

Generative KI und die neue Bewertungskrise in der Wissenschaft

Anne K. Krüger

Einleitung

Die Qualitätssicherung in der Wissenschaft steckt in der Krise. Bereits im Sommer 2025 titelte der *Guardian*: »Quality of scientific papers questioned as academics »overwhelmed« by the millions published« (Sample 2025). Von »der größten wissenschaftlichen Krise aller Zeiten« spricht auch der Neuropsychologe Bernhard Sabel, Mitinitiator der im November 2025 veröffentlichten Stockholm-Erklärung (Wiarda 2025). Hierin wird vor großen Qualitätseinbußen im Wissenschaftssystem gewarnt, wenn sich nichts am derzeitigen Publikationssystem ändere. Hintergrund dieser Krisendiagnosen ist die in letzter Zeit stark steigende Anzahl an Einreichungen bei wissenschaftlichen Zeitschriften, die diese vor die zunehmend größere Herausforderung stellt, sie angemessen zu begutachten. Wissenschaftliche Qualität scheint vor diesem Hintergrund nicht mehr begutachtbar oder aber in der Masse der Publikationen und Zeitschriften nicht mehr auffindbar zu sein.

Dieses Problem wird noch zusätzlich dadurch verschärft, dass Publikationen jene »Währung« darstellen, durch die die Chancen verteilt werden, auf welche Weise oder auch ob Wissenschaftler:innen überhaupt weiterhin Wissenschaft betreiben können. Wieviel jemand publiziert hat, gilt aktuell in vielen Disziplinen als ein zentrales Maß für »Leistung«, die wiederum mit Qualität gleichgesetzt wird. Damit stürzt die Begutachtungskrise des Publikationssystems letztlich auch die Bewertung von Wissenschaftler:innen in eine Krise. Diese Krisen sind allerdings keineswegs neu und auch ihr Zusammenhang ist so naheliegend wie bekannt. Jedoch spitzen sie sich unter dem Ein-

fluss der zunehmenden Nutzung von generativer KI in der wissenschaftlichen Praxis noch einmal zu. Generative KI verschärft damit die bislang diagnostizierten Probleme auf gravierende Weise.

Publikationskrise

Eine erste Publikationskrise wurde bereits in den 1950er Jahren diagnostiziert (Garfield 1955; de Solla Price 1963). Schon zu diesem Zeitpunkt war diese Diagnose von der Sorge getrieben, aufgrund des stetigen Wachstums des Wissenschaftssystems und des damit verbundenen Anstiegs an Publikationen den Überblick über relevantes Wissen zu verlieren. Um es den Bibliotheken in einer immer unübersichtlicher werdenden Publikationslandschaft zu ermöglichen, auf diese Weise die – zunächst für die Naturwissenschaften – zentralen Zeitschriften herauszufiltern, trieb Eugene Garfield die Messung von Zitationen als Maßstab für die Relevanz wissenschaftlicher Publikationen voran (Garfield 1964). Diese Zitationsanalyse wurde dann sehr schnell nicht mehr nur allein zur Bewertung von Zeitschriften, sondern auch als Instrument zur »Leistungsmessung« von Wissenschaftler:innen genutzt (Wouters 1999). Hieraus resultiert jener bis heute anhaltende Fokus auf die Anzahl von Publikationen und ihre Zitationen, die sich als Maßstab für die Bewertung von individuellen Wissenschaftler:innen, aber auch von Forschungsorganisationen und Universitäten bis hin zur »Leistungsfähigkeit« ganzer nationaler Wissenschaftssysteme etabliert hat.

Der Anstieg an Publikationen ist mittlerweile jedoch nicht mehr allein mit dem Wachstum der Wissenschaft an sich verbunden. Stattdessen hat die besagte Kopplung von Publikationen und Bewertung auch bei den Anbieter:innen wissenschaftlicher Zeitschriften zu problematischen Entwicklungen geführt. Große Verlagshäuser haben begonnen, nicht nur immer mehr Zeitschriften als mögliche Publikationsorte anzubieten, sondern auch dafür zu sorgen, dass einzelne Zeitschriften immer mehr Beiträge – insbesondere in Form von Sonderheften – publizieren. Damit hat der Publikationsdruck in der Wissenschaft für die Verlage zu stetig wachsenden Gewinnen geführt. Zunächst wurden diese Zeitschriften den Bibliotheken nicht mehr einzeln zum Abonnement angeboten. Es konnte also nicht mehr nach Qualität ausgewählt werden, welche Zeitschriften den Leser:innen zur Verfügung gestellt werden und welche auch nicht. Stattdessen wurde mittels immer teurer

werdender Publikationspakete der Absatz von kontinuierlich neu herausgebrachten Zeitschriften garantiert.

Hinzu kommt noch eine weitere zentrale Entwicklung im Publikationssystem. Es handelt sich hierbei um die wissenschaftspolitische Forderung von Open Access (siehe hierzu auch Gehring 2026). Dahinter steht die grundsätzliche Idee, freien Zugang zu öffentlich finanzierten wissenschaftlichen Publikationen zu ermöglichen und damit nicht zuletzt auch dem Monopol großer Verlagshäuser und ihren stetig teurer werdenden Lizenzen etwas entgegenzusetzen. Aktuell sind Wissenschaftsorganisationen dadurch vielerorts bei einem System angekommen, in dem nicht mehr für Lizenzen von Publikationspaketen gezahlt wird. Stattdessen wird jetzt fürs Publizieren gezahlt. Dadurch hat sich Open Access nicht nur zu einem wiederum äußerst lukrativen Geschäftsmodell für die großen Verlage entwickelt. Zudem wird das wissenschaftliche Publizieren selbst zum Kostenfaktor und damit auch zunehmend als ökonomische Last für die Bibliotheken jener Einrichtungen wahrgenommen, die besonders viele Publikationen hervorbringen. Denn sie sind für die Zahlung der Article Processing Charge (APC) zuständig, die für jede Veröffentlichung bei den in den DEAL-Verhandlungen eingeschlossenen Verlagen fällig wird (DEAL Open Access Services o.J.). Wenn wissenschaftliche Publikationen jedoch als ökonomisches Problem betrachtet werden, ist der Schritt dahin nicht mehr weit, dass das Veröffentlichen und Zugänglichmachen von Forschungsergebnissen nicht mehr allein von ihrer Qualität, sondern von der finanziellen Liquidität der jeweiligen Wissenschaftsorganisation abhängt. Dass dies gravierende Folgen für die Verbreitung von Forschungsergebnissen hätte, liegt auf der Hand. Und nicht zuletzt würden dadurch auch bereits bestehende Machtverhältnisse, die aktuell insbesondere durch die extrem hohe Befristungsquote (Bloch et al. 2021) und die damit einhergehenden Abhängigkeiten hervorgebracht werden, noch einmal eine ganz neue Dimension von Einfluss auf wissenschaftliche Inhalte gewinnen.

