



Wasserzeichen für KI-erzeugte Texte

Aussagekraft null?

Essay von Joachim Reinhart

Für das Auge unsichtbar aber maschinell lesbar: Wasserzeichen sind digitale Fingerabdrücke, die zeigen, ob ein Text mit KI erzeugt wurde. Sie sollen der Nachverfolgbarkeit dienen, etwa beim Erkennen von Plagiaten oder Fake-News. Zunächst klingt es nach der perfekten Lösung – aber sagen Wasserzeichen wirklich so viel aus, wie sie sollen?

Stand: Oktober 2026

www.bwcon-research.org

Einleitung

Die Aussagekraft des Wasserzeichens

Die Freude über den Liebesbrief war groß, aber dann kam die Enttäuschung: Er wurde mit Hilfe künstlicher Intelligenz geschrieben. Ein unsichtbares Wasserzeichen hat das verraten.

Ein Wasserzeichen gibt an, dass ein KI-System an der Erzeugung der konkreten sprachlichen Fassung beteiligt war. Aber sagt es auch aus, von wem die Gedanken stammen? Wie viel eigene Leistung der Absender erbracht hat? Empfindet der Absender die beschriebenen Gefühle tatsächlich? Und ist der Brief deshalb weniger authentisch als ein vollständig selbst verfasster?

Darauf gibt das Wasserzeichen keine Antwort. Und genau darin liegt ein bislang wenig beachtetes Problem: Wasserzeichen können Hinweise darauf geben, wie ein Text entstanden ist – aber nicht darauf, wem seine Gedanken, seine Leistung oder seine Aussage zuzurechnen sind.



Exkurs: Cyrano de Bergerac – ein KI-Problem aus dem Jahr 1897

Das Problem ist älter als ChatGPT. In Edmond Rostands 1897 uraufgeführtem „Cyrano de Bergerac“ liebt Cyrano Roxane, die sich wiederum zu Christian hingezogen fühlt. Weil Christian die passenden Worte fehlen, verfasst Cyrano für ihn Liebesbotschaften und Briefe. Dabei fließen zunehmend Cyranos eigene Gefühle in die Texte ein. Die Geschichte wirft eine überraschend moderne Frage auf: Was macht *eine* Äußerung eigentlich zu *meiner* Äußerung? Muss ich die Worte selbst gefunden haben, oder können auch fremde Worte authentisch sein, wenn sie ausdrücken, was ich wirklich denke und empfinde?

Mit generativer KI erhält diese Frage neue Aktualität. Stellen wir uns vor, Christian bittet ein Sprachmodell, seine Gefühle in einen Liebesbrief zu fassen. Die KI liefert einen treffenden Text, den er sorgfältig liest und voll und ganz unterschreibt. Daneben steht ein zweiter Mann, der seinen Liebesbrief selbst schreibt, aber kein Wort davon ernst meint. Welcher Brief ist authentischer? Welche Äußerung ist aufrichtiger? Ein Wasserzeichen könnte allenfalls darauf hinweisen, dass der erste Brief von einer KI formuliert wurde. Über die Ehrlichkeit der Gefühle würde es nichts aussagen. Die alte literarische Frage ist damit zu einer hochaktuellen Debatte geworden: Entscheidend ist nicht unbedingt, wer die Worte geschrieben hat, sondern ob sie tatsächlich das ausdrücken, was ein Mensch sagen möchte.

Passenderweise ist Cyrano längst selbst in der KI-Forschung angekommen: Als „Cyrano-Effekt“ wird untersucht, wie Menschen Sprachmodelle nutzen, um sich etwa im Online-Dating oder in der digitalen Kommunikation sprachlich besser darzustellen und wie sich das auf die Wahrnehmung von Authentizität auswirkt.

Problemstellung

Was ein KI-Wasserzeichen tatsächlich erkennt

Das alles ist kein Versagen der Wasserzeichentechnik. Sie wurde einfach nur für eine andere Aufgabe entwickelt.

