

Für die ZPO besteht bereits ein technikneutraler Instrumentenkasten. Art. 132 Abs. 2 und 3 ZPO erlaubt unter anderem den Umgang mit weitschweifigen sowie querulatorischen oder rechtsmissbräuchlichen Eingaben. Das Bundesgericht hat zur Weitschweifigkeit Kriterien entwickelt, die übermässige Wiederholungen, sachfremde Ausführungen und eine im Verhältnis zum Streitgegenstand nicht mehr gerechtfertigte Länge erfassen, zugleich aber die Komplexität des konkreten Falles berücksichtigen. Diese Kriterien funktionieren unabhängig davon, ob der Text durch eine Partei, eine Kanzlei oder ein Sprachmodell aufgebläht wurde (BGer 9C_440/2017).

V. Der neue Flaschenhals: öffentliche Verarbeitungskapazität

35

Wenn Rechtsmobilisierung billiger wird, entsteht die entscheidende Asymmetrie zwischen Sender und Empfänger. Eine

Partei kann mit KI zusätzlichen Text erzeugen, ohne proportional zusätzliche Arbeitsstunden aufzuwenden. Ein Gericht kann denselben Text bislang nicht ohne Weiteres ohne proportionalen Prüfaufwand verarbeiten, wenn Behauptungen, Begehren, Aktenstellen und Präjudizien individuell nachvollzogen werden müssen.

36

Die gelegentlich verwendete DoS-Metapher (Denial of Service, d. h. die Überlastung eines IT-Systems durch eine Flut automatisierter Anfragen) ist insoweit illustrativ, sollte aber nicht überstrapaziert werden. Es braucht keine missbräuchliche Absicht, damit eine Kapazitätsasymmetrie entsteht. Viele individuell rationale Nutzungsentscheidungen können kumuliert eine institutionelle Belastung erzeugen. Gerade darin liegt der transaktionskostenökonomische Punkt: Sinkt der Preis einer Aktivität auf einer Seite drastisch, während die Kosten ihrer Verarbeitung auf der anderen Seite weitgehend unverändert bleiben, verlagert sich der Engpass.

37

Die institutionelle Antwort sollte nicht darin bestehen, Rechtsmobilisierung künstlich wieder auf das Kostenniveau vor generativer KI anzuheben. Eine allgemeine «KI-Gebühr» wäre grob: Sie würde einen schwachen maschinell erstellten Schriftsatz anders behandeln als einen gleich schwachen menschlich erstellten und könnte gerade kleine, aber begründete Ansprüche erneut unwirtschaftlich machen.

38

Naheliegender ist eine mehrstufige Verarbeitung. KAPLOWS Theorie der optimalen mehrstufigen Streitentscheidung untersucht, wann zusätzliche Verfahrens- und Informationsstufen ihren Aufwand rechtfertigen. Die Übertragung auf moderne KI-Triage ist keine Aussage KAPLOWS über Sprachmodelle, sondern eine institutionelle Analogie: Die teuerste Prüfungsressource sollte nicht zwingend auf jeden

Fall in derselben Intensität und in derselben Reihenfolge angewendet werden (KAPLOW, 2017).

39

Eine gerichtliche Eingangsstufe könnte beispielsweise Zuständigkeit und formale Vollständigkeit strukturieren, Rechtsbegehren und Streitfragen extrahieren, Wiederholungen erkennen, zitierte Entscheide gegen authentische Datenbanken prüfen, Aktenstellen verknüpfen und eine überprüfbare Übersicht erzeugen. Entscheidend wäre nicht nur die Modellqualität, sondern die Architektur des Systems: Eine maschinelle Zusammenfassung dürfte keine Argumente stillschweigend verschwinden lassen; Aussagen müssten auf die jeweilige Originalstelle zurückgeführt werden können; unsichere oder nicht verifizierte Fundstellen müssten sichtbar bleiben.

40

Die institutionelle Pointe ist daher nicht der «1-Klick-Entscheid». Interessanter ist die überprüfbare 1-Klick-Strukturierung: Maschine zuerst dort, wo Informationen sortiert, verknüpft und auf offensichtliche Inkonsistenzen geprüft werden können; menschliche Aufmerksamkeit dort, wo Wertung, Sachverhaltswürdigung und rechtlich verantwortliche Entscheidung beginnen.