Die Notwendigkeit des Publizierens für die wissenschaftliche Karriere hat jedoch nicht nur »seriöse« Verlage motiviert, neue Zeitschriften herauszugeben und fürs Publizieren Geld zu nehmen. Die Entstehung von *Predatory Journals* macht deutlich, wie hier das Publikationssystem aufgrund des Publikationsdrucks auch zum Ort krimineller Praktiken geworden ist. Denn hier wird Peer-Review lediglich simuliert und gegen die Zahlung eines entsprechenden Betrags eine Veröffentlichung garantiert, was insbesondere vor dem Hintergrund des Anstiegs des *Open Access*-Publizierens und entsprechender

Gebühren für die Wissenschaftler:innen selbst immer weniger sichtbar wird. Das darin vorhandene kriminelle Potenzial zeigt sich darüber hinaus in den *Paper Mills*, die gegen entsprechende Summen die Anfertigung eines vollständigen Artikels anbieten, der dann in einem anerkannten Journal untergebracht wird. In beiden Fällen geht es also um die Simulation des wissenschaftlichen Publikationsprozesses, an dessen Ende aber immer eine Publikation steht, die als weitere Nummer in die Publikationsliste eingefügt und damit zum Kriterium für die quantitative »Leistungsmessung« genutzt werden kann.

Diese bereits existierenden Probleme stellen sich vor dem Hintergrund der sich schnell ausbreitenden Nutzung von generativer KI in der Wissenschaft auf neue Weise. Für *Open Access* ergibt sich hier das Problem, dass auch wissenschaftlicher »Text« massenhaft dazu genutzt wird, große Sprachmodelle zu trainieren, ohne dass ersichtlich ist, mit welchen Texten die KI konkret trainiert wurde. Tatsächlich geht die Nutzung von wissenschaftlichen Publikationen hierfür bereits weit über die offen verfügbaren Publikationen hinaus – wie verschiedene Urheberrechtsklagen und auch erste juristisch ausgehandelte Vergleiche etwa mit *Anthropic* zeigen (The Authors Guild o.J.). Darüber hinaus haben zahlreiche Verlagshäuser bereits begonnen, Nutzungsvereinbarungen mit Entwickler:innen von generativer KI abzuschließen, die dann Zugriff auf ihre Publikationen haben, ohne dass die Autor:innen von diesen Vereinbarungen selbst monetär profitieren (Ithaka S+R o.J.).

Wie eingangs bereits angesprochen, verstärkt die zunehmende Nutzung von generativer KI aber vor allem die Zunahme von Einreichungen und wissenschaftlichen Publikationen auf drastische Weise. Von Verlagen und Zeitschriften hört man allseits, dass die Zahl der Einreichungen von Artikeln rasant steigt. Das Verlagshaus Wiley weist beispielsweise in seiner Bilanz für das erste Quartal 2025 aus, dass 25 Prozent mehr Einreichungen als im Vorjahr zu verzeichnen waren (Wiley 2025). Und natürlich beschränkt sich die Nutzung von generativer KI nicht allein auf die »redliche« Wissenschaft. So mussten Taylor und Francis im Sommer 2025 bei ihrem Journal *Bioengineered* kurzfristig die Möglichkeit der Einreichung aussetzen, nachdem die Herausgeber:innen über 1.000 Einreichungen identifiziert hatten, die von den oben erwähnten *Paper Mills* kamen (Sample 2025; Brainard 2025). Die Möglichkeiten, Texte schneller oder vollständig von generativer KI produzieren zu lassen, hat also das bereits bestehende Problem mit den *Paper Mills* erheblich vergrößert, aber auch insgesamt den ohnehin kontinuierlichen Anstieg von Publikationen noch einmal rasant gesteigert. Für Verlage mag das aufgrund der für jeden Artikel anfallenden APC weiterhin sehr lukrativ sein. Für die

Wissenschaft drängt sich hier aber nicht nur in ökonomischer Hinsicht die Frage auf, was man in Zeiten von generativer KI noch damit gewinnt, wenn weiterhin die Anzahl von Publikationen als zentrales Maß für wissenschaftliche Qualität gilt, oder ob man damit nicht selbst vielmehr das wissenschaftliche Publikationssystem akut gefährdet.

Bewertungskrise

Der Zusammenhang von Publikations- und Bewertungssystem ist ebenfalls seit langem Gegenstand anhaltender Diskussion. Ihr Ausgangspunkt ist die Nutzung von zitationsbasierten Indikatoren zur Messung von »wissenschaftlicher Leistung« und den zahlreichen damit einhergehenden Mess- und Nutzungsproblematiken, angefangen bei unterschiedlichen disziplinären Zitationskulturen über die Nutzung des für Zeitschriften entwickelten *Journal Impact Factors* für die Bewertung einzelner Wissenschaftler:innen bis hin zur grundsätzlichen Frage nach der tatsächlichen Aussagekraft von Zitationen, wenn dabei die eigentlichen Inhalte einer Publikation nicht berücksichtigt werden. Dies wird bereits seit Jahren – wie beispielsweise in der *DORA-Deklaration* (DORA 2012) oder dem *Leiden-Manifest* (Hicks et al. 2015) – als Problem diskutiert. Hier werden stattdessen »responsible metrics« gefordert und es wird für einen verantwortungsbewussten und vor allem auch informierten Umgang mit dieser Form der Quantifizierung von »Leistung« in der Wissenschaft und ihrer Nutzung zur Bewertung von individuellen Wissenschaftler:innen argumentiert.

Darüber hinaus werden seit Anfang der 2000er Jahre »alternative Metriken« promotet, deren quantitative Erfassung und Berechnung durch die Digitalisierung des akademischen Publikationswesens möglich geworden sind. Diese sollen als bessere und vor allem unmittelbarere Indikatoren für das Rezipieren von Publikationen dienen und das auch über die wissenschaftliche Gemeinschaft hinaus. Denn hier wird nicht die Anzahl der Zitationen innerhalb eines bestimmten Zeitraums in der Vergangenheit ermittelt. Stattdessen wird auf aktuelle Download-Zahlen oder Erwähnungen in den sozialen Medien geschaut, die dann nicht nur Auskunft über die Rezeption innerhalb der Wissenschaft geben, sondern auch den »societal impact« messen sollen, also den Einfluss, den wissenschaftliche Ergebnisse auf die Gesellschaft haben (Priem et al. 2010; Haustein 2016).

Darüber hinaus werden in den letzten Jahren Alternativen zu einer solchen Quantifizierung von »Leistung« diskutiert. Die Niederlande haben bereits 2019 mit ihrem Programm *Recognition and Reward* den Versuch gestartet, ihr nationales Forschungsevaluationssystem zu reformieren und auch vielfältige andere Formen wissenschaftlicher »Leistung« in der Bewertung der Qualität von Forschung zu berücksichtigen. Dieser Gedanke steht auch hinter der *Coalition for Advancing Research Assessment* (CoARA), die 2022 aus dem *Agreement on Reforming Research Assessment* (Arentoft et al. 2022) hervorgegangen ist, das von Vertreter:innen der European University Association (EUA), Science Europe und der Europäischen Kommission formuliert und mittlerweile von 700 Wissenschaftsorganisationen unterzeichnet wurde. Mit der Unterzeichnung erfolgt die freiwillige Selbstverpflichtung, sich von quantitativen Leistungsindikatoren zu verabschieden und stattdessen wieder Inhalte zu bewerten.

