FREIE UNIVERSITÄT BERLIN

UNIVERSITÄTS BIBLIOTHEK

ChatGwP

Plagiate und Co. – Wissenschaftliches Fehlverhalten in der Praxis

Dr. Armin Glatzmeier | Universitätsbibliothek

Gliederung

1. Was ist eigentlich ‚Gute wissenschaftliche Praxis‘?
2. Was sind Plagiate?
3. Autorschaft – Verantwortung …und ein wenig LLM
4. Wie funktionieren Große Sprachmodelle (Large Language Models, LLMs)
5. Was ‚wissen‘ große Sprachmodelle
6. Rules for tools – genKI und GwP
7. Urheberrechtlich Aspekte bei der Nutzung von LLMs
8. Datenschutzrechtliche Aspekte bei der Nutzung von LLMs
9. Ethische Aspekte

Anhang: Was motiviert zu wissenschaftlichem Fehlverhalten?

Was ist eigentlich ‚Gute wissenschaftliche Praxis‘?

Fall 1 – „E pluribus unum“

Der aufstrebende Nachwuchspolitiker Andreas „Andi“ Maria de Bonnemontaigne ist über die Liste in den Bundestag eingezogen und will nun in seinem Wahlkreis für das sichere Direktmandat kandidieren. Leider hat die derzeitige Direktkandidatin Dr. Claudia Chrysantemis keinerlei Rücktrittsambitionen.

Daher beschließt der Jungjurist nach seinem ersten Staatsexamen, zumindest bei den akademischen Meriten gleichzuziehen, um die Parteifreundin in einer Kampfkandidatur abzulösen. Als Doktorvater kann Bonnemontaigne den renommierten Verfassungsrechtler Prof. Dr. Häferle gewinnen.

Er liest viel und zwischen Politik und Familie findet er auch etwas Zeit, einige Gedanken aufzuschreiben. Der Zeitdruck wächst. Einige Rechercheaufgaben delegiert er als mandatsbezogene Informationsanfragen an den Wissenschaftlichen Dienst des Bundestages (WD). Aus den Quellen und den WD-Gutachten stellt er ‚seinen‘ Text zusammen. Stolze 475 Seiten und 1218 Fußnoten später ist das Werk vollbracht. Es bringt ihm summa cum laude, die Veröffentlichung in einem hochangesehenen Verlag und das Direktmandat ein. Die Promotionsurkunde ist noch frisch, als Bonnemontaigne ins Kabinett berufen wird.

Der Absturz folgt zwei Jahre später: Mittlerweile Liebling des Boulevards, erfährt die bass erstaunte Öffentlichkeit, dass über 90 Prozent der Dissertation des Ministers Plagiate enthalten – Internetaktivisten hatten dies herausgefunden. Es folgt Empörung, Leugnung und ein schmachvoller Rücktritt.