VI. Vom Assistenten zum Entscheider: Wo die institutionelle Grenze verläuft

41

Wie eine solche Arbeitsteilung in der Praxis aussehen kann, zeigt die von LIU und LI dokumentierte Nutzung grosser Sprachmodelle am Shenzhen Intermediate People’s Court. Der dort beschriebene Ablauf war gerade nicht «KI entscheidet, Richter unterschreibt». Vielmehr traf der Richter zunächst

eine Ausgangsentscheidung; danach erzeugte das Sprachmodell einen Begründungsentwurf, den der Richter anschliessend überprüfte und überarbeitete (LIU/LI, 2024/2025).

42

Die Reihenfolge ist normativ entscheidend. Ein vollständig ausgearbeiteter maschineller Entscheidungsvorschlag, der vor der eigenen richterlichen Urteilsbildung vorliegt, kann kognitive Ankereffekte anders auslösen als eine Formulierungshilfe, die erst nach einer menschlichen Ausgangsentscheidung eingesetzt wird. «Human in the loop» ist deshalb eine zu grobe Kategorie. Entscheidend ist, wann, auf welcher Informationsbasis und mit welcher tatsächlichen Entscheidungsfreiheit der Mensch in den Prozess eingreift.

43

Daraus ergibt sich eine allgemeinere Regel: Menschliche Aufmerksamkeit gehört dorthin, wo rechtliche Unsicherheit, Tatsachenwürdigung, Ermessen, Grundrechtsrelevanz und erhebliche Fehlerfolgen sie besonders wertvoll machen. Das gilt für die Richterin ebenso wie für den Anwalt, der eine KI-gestützte Rechtsschrift unterzeichnet. Das ist etwas anderes als ein Sparprogramm. Es ist eine Theorie der Allokation knapper juristischer Aufmerksamkeit.

VII. Fazit: Die knappe Ressource ist verifizierte Urteilskraft

44

Generative KI macht nicht «das Recht» billig. Sie macht bestimmte Arbeitsschritte, mit denen Recht recherchiert, strukturiert, formuliert und mobilisiert wird, erheblich billiger – und erledigt sie zugleich schneller und besser.

45

Gerade weil maschinelle Vorarbeit immer besser wird, wächst eine neue Versuchung: plausiblen Output für geprüften Output zu halten. Plausible Sprache ist aber kein Verifikationsmechanismus. Quellenanbindung, bessere Modelle und durchdachte Schnittstellen können Fehler seltener und leichter erkennbar machen; die Verantwortung für juristische Richtigkeit verschwindet dadurch nicht. Sie bleibt bei der Person, die das Ergebnis unterzeichnet, freigibt oder vor Gericht vertritt.

46

Damit verändert sich die ökonomische Struktur der gesamten Anwaltschaft. Ob in der Rechtsabteilung, in der Grosskanzlei oder in der Einzelpraxis: Überall werden dieselben Arbeitsschritte automatisierbar, und überall gewinnen dieselben Leistungen an relativem Wert – Spezialwissen, Strategie, Verhandlung, Unabhängigkeit und Verantwortungsübernahme. Das spricht für Entbündelung und Anpassungsdruck, nicht für einen bereits vollzogenen Zusammenbruch einer gesamten Branche.

47

Auf der Bürgerseite kann dieselbe Entwicklung die Schwelle zu Rechtsinformation und Rechtsmobilisierung senken. Ob damit effektiver Rechtsschutz steigt, ist eine zweite Frage. Der Fall 4A_231/2026 illustriert den Unterschied zwischen der Fähigkeit, einen Schriftsatz zu produzieren, und der Fähigkeit, eine prozessual tragfähige Beschwerde zu verfassen.

48

Auf staatlicher Seite entsteht daraus die zentrale institutionelle Herausforderung. Es wäre widersprüchlich, private Rechtsmobilisierung künstlich wieder zu verteuern, nur damit die Justiz mit den Verarbeitungskosten des Papierzeitalters weiterarbeiten kann. Bestehende Verfahrensregeln können unzulängliche oder missbräuchliche

Eingaben grundsätzlich nach ihrer Qualität und Wirkung behandeln. Dort, wo technische Besonderheiten eigenständige Risiken erzeugen, können zusätzliche Governance-Regeln erforderlich sein. Gleichzeitig müssen jene Schritte der staatlichen Rechtsverarbeitung automatisierbarer werden, die sich ohne Preisgabe richterlicher Verantwortung strukturieren lassen.

49

Der neue Flaschenhals der Rechtsordnung ist nicht Text. Er ist verifizierte menschliche Urteilskraft.