Doch auch im Hinblick auf qualitative Leistungsbewertung verschärft die Nutzung von generativer KI deutlich die bereits bestehenden Problemlagen. Denn aufgrund der bereits geschilderten Effekte von generativer KI auf das wissenschaftliche Publikationssystem hat die Diskussion um das wissenschaftliche Bewertungssystem noch einmal an Fahrt gewonnen und die Rufe nach seiner Reform weitere prominente Unterstützung erfahren. In der bereits erwähnten, im November 2025 veröffentlichten *Stockholm-Deklaration* warnen Wissenschaftler:innen im Journal *Royal Society Open Science* der Königlich Schwedischen Akademie der Wissenschaften vor großen Qualitätseinbußen im Wissenschaftssystem, wenn sich nichts am derzeitigen kommerziellen Publikationssystem und den damit zentral verbundenen Anreizstrukturen für Fehlverhalten oder auch nur Massenproduktion ändere (Sabel, Larhammar 2025). In ihrem »call to action« wiederholen sie zentrale Forderungen, die sich auch in den bereits erwähnten Diskussionen um *Open Access* und die Reform des Bewertungssystems in der Wissenschaft finden lassen. Sie sprechen sich dafür aus, dass die Wissenschaft wieder selbst die Kontrolle über das wissenschaftliche Publikationswesen auf der Basis eines non-profit-Modells übernehmen sollte. Und sie verlangen, dass das wissenschaftliche Anreiz- und Bewertungssystem nicht mehr Quantität, sondern Qualität belohne. Um über reine Interessensbekundungen vorheriger Erklärungen hinauszukommen, fordern sie, betrügerische Zeitschriften durch eine unabhängige Kontrollorganisation zu kennzeichnen und konkrete Regulationen einzuführen, durch die auch gezielt rechtliche Maßnahmen eingeleitet werden können, um Betrug zu bestrafen.

Trotz dieser Versuche, die wissenschaftliche Leistungsbewertung zu reformieren, lässt sich jedoch gleichzeitig ein weiterer Trend beobachten. Entgegen allen Bemühungen, wieder mehr auf Qualität zu setzen, werden zunehmend mehr quantifizierende Daten über Forschung erfasst. Hierzu trägt die Digitalisierung des akademischen Publikationswesens wesentlich bei, die die »Datafizierung« wissenschaftlicher Praxis vorantreibt und es so ermöglicht, noch mehr Daten über den wissenschaftlichen Arbeitsprozess und seine Ergebnisse zu generieren (Krüger, Petersohn 2022). Daraus ist mittlerweile ein eigener Markt für Forschungsinformation entstanden, in dem Daten über wissenschaftliche Praxis von großen Datenanalysten wie Clarivate, Digital Science oder Elsevier in entsprechenden Datenbanken gesammelt und dazu genutzt werden, daraus Forschungsinformationsprodukte und -services zu entwickeln. Diese werden dann wiederum den Wissenschaftsorganisationen, aber auch Akteuren aus Politik und Wirtschaft zu hohen Preisen angeboten (Krüger 2024). Hierbei geht es nicht nur um das, was bereits 2021 unter anderem in einem Informationspapier des *Ausschusses für Wissenschaftliche Bibliotheken und Informationssysteme* der Deutschen Forschungsgemeinschaft als »Datentracking in der Wissenschaft« angeprangert wurde, durch das die Nutzungsdaten von Wissenschaftler:innen auf wissenschaftlichen Recherche- und Publikationsplattformen von deren Anbieter:innen abgefangen werden (DFG 2021). Vielmehr geht es bei »Forschungsinformation« um eine Vielzahl an miteinander verknüpften personalisierten Daten über wissenschaftliches Arbeiten und die daraus resultierenden Ergebnisse. Verknüpft werden hier beispielsweise Autor:innen, ihre Publikationen, deren Open-Access-Status und mögliche Drittmittelfinanzierung mit den Erscheinungsorten der Publikationen, den Ko-Autor:innen, ihren Arbeitsorten und deren geografischer Lage. Zusammen mit den Abstracts und Keywords werden daraus dann Parameter ermittelt wie beispielsweise die automatisierte Zuordnung von wissenschaftlichen Ergebnissen zu den *Sustainable Development Goals* der Vereinten Nationen oder die Ermittlung von internationalen Kooperationen durch Ko-Autorschaften.

Zu solchen gewinnorientierten Anbieter:innen ist eine Gegenbewegung entstanden, die den freien Zugang zu diesen Daten über öffentliche Forschung einfordert, um sowohl Informationen über Forschung als auch über Forschende frei zugänglich und offen einsehbar zu halten. Dies ist auch die zentrale Forderung der im April 2024 veröffentlichten *Barcelona-Deklaration* (Barcelona Declaration on Open Research Information et al. 2024), die von einer Gruppe von 25 Personen aus dem Kontext der Herstellung und Nut-

zung von Forschungsinformation verfasst wurde. Ziel dieser Erklärung ist es, dass sich Wissenschaftsorganisationen verpflichten, nur noch Daten und entsprechende digitale Infrastrukturen zu nutzen, die von nicht-gewinnorientierten Organisationen angeboten werden. Während man also versucht, einen freien Zugang zu Forschungsinformationsdaten zu ermöglichen und dabei davon spricht, für mehr Transparenz zu sorgen, bleibt hier jedoch die Frage weitgehend unterbelichtet, wofür diese Daten verwendet werden sollen. Es geht also in erster Linie um eine Reform der Infrastruktur und nicht der Bewertung selbst – auch wenn hier mittlerweile Verbindungen beispielsweise zu CoARA geknüpft werden.

Mit der sich ausbreitenden Nutzung von generativer KI stellt sich aber auch die Frage nach der Validität von Metriken noch einmal auf ganz neue Weise. Betrachtet man die »Rezeption« von wissenschaftlichen Publikationen durch generative KI, so wird deutlich, dass durch Large Language Models (LLMs) zwar wissenschaftliche Texte zur Generierung von Antworten genutzt, diese aber nicht notwendigerweise zitiert werden. Die Sichtbarkeit und Anerkennung der dahinterstehenden Publikationen und ihrer Autor:innen ist damit nicht mehr gegeben und eine Messung der Rezeption wie bei herkömmlichen oder alternativen Metriken nicht mehr möglich. Mit der absehbaren weiteren Verbreitung von generativer KI in der Wissenschaft und darüber hinaus werden diese bisherigen Metriken damit noch einmal mehr an ihrer ohnehin umstrittenen Aussagekraft verlieren, was auch für diejenigen deutlich werden müsste, die bislang noch an der Quantifizierbarkeit von wissenschaftlicher Leistung festgehalten haben.