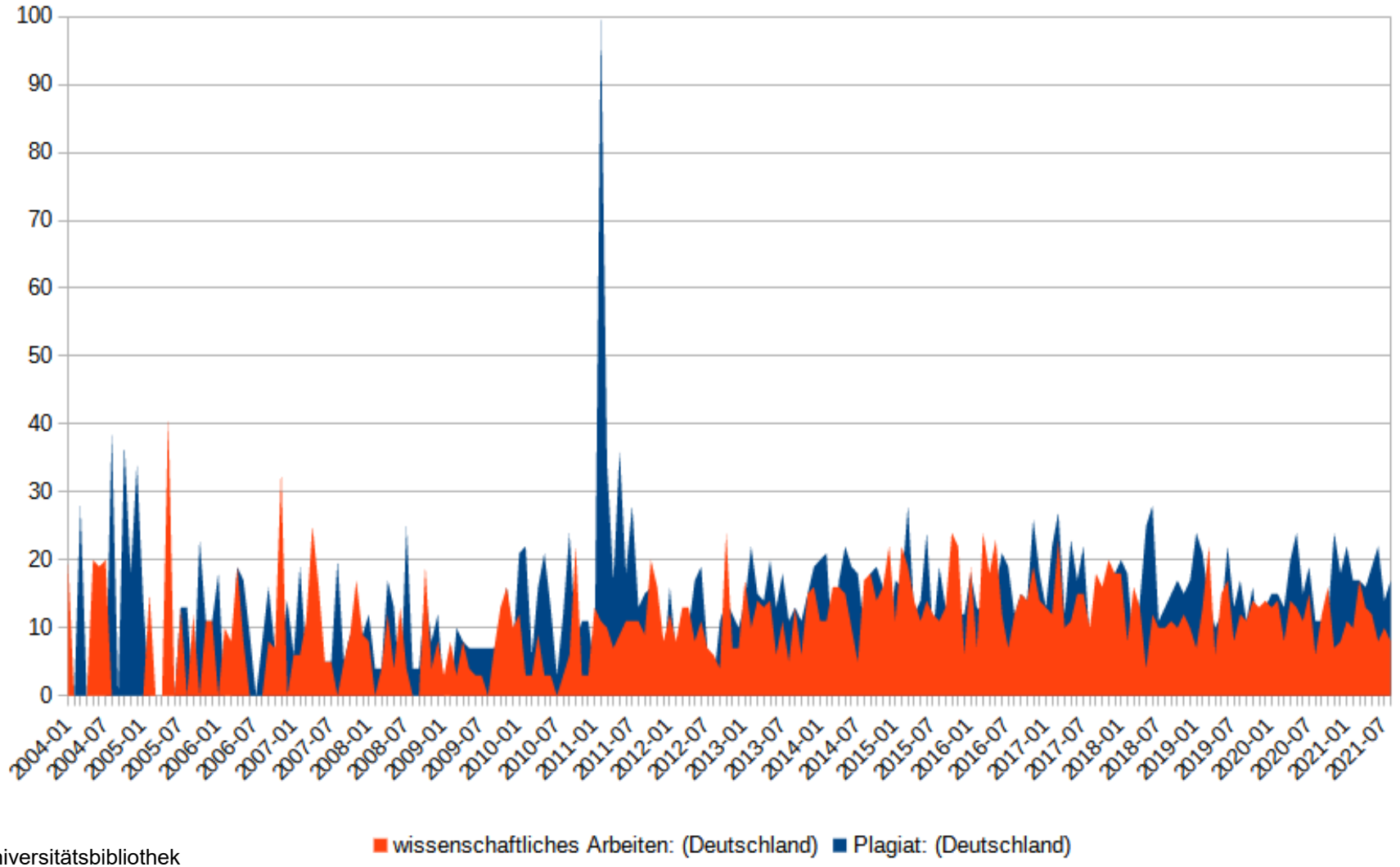
Folgen aus dem Fall Guttenberg (2011)

Unmittelbar

- Breite öffentliche Diskussion über Wissenschaftsplagiate
  - Wissenschaftlichen Konsequenzen (insb. Nicht-Verjährung)
  - Gesellschaftliche Konsequenzen (insb. Depromotion und politische Ämter)
- Höheres Interesse an „guter wissenschaftlicher Praxis“
- Diskussion über Whistleblower und die ‚Selbstreinigungskräfte‘ der Akademie

Guttenbergs Peak

Suchanfrage: https://trends.google.de/trends/explore?cat=958&date=all&geo=DE&q=%2Fm%2F0c8d1,wissenschaftliches%20Arbeiten (22.08.2021)



Folgen aus dem Fall Guttenberg (2011)

Unmittelbar

- Breite öffentliche Diskussion über Wissenschaftsplagiate
  - Wissenschaftlichen Konsequenzen (insb. Nicht-Verjährung)
  - Gesellschaftliche Konsequenzen (insb. Depromotion und politische Ämter)
- Höheres Interesse an „guter wissenschaftlicher Praxis“
- Diskussion über Whistleblower und die ‚Selbstreinigungskräfte‘ der Akademie