Große Sprachmodelle erzeugen Texte Schritt für Schritt. Für jeden nächsten Textbaustein berechnen sie Wahrscheinlichkeiten für verschiedene mögliche Fortsetzungen. Bei einem bekannten Verfahren von Kirchenbauer und Kollegen wird diese Auswahl geringfügig beeinflusst. Bestimmte grundsätzlich plausible Fortsetzungen werden nach einem verborgenen Schema etwas bevorzugt. Über einen längeren Text entsteht dadurch ein statistisches Muster. Ein entsprechender Detektor kann später untersuchen, ob dieses Muster ungewöhnlich deutlich vorhanden ist.

Googles SynthID-Text verfolgt ein technisch anderes, aber verwandtes Prinzip: Auch dort wird der Auswahlprozess während der Texterzeugung so beeinflusst, dass sich anschließend ein statistisches Signal erkennen lässt. Ein solches Wasserzeichen kann also eine sinnvolle und wichtige Frage beantworten: Passt die statistische Struktur dieses Textes zu einem entsprechend gekennzeichneten maschinellen Erzeugungsprozess?

Dabei handelt es sich um statistische Evidenz, nicht um einen unfehlbaren Herkunftsnachweis. Wasserzeichen können durch Umschreiben oder andere Veränderungen abgeschwächt werden, und die Forschung hat auch grundsätzliche Grenzen besonders robuster Verfahren aufgezeigt.

Aber selbst ein hypothetisch perfektes Wasserzeichen würde das eigentliche Problem nicht lösen. Denn es wüsste noch immer nicht, wie der Text zustande gekommen ist.

Dasselbe Etikett – völlig unterschiedliche Leistungen

Nehmen wir zwei Fachartikel. Autor A gibt einem Sprachmodell die Anweisung: „Verfasse einen kritischen Artikel über Wasserzeichen für KI-Texte.“ Das Modell entwickelt These, Struktur, Argumentation und Formulierung. Der Mensch übernimmt das Ergebnis nahezu unverändert.

Autor B entwickelt die Fragestellung, These und Beispiele selbst. Er nutzt das Sprachmodell für einzelne Passagen, prüft und verwirft Vorschläge und bringt neue Gedanken ein. Auch hier stammt die sprachliche Endfassung vom Sprachmodell. Technisch könnten damit beide Texte als KI-erzeugt erkannt werden. Die menschliche geistige Leistung hinter ihnen wäre jedoch grundverschieden.

Noch deutlicher wird das in einem dritten Fall: Eine Autorin schreibt einen Artikel vollständig selbst und lässt anschließend mittels KI die Grammatik, Zeichensetzung und den Stil verbessern – ausdrücklich ohne Aussagen oder Argumentation zu verändern. Die KI erzeugt die Endfassung. Ist dieser Artikel deshalb im gleichen Sinne „KI-erzeugt“ wie der erste?

Fünf Fragen statt einer

Das Grundproblem: Mit „menschlich“ oder „KI-erzeugt“ sollen mehrere Fragen gleichzeitig beantwortet werden. Mindestens fünf sollten unterschieden werden:

1. Wer hat die Wörter technisch erzeugt?
2. Wie ist der Text entstanden?
3. Von wem stammen die Gedanken und Entscheidungen?
4. Steht der Absender hinter dem Gesagten?
5. Wer übernimmt die Verantwortung für die getroffenen Aussagen?