Mit der Zukunft des Publikationswesens und seiner Verbindung zum wissenschaftlichen Bewertungssystem beschäftigt sich auch eine vor Kurzem vorgelegte Studie der Cambridge University Press (2025), die selbst als Verlagshaus zahlreiche Zeitschriften herausgibt. In dieser Studie geht es einerseits um die derzeitigen Entwicklungen des Open-Access-Publizierens. Andererseits wird hier unmittelbar auf den Zusammenhang zwischen dem Publikations- und dem Bewertungssystem in der Wissenschaft hingewiesen und dies unter den Vorzeichen der Herausforderungen durch generative KI problematisiert. Dabei unterstreicht diese Studie einmal mehr die große Unzufriedenheit mit dem aktuellen Bewertungssystem, in dem Quantität vor Qualität kommt. Gefragt wurde hier aber auch, inwieweit insbesondere generative KI als Problem für das Publikationssystem gesehen wird. Über die Hälfte der Befragten stimmte der Aussage zu, dass KI das Publikationssystem negativ beeinflussen wird, während lediglich 18 Prozent der Befragten dieser Aussage widersprachen.

Begutachtungskrise

Die doppelte Herausforderung des Peer-Review-Systems durch generative KI

Die Frage der Beeinflussung des Publikationssystems durch generative KI stellt sich jedoch nicht nur hinsichtlich der (Ko-)Produktion von Einreichungen bei wissenschaftlichen Zeitschriften, sondern auch hinsichtlich ihrer Begutachtung. Bereits seit Jahren wird darüber diskutiert, wie man die Begutachtung der stetig wachsenden Anzahl an Einreichungen gewährleisten kann, wenn der ›Goldstandard‹ der Qualitätssicherung auf einem Peer-Review basiert, in dem Fachkolleg:innen als Gutachtende die Qualität einer Einreichung bewerten. Ausgangspunkt ist, dass es für die Redakteur:innen wissenschaftlicher Fachzeitschriften immer schwieriger wird, adäquate Gutachter:innen zu finden (Adam 2025). Dies wird sowohl als Bedrohung für die Qualität der final publizierten Artikel als auch als Problem für ihre zeitnahe Veröffentlichung angesehen. Letzteres hat vermutlich auch das Verlagshaus Taylor & Francis auf die Idee gebracht, die Geschwindigkeit von Peer-Review zu einer zusätzlichen Einnahmequelle zu machen. Mit ihrem Angebot eines »accelerated publication«-Prozesses, in dem man für fünfstelligen US-Dollar-Beträge einen beschleunigten Begutachtungsprozess erhalten kann (Taylor & Francis, o.J.), haben sie einigen Unmut in der wissenschaftlichen Community erzeugt.

Als Ursache für diese Begutachtungskrise wird einerseits der kontinuierliche Anstieg sowohl der Zahl von Zeitschriften als auch von Publikationen gesehen. Andererseits wird das Problem der fehlenden Motivation auf Seiten der Gutachtenden genannt, die aufgrund dieses Anstiegs sehr viele Anfragen bekommen. Auf die fehlende Motivation ist mittlerweile mit verschiedenen Versuchen reagiert worden, für eine Anerkennung dieser Tätigkeit zu sorgen. Zeitschriften publizieren – zumeist am Jahresende – verbunden mit einer Danksagung gesammelt die Namen von Gutachtenden, um auf diese Weise der Begutachtungsarbeit Sichtbarkeit und den Gutachtenden Anerkennung zu verschaffen. Außerdem gibt es verschiedene Ansätze, Reviewtätigkeiten selbst zu einem relevanten Aspekt der Bewertung von Wissenschaftler:innen zu machen. Dazu gehört nicht nur die mittlerweile standardmäßige Auflistung in wissenschaftlichen Lebensläufen, für welche Publikationsorgane eine Wissenschaftler:in bereits begutachtet hat. Zudem stellen die oben bereits erwähnten Anbieter:innen von digitalen Infrastrukturen zur Erfassung von Daten über wissenschaftliche Praxis auch Möglichkeiten bereit, die eigenen Reviewtätigkeiten darzustellen und dafür Anerkennung zu

erfahren. So wurde beispielsweise die als Versuch der Reviewer-Anerkennung gestartete Plattform *Publons* mittlerweile in die bibliometrische Datenbank *Web of Science* integriert, die vielerorts als Grundlage für Forschungsinformation und -evaluation genutzt wird.

Aktuell wird an verschiedenen Orten noch eine weitere Lösung für die knappe Ressource »Begutachtung« eingeführt. Zur Unterstützung des menschlichen Peer-Reviews wird zunehmend generative KI eingesetzt. Im Gegensatz zur Nutzung von LLMs in der Herstellung von wissenschaftlichen Publikationen, die bereits seit der Veröffentlichung von ChatGPT weitgehend akzeptiert ist, wenn KI-Modelle nicht selbst als Autor:in gelistet und Anforderungen hinsichtlich einer Nutzungskennzeichnung eingehalten werden, hat generative KI zwar wesentlich länger gebraucht, um auch in der Begutachtung als legitim anerkannt zu werden. Tatsächlich sieht man aber seit einigen Monaten, dass generative KI immer häufiger – unter bestimmten Voraussetzungen – zur Begutachtung zugelassen oder sogar bereits aktiv in der Begutachtung auf LLMs zurückgegriffen wird.

Die *Association for the Advancement of Artificial Intelligence* (AAAI) hat beispielsweise bereits im Mai 2025 ein Pilotprojekt gestartet, in dem LLMs im Peer-Review eingesetzt werden. Allerdings sind die Grenzen hier eng gesteckt. So wird auf der Website versichert, dass diese Initiative zwar dazu gedacht ist, die Effizienz im Peer Review zu steigern, aber man trotzdem »the highest standards of scientific rigor and human oversight« beibehalte (AAAI 2025). Dennoch werden LLM-generierte Gutachten als eine Komponente des initialen Reviewprozesses betrachtet, die »an additional perspective alongside traditional human expert evaluations« darstelle (ebd.). Jedoch wird betont, dass die Entscheidung selbst immer bei den menschlichen Reviewer:innen verbleibe und auch keine automatisierte Entscheidungsfindung angestrebt werde.

Wie weit vorangeschritten bereits Entwickler:innen von KI-basierten Begutachtungs-Tools – zumindest hinsichtlich ihrer Versprechungen von Fähigkeiten der von ihnen entwickelten Produkte – sind, hat die Journalistin Miryam Naddaf in *Nature* zusammengetragen (Naddaf 2025a). Zwar geht es hier bislang vor allem um die Überprüfung formaler Kriterien oder die Erstellung von Zusammenfassungen. Aber bereits jetzt werden Systeme angeboten, durch die auch eine Bewertung der Neuheit der dargestellten Ergebnisse vorgenommen wird. Es deutet sich dementsprechend an, dass zumindest der Entwicklung solcher KI-basierter Funktionen keine Grenzen gesetzt sind. Auch vollständig automatisierte Reviews stellen hier kein Tabu mehr dar.