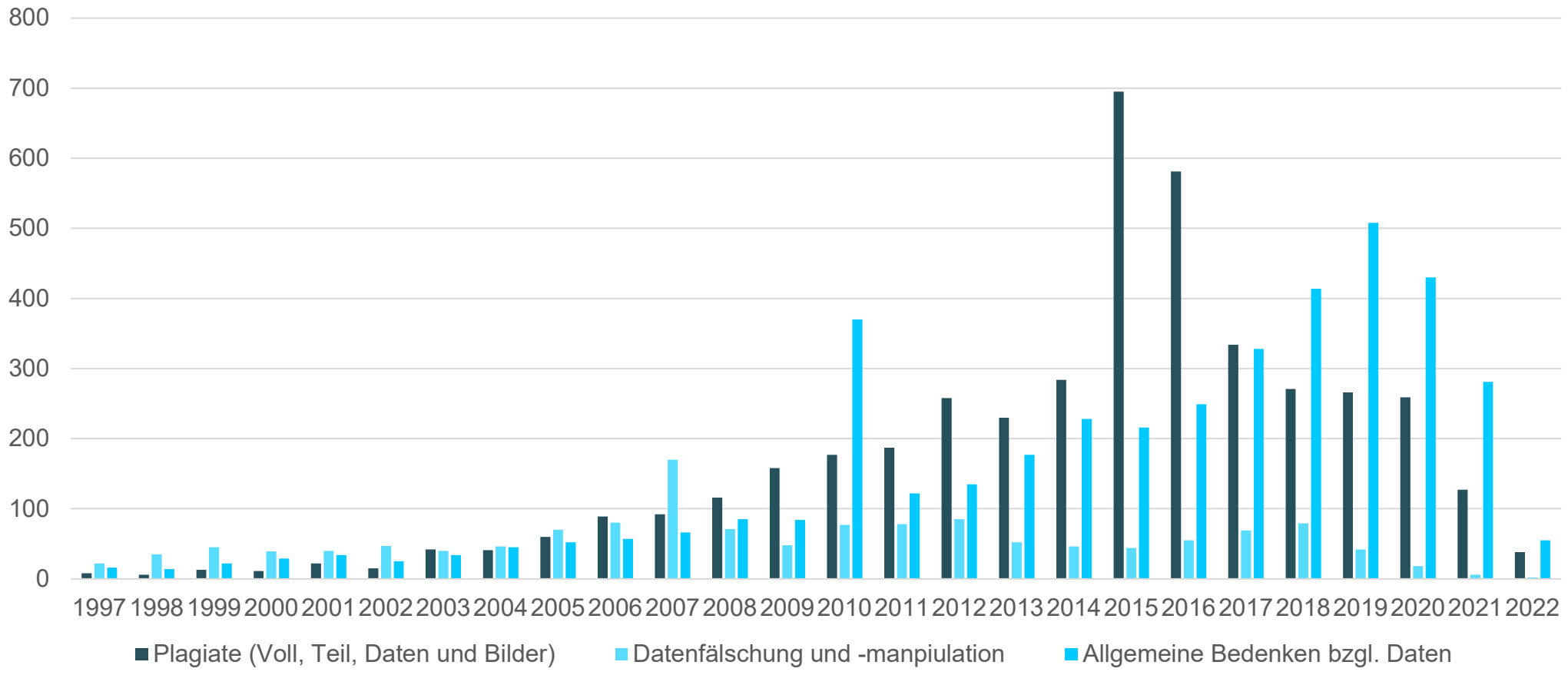
Kurz- und mittelfristig

- Diskussion über den Umfang, ab wann Plagiate sanktioniert werden sollen (im Vergleich etwa zu Silvana Koch-Mehrin, Jorgo Chatzimarkakis, Annette Schavan et al.)
- Verengung der Diskussion über wissenschaftliches Fehlverhalten auf Plagiate

Langfristig

- Änderungen in Promotionskollegs u.ä., Plagiats-Policies usw. sowie Hochschulrecht
- Novellierung DFG-Kodex

Retraction Database: Zurückgezogene Artikel



Fall 2 – Ein toxisches Duo

1988 wandte sich die Medizinstudentin Marion Brach mit der Bitte um Betreuung und Aufnahme in seine Arbeitsgruppe an den renommierten Krebsforscher Friedhelm Herrmann. Nach erstem Zögern stimmte Herrmann zu, als Brach anbot, sich durch zunächst unentgeltliche Arbeit ihren Platz in der Arbeitsgruppe zu verdienen.

Aus dieser ersten Begegnung entstand nicht nur eine überaus produktive Zusammenarbeit, in deren Verlauf sich Brach habilitierte, sondern auch eine private Beziehung, die beide Seiten im Nachhinein aus jeweils eigener Deutung als toxisch beschreiben.