50

Die Aufgabe des Rechtsstaats besteht nicht darin, KI daran zu hindern, private Rechtskosten zu senken. Sie besteht darin, diesen möglichen Wohlfahrtsgewinn zu erhalten, ohne die dadurch entstehenden Verifikations- und Verarbeitungskosten unkontrolliert auf Gerichte und Gegenparteien zu verlagern.

51

Oder zugespitzt: Wenn Maschinen juristische Texte in Sekunden skalieren können, muss der Rechtsstaat nicht jeden Verarbeitungsschritt weiterhin mit den menschlichen Grenzkosten des Papierzeitalters organisieren. Er darf Code mit Code beantworten – solange am Ende ein Mensch für das verantwortlich bleibt, was nur ein Mensch legitim entscheiden darf.

VII. Was das für Medien konkret bedeutet

52

Für Medienunternehmen ist die beschriebene Kostenverschiebung besonders greifbar, weil die Juristinnen und Juristen, die sie beraten – in der hauseigenen Rechtsabteilung ebenso wie in spezialisierten Kanzleien –, an zwei Stellen gleichzeitig

betroffen sind: **Sie produzieren selbst mehr juristische Vorarbeit zu geringeren Kosten, und sie müssen mit mehr, schneller und professioneller erzeugter Rechtskommunikation der Gegenseite umgehen.**

53

Auf der Produktionsseite liegen naheliegende Anwendungen etwa in der Vorbereitung von Pre-Publication-Reviews oder bei ersten Entwürfen von Antworten auf Beanstandungen. Das sollte allerdings nicht mit einer automatisierten Publikationsfreigabe verwechselt werden. Medienrechtliche Entscheidungen sind häufig gerade dort schwierig, wo die ökonomische Standardisierung an ihre Grenze stösst: bei der Gewichtung von Informationsinteresse und Persönlichkeitsschutz, bei Unsicherheit über Tatsachengrundlagen, beim Quellenschutz, bei drohenden vorsorglichen Massnahmen oder bei Veröffentlichungen mit aussergewöhnlich hohem Reputations- und Schadenspotential. In solchen Situationen liegt der Wert juristischer Beratung nicht primär darin, schneller Text zu analysieren und zu produzieren. Er liegt darin, Unsicherheit einzuordnen, Risikobereitschaft mit der Redaktion abzugleichen und für die Entscheidung Verantwortung zu übernehmen.

54

Besonders relevant ist deshalb eine risikobasierte Arbeitsteilung:

Typischer Vorgang im Medienrecht	**Sinnvolle KI-Unterstützung**	**Was menschlich bleiben sollte**
Standardvertrag, Freelancer- oder Lizenzvertrag	Klauselvergleich, Abweichungsliste, erster Änderungsvorschlag	ungewöhnliche Rechteübertragung, Haftung, strategische Verhandlung

Typischer Vorgang im Medienrecht	Sinnvolle KI-Unterstützung	Was menschlich bleiben sollte
Pre-Publication-Review	Checkliste möglicher Rechtsfragen, Zuordnung relevanter Passagen	Interessenabwägung und Freigabe risikoreicher Inhalte
Beanstandungs- oder Anwaltsschreiben	Extraktion der Vorwürfe, Vergleich mit Artikel und Quellen, erster Antwortentwurf	Anerkennung von Fehlern, Vergleichsstrategie, Eskalationsentscheidung
Prozess oder vorsorgliche Massnahme	Aktenstruktur, Chronologie, Quellenprüfung, Entwurfsmodule	Prozessstrategie, Beweiswürdigung, verbindliche Eingaben

55

Der zweite Effekt könnte mittelfristig sogar wichtiger werden: Nicht nur das Medienunternehmen kann juristische Texte billiger produzieren. Seine Anspruchsteller können es ebenfalls. Eine betroffene Person benötigt künftig möglicherweise keine anwaltliche Erstberatung mehr, um ein formal überzeugendes Beanstandungsschreiben, ein umfangreiches Auskunftsbegehren oder einen Entwurf für eine Eingabe zu erzeugen. Das bedeutet keineswegs, dass diese Ansprüche unbegründet wären. Es bedeutet, dass die bisherige Produktionskostenschwelle als Mengenfilter schwächer wird.

56

Für alle, die Medien rechtlich beraten, spricht deshalb viel dafür, eingehende Kommunikation nicht mehr primär nach ihrer

Länge oder professionellen Sprache zu priorisieren. Ein zehnseitiges KI-unterstütztes Schreiben kann rechtlich trivial sein; eine dreizeilige Nachricht kann ein erhebliches Risiko enthalten. Die Triage sollte deshalb stärker nach Rechtsfolge, Frist, Irreversibilität, Beweislage und Eskalationspotential gewichten als nach dem sichtbaren Produktionsaufwand des Gegenübers.