Eine solche Verschiebung von Grenzen zeigt sich auch an der Organisation *openRxiv*, die die beiden Repositorien *bioRxiv* und *medRxiv* betreibt, auf denen Pre-prints, also Publikationen, die noch nicht bei einer Zeitschrift publiziert wurden, vorab veröffentlicht werden können. Diese Repositorien haben aktuell ebenfalls mit einer Flut an Einreichungen zu kämpfen, die überprüft werden, bevor sie zugänglich gemacht werden. *OpenRxiv* hat vor diesem Hintergrund ein Pilotprojekt gestartet, das alle biomedizinischen Einreichungen einer KI-generierten Begutachtung unterzieht, die bereits nach 30 Minuten ein Ergebnis ausspuckt und dabei nicht nur formale Kriterien checkt, sondern auch Originalität bewerten oder logische Fehler identifizieren soll und sogar Verbesserungsvorschläge macht. Der Neurobiologe Giorgio F. Gilestro schreibt dazu in einem Beitrag in *Nature*, dass dies zwar gute Nachrichten für alle sein dürften, die sonst monatelang auf eine Begutachtung gewartet haben. Jedoch dürfe man nicht Effizienz gegen Validität eintauschen und damit das Vertrauen in die Qualität der Begutachtung und in das Begutachtungssystem insgesamt untergraben (Gilestro 2025). Denn LLMs machen nicht nur Fehler, sondern können aufgrund ihrer Funktionsweise immer nur Mittelwerte und deren Abweichungen erkennen. Inwieweit es sich bei solchen Abweichungen also tatsächlich um innovative Ideen oder eher um »kalten Kaffee« handelt, lässt sich auf diese Weise nicht bewerten. Darüber hinaus stellt sich hier die Frage nach der Grundlage, auf der eine solche KI-generierte Bewertung erstellt wird. Aus welchen Trainingsdaten sich die Bewertung speist, ist nicht nur für die Validität der inhaltlichen Kritik von Belang, sondern auch hinsichtlich der Berücksichtigung von Ansätzen, die (noch) nicht zum Mainstream der Forschung gehören und deshalb aufgrund der wahrscheinlichkeitsbasierten Funktionsweise der Modelle nicht beachtet werden. Hinzu kommt, dass anzunehmen ist, dass in diese Modelle Texte eingeflossen sind, die aus *Predatory Journals* stammen und daher zusätzlich zu einer Verzerrung der Inhalte beitragen. Solche Probleme sind bereits aus der Zitationsanalyse bekannt. Diese basiert zwar nicht auf der Verarbeitung von großen Mengen unstrukturierter Texte, sondern auf bibliometrischen Datenbanken. Doch auch diese Datenbanken sind schon lange hinsichtlich der Auswahl ihrer Inhalte, das heißt der darin berücksichtigten Publikationen und Publikationsorte, in der Kritik. Denn diese Auswahl entscheidet darüber, welche Forschung und welche Forschenden als besonders oft zitiert gelten – was sich je nach Datenbank deutlich unterscheiden kann (Stahlschmidt, Stephen 2022). Außerdem ist zu vermuten, dass in das Training der KI-Modelle Artikel eingeflossen sind, die zwar aufgrund von nachgewiese-

nen Problemen wieder zurückgezogen wurden, aber dennoch weiterhin in diesen Modellen enthalten sind (Gunter 2026). Es ist also vollkommen ungeklärt, auf welcher wissenschaftlichen Grundlage durch generative KI Bewertungen erzeugt werden und über welche inhaltliche Tiefe solche Beurteilungen verfügen, die über Allgemeinplätze hinausgehen – auch wenn der Duktus KI-generierter Texte dies gern vergessen lässt.

Diese Problematik hat sich auch bei den Einreichungen für die jährlich stattfindende *International Conference on Learning Representations* gezeigt (Naddaf 2025b). Nach ersten Vermutungen, dass Konferenzeinreichungen vollständig durch generative KI beurteilt worden sein könnten, hat eine großflächige Analyse der eingereichten Gutachten ergeben, dass rund 21 Prozent der Gutachten vollständig KI-generiert waren. *Nature*-Journalistin Naddaf berichtet, dass dies zu großem Unmut unter den Forschenden geführt habe, weil unklar gewesen sei, wie sie mit Gutachten umgehen sollten, die teils relevante Punkte, teils aber auch völligen Nonsense enthielten. Dieses Beispiel macht deutlich, dass ein weiterer zentraler Aspekt im wissenschaftlichen Peer-Review die Erklärbarkeit von Bewertungen ist. Dieser Punkt ist besonders vor dem Hintergrund relevant, dass das bisherige Peer-Review-System immer wieder aufgrund seiner Intransparenz in den Entscheidungsprozessen kritisiert worden ist (Hesselmann, Schendzielorz, Krüger 2021). Deshalb wird hier bereits seit Jahren immer wieder darüber diskutiert, wie durch mehr Sichtbarkeit, zum Beispiel der Gutachtenden, und mehr Einsehbarkeit, zum Beispiel in die Gutachten, Machtverhältnisse aufgebrochen und inhaltliche Einflussnahmen, die nicht wissenschaftlich begründet sind, vermieden werden können. Die Herstellung von mehr Transparenz ist damit eine der zentralen Forderungen, durch die rein wissenschaftlich motivierte Bewertungen garantiert werden sollen, um auf diese Weise sowohl die daraus hervorgegangene Publikation als auch das jeweilige Peer-Review-Verfahren selbst zu legitimieren. Zwar gibt es immer wieder Stimmen, die genau aus diesem Grund für den Einsatz von KI in der Begutachtung werben. Argumentiert wird, dass den subjektiv verzerrten Entscheidungen menschlicher Gutachtender – vermeintlich – objektive Technik entgegengestellt und der Review-Prozess dadurch besser und fairer (oder zumindest »genauso gut oder schlecht«) gemacht werden könne. Abgesehen davon, dass gerade die unterschiedlichen subjektiven Expertisen und Erfahrungen in der Erforschung eines bestimmten Gegenstands in einem entsprechend durch die Herausgeber:innen kuratierten Review-Prozess fruchtbare Anregungen zur Weiterentwicklung eines eingereichten Artikels enthalten können, ist es je-

doch genau dieses Element der Transparenz und der damit einhergehenden Legitimität der Begutachtung, das mit dem Einsatz von generativer KI nun vollständig in Frage gestellt wird. Denn Autor:innen oder auch Herausgeber:innen können zu keinerlei Einschätzung darüber gelangen, wie ein KI-generiertes Gutachten zustande kommt. Während menschliche Gutachter:innen gerade aufgrund ihrer eigenen spezifischen Schwerpunkte und Kenntnisse angefragt werden, kann nicht erklärt und nachvollziehbar gemacht werden, welche Inhalte und deren Gewichtungen zur Bewertung durch ein LLM beigetragen haben. Derartige Antworten auf Rückfragen sind technisch nicht möglich, da die Bewertungen nicht inhaltlich, sondern stochastisch bedingt sind, weshalb die Frage nach dem ursächlichen Sinn ebenfalls nur mittels der wahrscheinlichkeitsbasierten Auswahl von Zeichen gestellt werden kann. Die derzeitigen LLMs sind zwar so gut, dass dies sprachlich auf den ersten Blick nicht unbedingt auffallen mag. Und natürlich können auch durch die zugrundeliegende Reproduktion von Mustern für die Rezipient:innen sinnhafte Inhalte entstehen. Da es im Peer-Review jedoch darum geht, dass hierdurch nicht nur die Einhaltung von formalen Standards überprüft und bereits bestehende Erkenntnisse reproduziert, sondern vielmehr die Innovativität eines Arguments oder die Neuartigkeit eines Resultats eingeschätzt werden soll, ist mehr als fraglich, wie dies durch eine KI-generierte Begutachtung geschehen kann. Insbesondere, wenn ein Review-Prozess über mehrere Runden geht, ist der reflexive Bezug eines LLMs auf die von ihm selbst produzierten Inhalte schlichtweg technisch nicht möglich.