Die Arbeitsbedingungen in der AG waren prekär: Hohe Arbeitslast bei hoher Personalfluktuation führten dazu, dass oft unklar blieb wer, welche Daten mit welchen Methoden erhoben oder ausgewertet hatte.

Am Ende der Zusammenarbeit stand der bis dahin größte Skandal der deutschen Krebsforschung:

Herrmann wurde nachgewiesen, dass er einen Forschungsantrag, den er als Gutachter zunächst als nicht förderungswürdig befand, fast identisch bei der Thyssen-Stiftung eingereicht hatte. Im Ergebnis brachte ihm dies Drittmittel in Höhe von rund 200.000 D-Mark.

Darüber hinaus ergab eine Prüfung im Auftrag der DFG Beanstandungen bei 94 von 347 Publikationen wegen Datenfälschung und/oder -manipulation. Bei 121 weiteren konnte ein Anfangsverdacht nicht vollständig entkräftet werden. Die übrigen 132 Schriften hielten einer Überprüfung durch einen eigens eingesetzte Task-Force stand.

Herrmann und Brach verloren ihre Professuren. Ein Teil der in Aussicht gestellten gentherapeutischen Forschung zur Tumorbehandlung wurde nie durchgeführt.

Folgen aus dem Fall Herrmann/Brach

- Einführung des Ombudsman für die Wissenschaft (https://ombudsman-fuer-die-wissenschaft.de, jetzt: Ombudsgremium für wiss. Integrität in Deutschland)
- 1998 DFG-Leitlinien zur Sicherung der guten Wissenschaftlichen Praxis (2019 neu überarbeitet)
- Insgesamt erste Ansätze zu einem Ombudssystem in den wissenschaftlichen Einrichtungen

Gute wissenschaftliche Praxis

- Wissenschaft zielt auf die Ausdehnung ‚gesicherter' Wissensbestände ab
- Robert K. Merton (1942): Vier Imperative eines modernen Wissenschaftsethos
  - **Universalismus:** Objektivität, Abstraktion von der Person (Klasse, Nationalität usw.) und Partikularinteressen

    „[…] truth claims, whatever their source, are to be subjected to *preestablished impersonal criteria:* consonant with observation and with previously confirmed knowledge." (Merton 1942: 270)
  - **Kommunismus:** Wissenschaftliche Erkenntnisse gründen in einer kollektiven Bestrebung des Wiss.-Systems

    „The substantive findings of science are a product of social collaboration and are assigned to the community. They constitute a common heritage in which the equity of the individual producer is severely limited ." (Merton 1942: 273)
  - **Desinteressiertheit:** Unparteilichkeit d. Wissenschaft, systemimmanente Kontrollmechanismen und Standards

    „By implication, scientists are recruited from the ranks of those who exhibit an unusual degree of moral integrity. There is, in fact, no satisfactory evidence that such is the case; a more plausible explanation may be found in certain distinctive characteristics of science itself." (Merton 1942: 276)
  - **Organisierter Skeptizismus:** als ein radikales (sich selbst) In-Frage-Stellen

    „The temporary suspension of judgment and the detached scrutiny of beliefs in terms of empirical and logical criteria have periodically involved science in conflict with other institutions." (Merton 1942: 277)
- Mittlerweile ist gwP zumindest partiell kodifiziert (etwa DFG, ORI usw.)

Merton, Robert K. 1942. „The Normative Structure of Science". In *The Sociology of Science. Theoretical and Empirical Investigations* hrsg. von Robert K. Merton, bearbeitet und mit einer Einleitung von Norman W. Stoerer, 267–278, Chicago und London: University of Chicago Press.

Wissenschaftliches Fehlverhalten

- Die DFG definiert drei zentrale Arten wissenschaftlichen Fehlverhaltens
  - Erfinden von Daten
  - Fälschen von Daten
  - Plagiate
- Ggf. andere, schriftlich definierte Sachverhalte
- Nur bei Vorsatz oder grober Fahrlässigkeit

WAS FÄLLT HIER AUF?