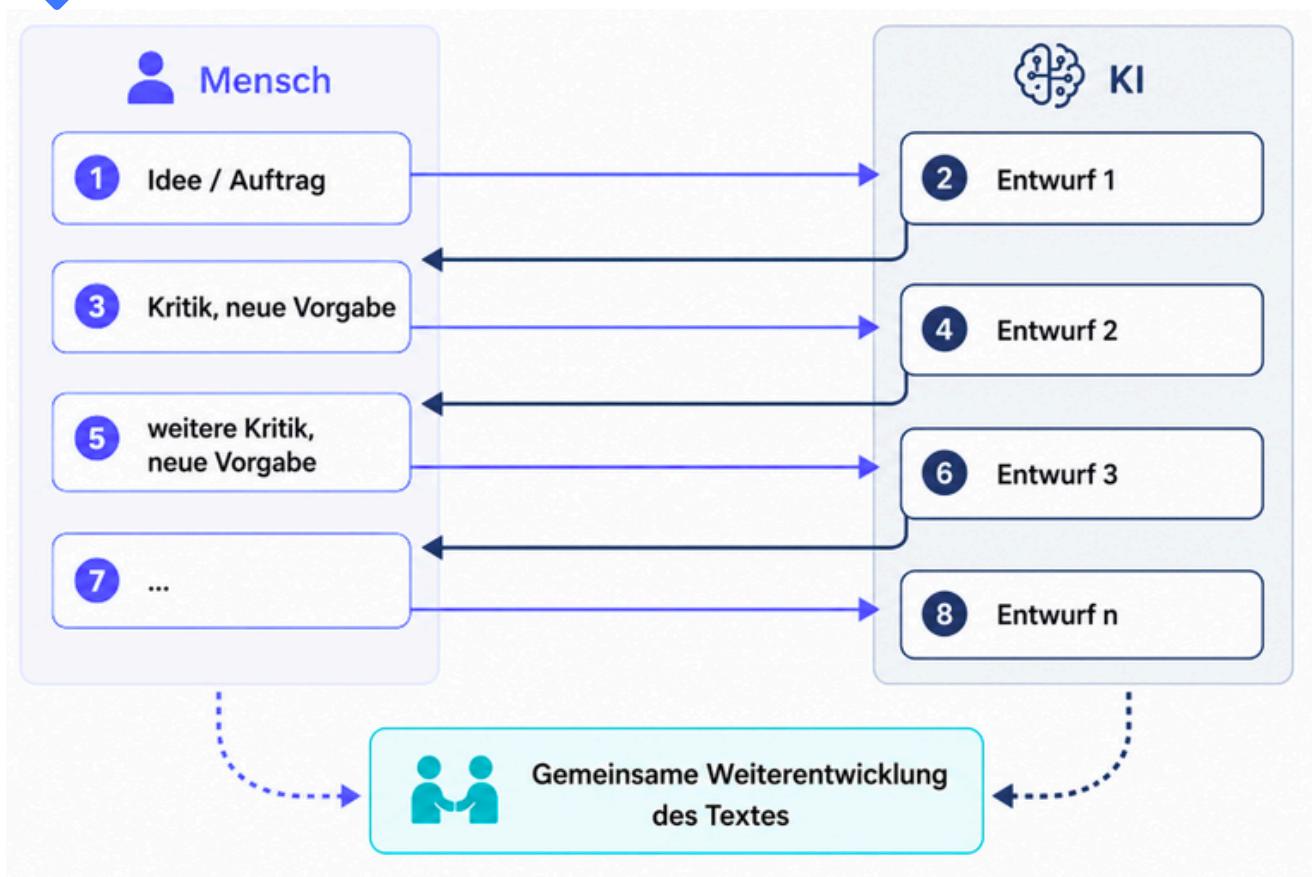
Diese fünf Fragen können zu völlig unterschiedlichen Antworten führen. Ein Wasserzeichen beantwortet im Wesentlichen nur die erste.

Doch selbst „Wer hatte welche Idee?“ kann zu einfach sein

Bei intensiver Zusammenarbeit zwischen Mensch und KI entsteht noch ein weiteres Problem. Wir stellen uns den Ablauf häufig so vor:



Anspruchsvollere Nutzung verläuft jedoch mehr in einem Rückkopplungsprozess:



In diesem Prozess werden sowohl Mensch als auch KI vom gegenseitigen Austausch zu neuem Input, neuen Ideen und Entscheidungen gelenkt. Wem gehört letztendlich dann das Ergebnis?

Systemtheoretische und kybernetische Ansätze beschäftigen sich seit Langem mit der Bedeutung solcher Wechselwirkungen und Rückkopplungen. Ihr wichtiger Gedanke für die gegenwärtige Diskussion lautet: Ein Ergebnis muss nicht einfach die Summe voneinander unabhängiger Einzelbeiträge sein. Bestimmte Eigenschaften können erst in der Wechselwirkung entstehen.