Auswirkungen auf die wissenschaftliche (Begutachtungs-)Praxis

Die gestiegene Anzahl an Einreichungen betrifft jedoch nicht nur Zeitschriften, Repositorien und Konferenzen, sondern auch die Drittmittelgeber:innen. Auch die Deutsche Forschungsgemeinschaft (DFG) hat als zentrale Drittmittelgeberin erstmals bereits im September 2023 Regularien zur Nutzung von generativer KI in der Begutachtung veröffentlicht (DFG 2023). Die Nutzung von generativer KI wurde zunächst verboten. Den Hintergrund hierfür bildeten datenschutzrechtliche Bedenken, aufgrund derer vermieden werden sollte, dass Anträge bei Anbieter:innen von LLMs hochgeladen werden, die diese dann als Trainingsdaten nutzen, wodurch man selbst gegen Urheberrecht und Datenschutz verstoßen würde. Da LLMs mittlerweile auch lokal gehostet werden können, sodass die genutzten Daten

nicht an die Entwickler:innen weitergeleitet werden, entfällt dieses Problem jedoch. Dies hat die DFG im Dezember 2025 dazu bewogen, die Nutzung von generativer KI in der Begutachtung zuzulassen, solange die Gutachtenden »die Nutzung von KI unter Nennung der spezifischen Einsatzzwecke offenlegen«. Auch ist eine »kritische Prüfung aller KI-generierten Inhalte [...] erforderlich und eine unreflektierte Übernahme von Inhalten verboten« (DFG 2025).

Dieses Zugeständnis an die Praxis einer KI-unterstützten Begutachtung, von der man augenscheinlich annimmt, sie nicht anders handhaben geschweige denn kontrollieren zu können, hat in der wissenschaftlichen Community zu einigem Protest geführt (Wiarda 2026). Allerdings muss man zunächst zwischen dem Begutachtungsprozess insgesamt und der darin erfolgenden Bewertung trennen. Denn letztere soll laut DFG weiterhin bei den menschlichen Gutachtenden verbleiben: »Die inhaltliche Verantwortung liegt unverändert bei der Person, die das Gutachten erstellt« (DFG 2026: 4). Dennoch drängt sich auch hier unmittelbar die Frage auf, inwieweit sich der Einfluss von generativer KI auf die eigentliche Bewertung überhaupt begrenzen lässt beziehungsweise tatsächlich reflektiert werden kann.

Dass dies nicht allein eine Frage der Konstruktions- und Funktionsweise der Technik selbst ist, sondern fundamental mit ihren Nutzer:innen zusammenhängt, hat die geistes- und sozialwissenschaftliche Wissenschafts- und Technikforschung umfänglich gezeigt, die sich bereits seit Jahrzehnten mit dem Einfluss von Technik auf menschliche Praxis beschäftigt.¹ Die hier erlangten Erkenntnisse lassen sich in vielerlei Hinsicht unmittelbar auf die Nutzung von LLMs in der Begutachtung übertragen. Dazu gehört vor allem anderen die zentrale Einsicht, dass *immer*, wenn Technik zum Einsatz kommt, diese nicht als reines Werkzeug zur Optimierung von durch Menschen intendierten Ergebnissen verstanden werden kann. Stattdessen verändert Technik sowohl Arbeitspraktiken als auch die Kriterien für ein gutes Arbeitsergebnis, was weit über geläufige Vorstellungen einer gezielten Optimierbarkeit bestimmter Handlungsabläufe durch Technik hinausgeht. Im Fall von generativer KI wird dies zusätzlich dadurch verstärkt, dass diese Technologie nicht nur in menschliche Praktiken eingreift, indem sie standardisierte Routinehandlungen übernimmt und automatisiert.² Darüber hinaus wird generative KI jetzt auch in der ›Wissensarbeit‹ genutzt, also in der Ge-

1 Für einen Einblick siehe Bauer, Voigt, Lemke (2017).

2 Und dabei das, dessentwegen sie eingeführt wird, an ihre eigene technische Funktionsweise anpasst.

nerierung von sinnhaften Inhalten, die bislang menschlichen Expert:innen vorbehalten war (Pakarinen, Huising 2025).

Für das Peer-Review bedeutet dies vor allem zweierlei: Erstens geht es darum, dass hier eine Nutzung von generativer KI als inhaltliche ›Ideengeberin‹ erfolgen kann. Zwar definiert die DFG zunächst sehr eindeutig »den Einsatz von KI als unterstützendes Werkzeug, nicht als Ersatz menschlicher Beurteilung« (DFG 2026: 3). Die Aufgaben, für die generative KI genutzt werden darf, werden hier auf die »sprachliche und strukturelle Unterstützung«, die »Literaturrecherche« und die »formale Textprüfung« beschränkt (ebd.) und dagegen die »wissenschaftliche Analyse, Bewertung und Empfehlung« als jene Aufgaben definiert, die »ausschließlich bei den Gutachter*innen« verbleiben (ebd.: 4).

Aber nicht nur vor dem oben erwähnten Hintergrund stellt sich die Frage, wo konkret die Beeinflussung durch KI einsetzt, was zumindest in einigen Disziplinen schon bei der Literaturrecherche sein kann, die von der DFG explizit als Beispiel für eine erlaubte Nutzung im Reviewprozess genannt wird. Zudem verschwimmen diese selbstgezogenen Grenzen aber auch in dem Positionspapier selbst. Dies beginnt bereits damit, dass die Nutzung von generativer KI in der Begutachtung als Anpassung »an den Stand technischer Innovationen« und Möglichkeit der »Effizienzsteigerung« (ebd.: 3) legitimiert wird. Zwar wird zunächst jegliche Nutzung für Bewertungszwecke abgelehnt, dann aber immer wieder eine »kritische Prüfung« angemahnt, um Verzerrungen zu vermeiden, die »die Objektivität und Vielfalt wissenschaftlicher Bewertungen beeinträchtigen« (ebd.) – womit sicherlich nicht allein durch die KI korrigierte Grammatikfehler gemeint sind. Es ist gerade diese Unschärfe in der Eingrenzung der erlaubten Nutzbarkeit, die hier eher Tür und Tor öffnet, als ein klares Votum gegen die Nutzung für KI-Bewertungen darzustellen. Dies liegt weit jenseits einer »kontrollierte[n] Integration digitaler Werkzeuge in den Begutachtungsprozess« (ebd.). Und wenn es schließlich heißt, dass KI-generierte Inhalte »nicht unreflektiert übernommen«, sondern »auf Richtigkeit, Verzerrungen und fachliche Angemessenheit geprüft werden« (ebd.: 4) müssen, ist die Verwirrung komplett.