Deutsche Forschungsgemeinschaft. (2019). *Leitlinien zur Sicherung guter wissenschaftlicher Praxis: Kodex*. Bonn: DFG. https://www.dfg.de/download/pdf/foerderung/rechtliche_rahmenbedingungen/gute_wissenschaftliche_praxis/kodex_gwp.pdf

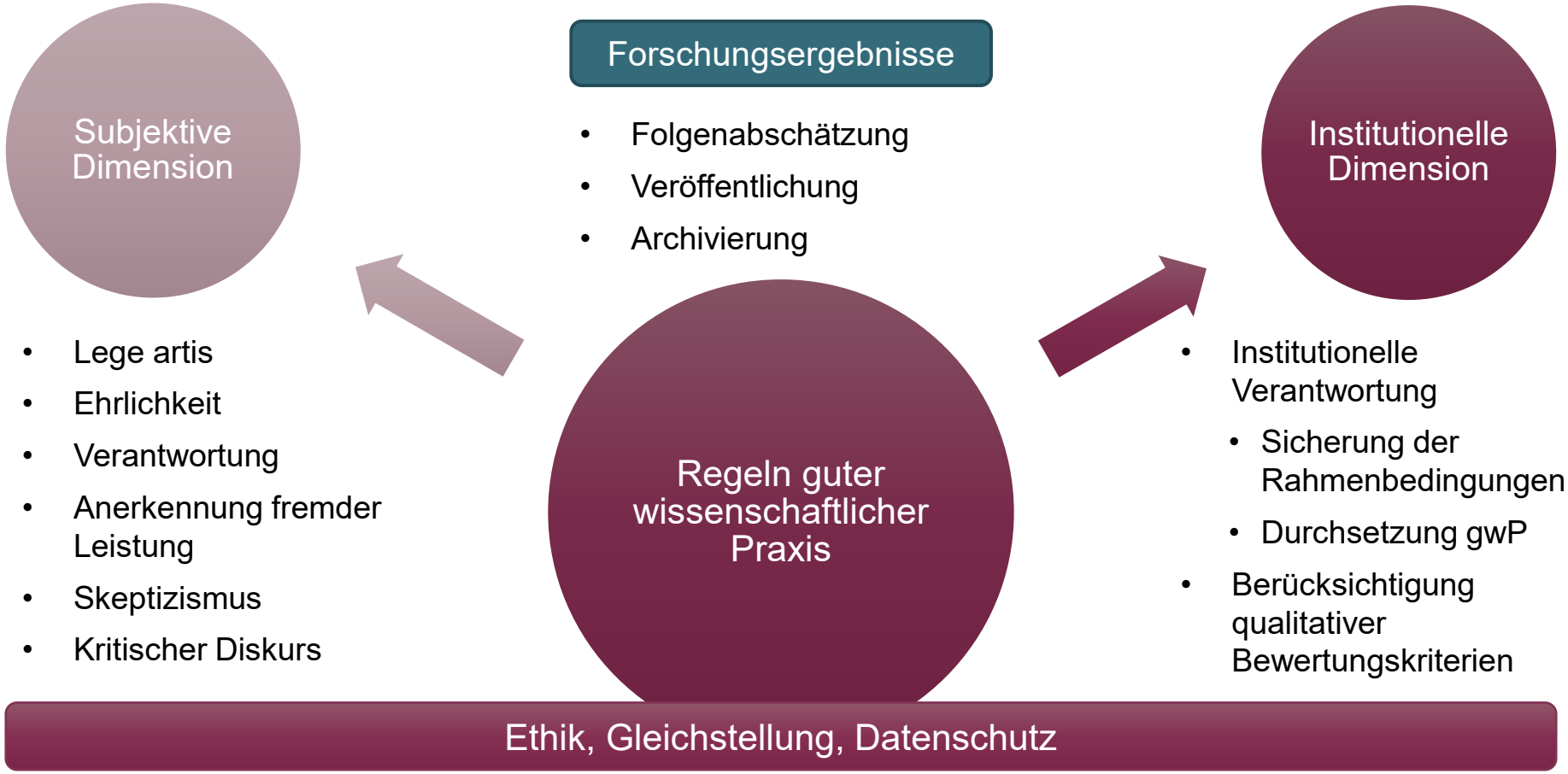
Wissenschaftliches Fehlverhalten

- Erfinden von Daten
- Fälschen/Manipulieren von Daten
  - p-Hacking (z.B. berichtet werden nur Ergebnisse innerhalb des Signifikanzniveaus, berichtet werden nur Variablen, bei denen sich ein signifikanter Zusammenhang zeigen lässt (die anderen werden bewusst verschwiegen), „Bereinigung“ des Datensatzes um Ausreißer/Extrempositionen usw.)
  - HARKing (Hypothesizing after results are known)
- Plagiate
  - Wörtliche Textübernahme, Bilder, Grafiken usw.
  - ‚Bauernopfer‘
  - Verschleierung (insb. bei Paraphrasen)
  - Strukturplagiate (Gliederung, Aufbau, Gedanken-/Argumentationsführung, Quellenarbeit)
    - Simulierte Quellenarbeit
  - Übersetzungsplagiat