Für die Zusammenarbeit mit Sprachmodellen ist das eine entscheidende Einsicht. Die relevante Frage lautet deshalb möglicherweise nicht nur: Wer hat welchen Beitrag geliefert? Sondern zusätzlich: Was ist erst im Zusammenspiel entstanden? Ein Wasserzeichen der Endfassung kann auch diese Frage nicht beantworten.

Das Wasserzeichen misst – wir geben dem Ergebnis Bedeutung

Damit wird ein weiterer Unterschied wichtig. Der technische Detektor trifft zunächst eine vergleichsweise enge Unterscheidung:

- **Wasserzeichen gefunden / Wasserzeichen nicht gefunden**

Menschen und Institutionen machen daraus jedoch möglicherweise ganz andere Unterscheidungen:

- Eine Hochschule könnte daraus schließen: Eigenleistung / keine Eigenleistung
- Ein Leser: authentisch / nicht authentisch
- Ein Verlag: offenkundigspflichtig / nicht offenkundigspflichtig
- Urheberrechtliche Frage: menschlich geschaffen / nicht menschlich geschaffen

Niklas Luhmann beschreibt Kommunikation nicht als einfache Übertragung eines fertigen Inhalts von einem Sender zu einem Empfänger. Kommunikation entsteht erst im Zusammenhang von Information, Mitteilung und Verstehen.

Auf unser Problem übertragen bedeutet das: Das Wasserzeichen liefert einen technischen Befund. Was dieser Befund bedeutet, hängt vom jeweiligen Zusammenhang ab. Beim Liebesbrief interessiert die Aufrichtigkeit, bei einer Prüfung die Eigenleistung. In der Wissenschaft die Entstehung von Erkenntnis und die Verantwortung für sie. Im Urheberrecht die menschliche schöpferische Leistung. Das technische Signal kann dasselbe sein. Die gesellschaftlich relevante Frage ist es nicht.



Urheberrecht

Das Urheberrecht zeigt, wie schwierig diese Unterscheidungen sind

Wenn die Ideen und Anweisungen vom Menschen stammen, dann sei eben der Mensch der Autor des gesamten KI-Textes.

Das US Copyright Office hat sich 2025 ausführlich mit dem Thema beschäftigt. Der Einsatz generativer KI schließt menschlichen Urheberrechtsschutz nicht aus. Menschliche Beiträge, schöpferische Auswahl, Anordnung und Bearbeitungen können geschützt werden. Entscheidend ist aber die tatsächliche menschliche Kontrolle über die konkrete Ausdrucksform. Ausführliche Anweisungen allein reichen bei heutigen KI-Systemen dagegen in der Regel nicht aus.

Es muss daher unterschieden werden zwischen geistigem Beitrag, technischer Texterzeugung und urheberrechtlicher Autorschaft. Ein Mensch kann die entscheidende Idee und Argumentation liefern, ohne automatisch Urheber jedes von einer KI formulierten Satzes zu sein. Umgekehrt beweist die maschinelle Erzeugung eines Satzes nicht, dass der Mensch keinen erheblichen geistigen Beitrag zum Gesamtwerk geleistet hat.

Auch der europäische Gesetzgeber unterscheidet

Artikel 50 der europäischen KI-Verordnung enthält Transparenzpflichten für bestimmte KI-erzeugte oder manipulierte Inhalte. Die EU-Leitlinien unterscheiden dabei zwischen geringfügiger Bearbeitung und substanzieller Erzeugung oder Veränderung. Die KI-Verordnung erkennt damit etwas Wichtiges an: KI-Beteiligung ist nicht gleich KI-Beteiligung.



Und auch die Wissenschaft interessiert sich für das „Wie“

Das International Committee of Medical Journal Editors fordert, dass nicht nur die reine KI-Nutzung offengelegt wird, sondern auch eine Beschreibung, wie sie eingesetzt wurde. Die Verantwortung für Richtigkeit, Integrität und Veröffentlichung verbleibt beim Menschen.