Wenn sich auf diese Weise jedoch die Nutzung von generativer KI in die Begutachtung einschleicht, stellt sich gleich noch ein zweites Problem, das tatsächlich noch gravierenderen Einfluss auf wissenschaftliche Qualität haben kann, die gerade durch das Peer-Review hergestellt und gesichert werden soll. Diese Einflussnahme findet bereits vor der Begutachtung beim Verfassen von Einreichungen statt – unabhängig davon, ob Artikel oder

Forschungsantrag. Schon allein das Wissen um die Nutzung von generativer KI bei der Begutachtung kann dazu führen, dass diese Texte so aufbereitet werden, dass sie durch KI besonders gut bewertbar gemacht werden. Während in bestimmten Forschungsrichtungen ohnehin eine starke Standardisierung in der Präsentation von Ergebnissen und Forschungsvorhaben Einzug gehalten hat, sind solche nach ein und demselben Muster aufgebauten Texte in anderen Fachgebieten bislang nur sehr begrenzt üblich oder überhaupt möglich. Der Einsatz von generativer KI in der Begutachtung kann jedoch dazu führen, dass man sich noch einmal genau überlegt, was die Forschungsfrage sein und wie das Forschungsdesign aussehen soll, wofür dann generative KI bereits vor der Einreichung zur Überprüfung der eigenen Texte herangezogen wird. Dadurch läuft das Wissenschaftssystem aktuell Gefahr, dass nicht nur die Form wissenschaftlicher Wissenspräsentation, sondern auch die Pluralität unterschiedlicher Ansätze, Konzeptionen und Perspektiven einem KI-generierten Standard geopfert wird – auch wenn er je nach GPT-Modell jeweils unterschiedlich ausfallen könnte.

Krise durch Wettbewerb

Die drei hier genannten Krisen – Publikationskrise, Bewertungskrise und Begutachtungskrise – gab es unzweifelhaft bereits vor der Entwicklung von generativer KI. Und ganz unzweifelhaft gibt es auch eine Vielzahl von positiven Effekten, die hier jetzt sicherlich zu kurz gekommen sind. Doch machen die aktuellen Entwicklungen, die durch generative KI mit hervorgerufen werden, deutlich, dass sich diese bestehenden Krisen zuspitzen, gerade weil sie so unmittelbar miteinander verzahnt sind und sich in dem Punkt der wissenschaftlichen Qualitätssicherung treffen. Wie schon an verschiedenen Stellen angemerkt wurde, ist vieles davon erst einmal nicht KI-spezifisch, sondern schon lange hausgemacht. Dennoch werden diese Krisen aktuell durch die Nutzung von generativer KI verstärkt, was vielerorts als eine Zuspitzung erlebt wird, die nach deutlicheren Maßnahmen als bisher verlangt.

Dreh- und Angelpunkt ist dabei eindeutig das aktuelle Bewertungssystem in der Wissenschaft, das in immer mehr Disziplinen die Anzahl von Publikationen und Drittmitteln zur Voraussetzung für wissenschaftliche Karrieren oder gar den Verbleib in der Wissenschaft an sich macht. Der dadurch entstehende Publikationsdruck ist ein entscheidender Treiber für die un-

übersichtlich gewordene Publikationslandschaft, aus der nicht nur kommerzielle Anbieter:innen massive Gewinne abschöpfen, sondern auch die Nutzung generativer KI-Werkzeuge³ sowohl zur Erstellung von Inhalten als auch zu deren Bewertung erfolgt. Deshalb bleibt die Frage zentral, wie vor diesem Hintergrund ein zukünftiges Bewertungssystem in der Wissenschaft aussehen könnte, wenn die Anzahl von Publikationen aufgrund ihrer Inflation als ›Währung‹ ausgedient hat. CoARA oder auch die Stockholm-Deklaration setzen hier – wie bereits erwähnt – auf Qualität vor Quantität. Doch gleichzeitig zeigt die Zulassung von KI im qualitativen Review-Prozess, dass sich aus der Nutzung von generativer KI zusätzliche Probleme ergeben, die wissenschaftliche Qualität gefährden.

Doch vielleicht muss diese Frage anders gestellt werden: Hat hier nicht vielmehr ein System ausgedient, das permanent auf Bewertung und Wettbewerb setzt und dadurch diese Krisen überhaupt erst erzeugt? Sollten wir nicht weniger über eine Reform der Bewertung nachdenken, als vielmehr darüber, wo wir Bewertung wirklich brauchen? Könnten nicht bereits Personalstrukturen, die eine wesentlich frühere Absicherung der eigenen wissenschaftlichen Karriere ermöglichen, oder ausreichend Grundmittel, die Forschung auch ohne Drittmittelantrag möglich machen, hier Publikations- und Antragsdruck verringern? Vor dem Hintergrund der aktuellen Entwicklungen sollten wir nicht nach weiteren technischen Lösungen für ein durch Technik erzeugtes Problem suchen, sondern die grundlegende Frage danach stellen, wo man tatsächlich weiterhin Bewertung und den dahinterstehenden Wettbewerb braucht.

Literatur

- AAAI – Association for the Advancement of Artificial Intelligence 2025: AAAI Launches AI-Powered Peer Review Assessment System. <https://aaai.org/aaai-launches-ai-powered-peer-review-assessment-system/>, letzter Aufruf am 25. Juni 2026.
- Adam, David 2025: The peer-review crisis: how to fix an overloaded system. *Nature*, vol. 644, no. 8075, 24–27, doi: 10.1038/d41586-025-02457-2.
- Arentoft, Michael et al. 2022: Agreement on Reforming Research Assessment. doi: 10.5281/zenodo.13480728.

3 Die mittlerweile von denselben Anbieter:innen entwickelt und vertrieben werden wie zuvor die Zeitschriften und dann die Dateninfrastrukturen.