57

Die dritte Konsequenz betrifft die Verifikationsarchitektur. Medienunternehmen und die Kanzleien, die sie begleiten, verfügen häufig über etwas, das generische Sprachmodelle nicht besitzen: eigene Vertragsbestände, frühere Rechtsgutachten, interne Weisungen, dokumentierte Freigabeentscheide und branchenspezifisches Erfahrungswissen. Der eigentliche Produktivitätsgewinn dürfte deshalb weniger aus einem allgemeinen Chatbot entstehen als aus einem System, das freigegebene interne und externe Quellen auffindbar macht und jede wesentliche Aussage auf ihre Grundlage zurückführt. Solche überprüfbaren Quellenbezüge senken die Kosten der Verifikation erheblich – die menschliche Kontrolle ersetzen sie nicht.

58

Für die juristische Arbeit im Medienbereich ergibt sich daraus eine pragmatische Konsequenz: Juristen sollten möglichst viel ihrer knappen Aufmerksamkeit auf jene Fragen konzentrieren, bei denen ein falscher Entscheid nicht einfach durch eine weitere Textiteration korrigiert werden kann.

59

Die gleiche Logik gilt damit im Kleinen für die medienrechtliche Beratung wie im Grossen für die Justiz: Wenn die Produktion skaliert, wird verlässliche Triage und Verifikation zum entscheidenden Organisationsproblem. Der Wettbewerbsvorteil liegt langfristig weniger darin, wer am meisten juristischen Text erzeugen kann. Entscheidend dürfte

sein, wer am schnellsten erkennt, welcher Teil dieses Textes menschliche Urteilskraft verdient.

Literatur:

KAPLOW, Louis, *The Value of Accuracy in Adjudication: An Economic Analysis*, 23 Journal of Legal Studies 307–401 (1994). Theoretische Grundlage für die Abwägung zwischen Genauigkeit und den dafür eingesetzten Ressourcen.

KAPLOW, Louis / SHAVELL, Steven, *Accuracy in the Assessment of Damages*, 39 Journal of Law & Economics 191–210 (1996). Zu privaten und gesellschaftlichen Anreizen für zusätzliche Genauigkeit.

SHAVELL, Steven, *Economic Analysis of Litigation and the Legal Process*, NBER Working Paper No. 9697 (2003). Zentrale Grundlage zu privaten und gesellschaftlichen Prozessanreizen, negativen Klagewerten und Kostenexternalitäten.

KAPLOW, Louis, *Optimal Multistage Adjudication*, 33 Journal of Law, Economics & Organization 613–652 (2017). Theoretischer Bezugspunkt für die Idee mehrstufiger, informationsabhängiger gerichtlicher Verarbeitung.

SUSSKIND, Richard, *The End of Lawyers? Rethinking the Nature of Legal Services*, Oxford University Press, Oxford 2008. Grundlegend zur Commoditisierung und Entbündelung juristischer Dienstleistungen.

LIU, John Zhuang / LI, Xueyao, *How Do Judges Use Large Language Models? Evidence from Shenzhen*, 16 Journal of Legal Analysis 235–262 (2024/2025). Relevant für die tatsächliche Sequenz Mensch → KI-Begründungsentwurf → menschliche Revision.

Transparenzhinweis des Autors:

Bei der Erstellung dieser Arbeit wurden KI-gestützte Sprachmodelle als methodische und redaktionelle Hilfsmittel eingesetzt. Die grundlegende Ideenfindung, die Entwicklung der Thesen sowie die inhaltliche Ausarbeitung erfolgten ausschließlich durch die Autorenschaft. Claude (Anthropic) wurde primär zur Strukturierung und zur sprachlichen Optimierung der Textentwürfe verwendet. Zur redaktionellen Endkontrolle und Überprüfung der Lesbarkeit kamen zudem ChatGPT (OpenAI) und Gemini (Google) zum Einsatz. Der Autor hat alle Textpassagen sorgfältig geprüft, final angepasst und übernimmt die vollumfängliche Verantwortung für den Inhalt, die Richtigkeit und die Originalität dieser Publikation.



Dieses Werk ist lizenziert unter einer Creative Commons Namensnennung – Weitergabe unter gleichen Bedingungen 4.0 International Lizenz.