Fazit & Ausblick

Wer hat eigentlich diesen Beitrag geschrieben?

Damit wird die Diskussion selbst zum Beispiel. Auch der Text, den Sie gerade lesen, ist im Zusammenspiel von Mensch und KI entstanden. Die ursprüngliche Frage nach der Aussagekraft von Textwasserzeichen stellte der menschliche Autor. Von ihm stammt auch der Titel „Wasserzeichen für KI-erzeugte Texte: Aussagekraft null?“. Er brachte auch den Exkurs zu Cyrano de Bergerac in die Diskussion ein, ging auf Vorschläge ein und korrigierte Interpretationen. Das Sprachmodell entwickelte daraufhin neue Fassungen und der Beitrag veränderte sich bei diesem Zusammenspiel. Wer hat diesen Beitrag also geschrieben?

Das lässt sich nur beantworten, wenn wir vorher klären, was wir mit „geschrieben“ meinen.

Die konkrete sprachliche Fassung wurde in großen Teilen von einem Sprachmodell erzeugt. Fragestellung, Titel und mehrere entscheidende konzeptionelle Eingriffe stammen vom Menschen, detailliertes Feedback und Korrekturen von einem reviewenden Kollegen.

Andere Gedanken entwickelten sich erst in der wiederholten Wechselwirkung. Die Auswahl der endgültigen Fassung und die Verantwortung für eine Veröffentlichung liegen beim menschlichen Autor.

Wäre das verwendete Sprachmodell mit einem entsprechenden Textwasserzeichen ausgestattet und bliebe dieses bei der weiteren Bearbeitung erhalten, könnte ein Detektor möglicherweise die maschinelle Erzeugung der konkreten sprachlichen Fassung erkennen.

Dieser Beitrag demonstriert damit selbst das Problem, von dem er handelt.



Also doch: Aussagekraft null?

Nein. Das wäre eine ebenso falsche Schlussfolgerung. Ein technisch gutes Wasserzeichen kann wertvolle Informationen über die maschinelle Herkunft eines Textes liefern. Bei massenhaft erzeugten Inhalten, Desinformation oder Herkunftsnachweisen kann diese Information ausgesprochen relevant sein.

Problematisch wird es, wenn aus „Diese sprachliche Ausgabe weist Merkmale eines gekennzeichneten KI-Erzeugungsprozesses auf“ unbemerkt „Dieser Text stammt von einer KI“ und daraus schließlich „Der angegebene Autor hat diesen Text nicht selbst erarbeitet“ wird. Diese drei Aussagen sind nicht gleichbedeutend. Die Schwäche des Wasserzeichens besteht also nicht darin, dass es technisch schlecht misst, sondern, dass aus einer Information über die technische Herkunft Aussagen über geistige Eigeneistung, Autorschaft, Aufrichtigkeit oder Verantwortung abgeleitet werden.

Vom Wasserzeichen zum Herkunftsprotokoll?

Neben dem einfachen Etikett „KI-erzeugt“ wäre eine Beschreibung des Entstehungsprozesses aussagekräftiger:

- **menschlicher Ausgangstext → KI-gestützte sprachliche Überarbeitung → menschliche Endredaktion**
- **oder: menschliche Recherche und Argumentation → KI-gestützte Formulierung → menschliche Überarbeitung**
- **oder: überwiegend KI-erzeugter Entwurf → menschliche Auswahl und redaktionelle Bearbeitung**

Mit Standards wie C2PA und den darauf beruhenden Content Credentials gibt es bereits Ansätze, Entstehung und Bearbeitung digitaler Inhalte technisch nachvollziehbar zu dokumentieren. Auch solche Herkunftsinformationen können jedoch nicht feststellen, von wem ein Gedanke stammt oder wer sich eine Aussage tatsächlich zu eigen macht.

Doch selbst ein perfektes Herkunftsprotokoll hätte Grenzen. Denn bei intensiver Zusammenarbeit entstehen Gedanken nicht immer unabhängig voneinander. Eine Idee kann eine Reaktion auf einen maschinellen Vorschlag sein, der seinerseits auf einer vorherigen menschlichen Kritik beruht.