- Barcelona Declaration on Open Research Information / Kramer, Bianca / Neylon, Cameron / Waltman, Ludo 2024: Barcelona Declaration on Open Research Information. <https://zenodo.org/records/10958522>.
- Bauer, Susanne / Voigt, Torsten / Lemke, Thomas (Hg.) 2017: Science and Technology Studies. Klassische Positionen und aktuelle Perspektiven. Berlin: Suhrkamp.
- Bloch, Roland / Krüger, Anne, K. / Hartl, Jacob / Würmann, Carsten / Cloppenburg, Jan / Keller, Andreas 2021: Arbeits- und Beschäftigungsbedingungen an Hochschulen in Deutschland. Politische Steuerungsversuche und ihre Effekte seit 2006. Bielefeld: wbv Media.
- Brainard, Jeffrey 2025: Journal plagued with problematic papers, likely from paper mills, pauses submissions. Science Insider, 3 July 2025, doi: 10.1126/science.z3lm5o3.
- Cambridge University Press & Assessment 2025: Publishing Futures: Working Together to Deliver Radical Change in Academic Publishing. <https://www.cambridge.org>, letzter Aufruf am 25. Juni 2026.
- DEAL Open Access Services o.J.: Verträge. www.deal-konsortium.de/vertraege/, letzter Aufruf am 25. Juni 2026.
- Declaration on Research Assessment (DORA) 2012: Erklärung von San Francisco zur Forschungsbewertung. <https://sfdora.org/read/read-the-declaration-deutsch/>, letzter Aufruf am 25. Juni 2026.
- de Solla Price, Derek 1963: Little Science, Big Science. New York: Columbia University Press.
- DFG – Deutsche Forschungsgemeinschaft 2021: Datentracking in der Wissenschaft: Aggregation und Verwendung bzw. Verkauf von Nutzungsdaten durch Wissenschaftsverlage. Ein Informationspapier des Ausschusses für Wissenschaftliche Bibliotheken und Informationssysteme der Deutschen Forschungsgemeinschaft, doi: 10.5281/zenodo.5900759.
- DFG – Deutsche Forschungsgemeinschaft 2023: Stellungnahme des Präsidiums der Deutschen Forschungsgemeinschaft (DFG) zum Einfluss generativer Modelle für die Text- und Bilderstellung auf die Wissenschaften und das Förderhandeln der DFG. www.dfg.de/resource/, letzter Aufruf am 25. Juni 2026.
- DFG – Deutsche Forschungsgemeinschaft 2025: Künstliche Intelligenz in der Begutachtung. Information für die Wissenschaft Nr. 102, 15. Dezember 2025. www.dfg.de/de/aktuelles/, letzter Aufruf am 25. Juni 2026.
- DFG – Deutsche Forschungsgemeinschaft 2026: Künstliche Intelligenz in der wissenschaftlichen Begutachtung. Position und Perspektiven der DFG. <https://zenodo.org/records/18886257>.
- Garfield, Eugene 1955: Citation Indexes for Science. A New Dimension in Documentation through Association of Ideas. Science, vol. 122, no. 3159, 108–111, doi: 10.1126/science.122.3159.108.
- Garfield, Eugene 1964: »Science Citation Index« – A New Dimension in Indexing. Science, vol. 144, no. 3619, 649–654, doi: 10.1126/science.144.3619.649.

- Gehring, Petra 2026: Shining Diamond? Zum Publizieren bei Bibliotheken als Ausweg aus der Krise des wissenschaftlichen Publikationssystems. *Merkur*, 80. Jg. Heft 920, 20–33.
- Gilestro, Giorgio, F. 2025: AI reviewers are here – we are not ready. *Nature*, vol. 648, no. 8092, 9. doi: 10.1038/d41586-025-03909-5.
- Gunter, Darrell 2026: Why scholarly publishing needs a neutral governance body for the AI age. *Research Information*, 14 January 2026, <https://www.researchinformation.info/analysis-opinion/>, letzter Aufruf am 25. Juni 2026.
- Haustein, Stefanie 2016: Grand Challenges in Altmetrics. Heterogeneity, Data Quality and Dependencies. *Scientometrics*, vol. 108, no. 1, 413–423, doi: 10.1007/s11192-016-1910-9.
- Hesselmann, Felicitas / Schendzielorz, Cornelia / Krüger, Anne K. 2021: Sichtbarkeitskonstellationen im Journal Peer Review – Konsequenzen von In/Transparenz in wissenschaftlichen Bewertungsverfahren. In Oliver Berli / Stefan Nicolae / Hilmar Schäfer (Hg.), *Bewertungskulturen. Soziologie des Wertens und Bewertens*. Wiesbaden: Springer VS, 71–92.
- Hicks, Diana / Wouters, Paul / Waltman, Ludo / De Rijcke, Sarah / Rafols, Ismael 2015: Bibliometrics. The Leiden Manifesto for Research metrics. *Nature*, vol. 520, no. 7548, 429–431, doi: 10.1038/520429a.
- Ithaka S+R o. J.: Generative AI Licensing Agreement Tracker. <https://sr.ithaka.org/our-work/generative-ai-licensing-agreement-tracker/>, letzter Aufruf am 25. Juni 2026.
- Krüger, Anne, K. 2024: Neue Daten, neues Wissen? Konsequenzen der Digitalisierung für die Leistungsbewertung in der Wissenschaft. In Jörg Pohle / Rainer Fischbach / Klaus Lenk (Hg.), *Die gesellschaftliche Macht digitaler Technologien. Zwischen Wellen technologischen Überschwangs und den Mühen der Ebene*. Marburg: Metropolis-Verlag, 255–277.
- Krüger, Anne, K. / Petersohn, Sabrina 2022: From Research Evaluation to Research Analytics. The digitization of academic performance measurement. *Valuation Studies*, vol. 9, no. 1, 11–46.
- Naddaf, Miryam 2025a: AI is Transforming Peer Review – and many scientists are worried. *Nature*, vol. 639, no. 8056, 852–854, doi: 10.1038/d41586-025-00894-7.
- Naddaf, Miryam 2025b: Major AI Conference Flooded with Peer Reviews Written Fully by AI. *Nature*, vol. 648, no. 8093, 256–257, doi: 10.1038/d41586-025-03506-6.
- Pakarinen, Pauli / Huising, Ruthanne 2025: Relational Expertise. What Machines Can't Know. *Journal of Management Studies*, vol. 62, no. 5, 2053–2082, doi: 10.1111/joms.12915
- Priem, Jason / Taraborelli, Dario / Groth, Paul / Neylon, Cameron 2010: Altmetrics: A manifesto, doi: 10.5281/zenodo.12684249.
- Sabel, Bernhard / Larhammar, Dan 2025: Reformation of science publishing: the Stockholm Declaration. *Royal Society Open Science*, vol. 12, no. 11, 251805.

- Sample, Ian 2025: Quality of scientific papers questioned as academics overwhelmed by the millions published. *The Guardian*, 13 July 2025.
- Stahlschmidt, Stephan / Stephen, Dimity 2022: From indexation policies through citation networks to normalized citation impacts: Web of Science, Scopus, and Dimensions as varying resonance chambers. *Scientometrics*, vol. 127, no. 5, 2413–2431.
- Taylor & Francis o.J.: Accelerated publication, <https://taylorandfrancis.com/medical-publication-professionals/>, letzter Aufruf am 25. Juni 2026.
- The Authors Guild o.J.: Bartz v. Anthropic Settlement: What Authors Need to Know. www.authorsguild.org/advocacy/, letzter Aufruf am 25. Juni 2026.
- Wiarda, Jan-Martin 2025: »Die größte wissenschaftliche Krise aller Zeiten«. *Der Wiarda-Blog*, 5. November 2025. www.jmwiarda.de/blog/, letzter Aufruf am 23. Juni 2026.
- Wiarda, Jan-Martin 2026: Aufregung über ein Symptom. *Der Wiarda-Blog*, 12. Januar 2026. www.jmwiarda.de/blog/, letzter Aufruf am 25. Juni 2026.
- Wiley 2025: AI Demand Drives Wiley's First Quarter 2026 Results. <https://newsroom.wiley.com/press-releases/>, letzter Aufruf am 25. Juni 2026.
- Wouters, Paul 1999: *The Citation Culture*. PhD Thesis, University of Amsterdam, Digital Academic Repository. <https://dare.uva.nl/>, letzter Aufruf am 25. Juni 2026.