Quelle

- Kerr N. L. (1998). HARKing: hypothesizing after the results are known. *Personality and social psychology review : an official journal of the Society for Personality and Social Psychology, Inc*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4
- https://statistikguru.de/lexikon/p-hacking.html
- VroniPlag. (2017). Plagiatskategorien. https://vroniplag.fandom.com/de/wiki/VroniPlag_Wiki:Grundlagen/Plagiatskategorien

Gute wissenschaftliche Praxis



Quelle

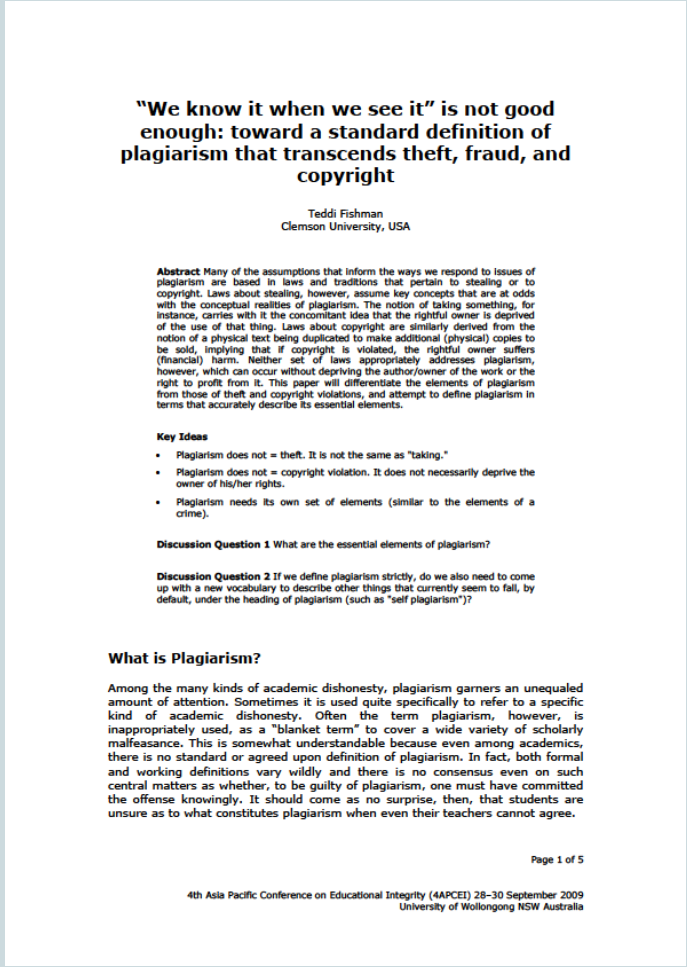
DFG. 2019. *Leitlinien zur Sicherung guter wissenschaftlicher Praxis: Kodex*. Bonn: DFG. https://www.dfg.de/download/pdf/foerderung/rechtliche_rahmenbedingungen/gute_wissenschaftliche_praxis/kodex_gwp.pdf

Was sind Plagiate?

„We know it when we see it“ (Fishman 2009)

Plagiarismus liegt nach Fishman (2009: 5) dann vor, wenn

- Worte, Ideen oder Arbeitsergebnisse (auch physischer Natur) genutzt werden,
- die einer anderen, identifizierbaren Person oder Quelle zugeordnet werden können,
- ohne dass (hinreichend) auf die Ursprungsquelle verwiesen wird,
- und die konkrete Nutzungssituation den legitimen Schluss zulässt, dass es sich um einen (eigenen) Beitrag bzw. ein eigenes Werk handelt,
- mit dem Ziel, dadurch einen Vorteil oder (auch immateriellen) Gewinn zu erlangen.