Deshalb lautet die eigentlich interessante Frage möglicherweise nicht mehr: Mensch oder KI? Auch nicht nur: Wie viel Mensch und wie viel KI?

Sondern: Wie ist dieser Text entstanden, welche Rolle spielten Mensch und Maschine dabei – und welche Schlussfolgerung wollen wir daraus überhaupt ziehen?

Cyrano musste seine Briefe noch im Verborgenen schreiben. Heute kann ein Sprachmodell seine Rolle übernehmen.

Aber die alte Frage bleibt: Was macht *eine* Äußerung eigentlich zu *meiner* Äußerung?



- Manchmal interessiert uns, wer die Worte erzeugt hat.
- Manchmal, wer die Gedanken entwickelt hat.
- Manchmal, wer sie sich zu eigen macht.
- Und manchmal schlicht, wer sie wirklich meint.

Ein Wasserzeichen kann uns bei der ersten Frage helfen. Die anderen müssen wir bis auf weiteres wohl selbst beantworten.

**KI verstehen. KI nutzen.
KI richtig machen.**



Holen Sie sich Orientierung
in der neuen KI-Sprechstunde
von bwcon!

Sie haben Fragen zum Thema? Kontaktieren Sie uns!



Joachim Reinhart

Projektleiter Innovations- & Forschungsprojekte



+49 176 47669984



reinhart@bwcon.de

www.bwcon-research.org

Literatur

Ante, L. (2026): The Cyrano effect: LLM-assisted impression management and authenticity in online dating. Telematics and Informatics, 108, 102422. <https://www.sciencedirect.com/science/article/pii/S0736585326000584>

Coalition for Content Provenance and Authenticity (C2PA) (2026). C2PA Technical Specification, Version 2.4. April 2026. https://spec.c2pa.org/specifications/specifications/2.4/specs/C2PA_Specification.html

Dathathri, S. et al. (2024): Scalable watermarking for identifying large language model outputs. Nature, 634, 818–823. <https://pmc.ncbi.nlm.nih.gov/articles/PMC11499265/>

Europäische Kommission (2026): Guidelines on transparency obligations for providers and deployers of certain AI systems. <https://digital-strategy.ec.europa.eu/en/policies/guidelines-ai-transparency-obligations>

Europäische Union (2024): Verordnung (EU) 2024/1689 über künstliche Intelligenz (AI Act), insbesondere Art. 50. https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=OJ:L_202401689

Hwang, A. et al. (2025). "It was 80% me, 20% AI": Seeking authenticity in co-writing with large language models. Proceedings of the ACM on Human-Computer Interaction, 9(2), Article CSCW122, 1–41. DOI: 10.1145/3711020. <https://dl.acm.org/doi/epdf/10.1145/3711020>

International Committee of Medical Journal Editors (ICMJE) (2026). Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals, Abschnitt „Use of Artificial Intelligence in Publishing“, insbesondere „Use of AI by Authors“. Aktualisiert Januar 2026. <https://www.icmje.org/icmje-recommendations.pdf>

Kirchenbauer, J. et al. (2023): A Watermark for Large Language Models. Proceedings of the 40th International Conference on Machine Learning, PMLR 202, 17061–17084. <https://proceedings.mlr.press/v202/kirchenbauer23a/kirchenbauer23a.pdf>

Luhmann, N. (1992): What is Communication? Communication Theory, 2(3), 251–259. <https://luhmann.ir/wp-content/uploads/2021/07/What-is-Communication.pdf>

Rostand, E. (1898). Cyrano de Bergerac: Comédie héroïque en cinq actes, en vers. Paris: Charpentier et Fasquelle.

U.S. Copyright Office (2025): Copyright and Artificial Intelligence, Part 2: Copyrightability. <https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-2-Copyrightability-Report.pdf>

Zhang, H. et al. (2024): Watermarks in the Sand: Impossibility of Strong Watermarking for Language Models. Proceedings of the 41st International Conference on Machine Learning. <https://arxiv.org/pdf/2311.04378>