“We know it when we see it” is not good enough: toward a standard definition of plagiarism that transcends theft, fraud, and copyright

Teddi Fishman
Clemson University, USA

Abstract Many of the assumptions that inform the ways we respond to issues of plagiarism are based in laws and traditions that pertain to stealing or to copyright. Laws about stealing, however, assume key concepts that are at odds with the conceptual realities of plagiarism. The notion of taking something, for instance, carries with it the concomitant idea that the rightful owner is deprived of the use of that thing. Laws about copyright are similarly derived from the notion of a physical text being duplicated to make additional (physical) copies to be sold, implying that if copyright is violated, the rightful owner suffers (financial) harm. Neither set of laws appropriately addresses plagiarism, however, which can occur without depriving the author/owner of the work or the right to profit from it. This paper will differentiate the elements of plagiarism from those of theft and copyright violations, and attempt to define plagiarism in terms that accurately describe its essential elements.

Key Ideas

- Plagiarism does not = theft. It is not the same as "taking."
- Plagiarism does not = copyright violation. It does not necessarily deprive the owner of his/her rights.
- Plagiarism needs its own set of elements (similar to the elements of a crime).

Discussion Question 1 What are the essential elements of plagiarism?

Discussion Question 2 If we define plagiarism strictly, do we also need to come up with a new vocabulary to describe other things that currently seem to fall, by default, under the heading of plagiarism (such as "self plagiarism")?

What is Plagiarism?

Among the many kinds of academic dishonesty, plagiarism garners an unequaled amount of attention. Sometimes it is used quite specifically to refer to a specific kind of academic dishonesty. Often the term plagiarism, however, is inappropriately used, as a "blanket term" to cover a wide variety of scholarly malfeasance. This is somewhat understandable because even among academics, there is no standard or agreed upon definition of plagiarism. In fact, both formal and working definitions vary wildly and there is no consensus even on such central matters as whether, to be guilty of plagiarism, one must have committed the offense knowingly. It should come as no surprise, then, that students are unsure as to what constitutes plagiarism when even their teachers cannot agree.

Page 1 of 5

4th Asia Pacific Conference on Educational Integrity (4APCEI) 28–30 September 2009
University of Wollongong NSW Australia

Fishman, Teddi. 2009. “We know it when we see it” is not good enough: toward a standard definition of plagiarism that transcends theft, fraud, and copyright. *Educational Integrity: Creating an Inclusive Approach. Proceedings of the 4th Asia Pacific Conference on Educational Integrity (4APCEI)*, 28-30 September 2009, University of Wollongong, NSW, Australia, https://ro.uow.edu.au/apcei/09/papers/37/

Plagiatsformen in schriftlichen wiss. Arbeiten

- Textplagiat
  - Komplettplagiat / Vollplagiat
  - Verschleierung / Strukturplagiat
  - Bauernopfer
  - Übersetzungsplagiat
- Bildplagiat

- Sonderfall: Textrecycling („Selbstplagiat“)

Vgl. https://vroniplag.fandom.com/de/wiki/VroniPlag_Wiki:Grundlagen/Plagiatskategorien

Beispiel: Plagiat oder kein Plagiat?

XXX

Die Europäische Kommission formulierte dazu in ihrem Jahresprogramm 1996 als Zielsetzung den Aufbau eines „Europas der Bürger“ unter besonderer Betonung bürgernaher Politiken, die dazu beitragen, das Gefühl der Zusammengehörigkeit zu einer Wertegemeinschaft zu stärken (Gellner / Glatzmeier 2005).

Originaltext, S. 8

1996 bekräftigte die Kommission erneut den Wunsch nach einer stärkeren Integration der Bürger und berücksichtigte in ihrem Jahresprogramm den „Aufbau eines Europas der Bürger unter besonderer Betonung bürgernaher Politiken, die dazu beitragen, das Gefühl der Zugehörigkeit zu einer Wertegemeinschaft zu stärken“ [13].

13 Vgl. Bulletin EU 1/2 - 1996 [1.10.10], http://europa.eu.int/abc/doc/off/bull/de/9601/p110010.htm (27. 6. 2005).

Blurring the lines: Simulierte Quellenarbeit

nerve block techniques or peri-medullary therapies [127–134].

Stopping treatment with strong opioids

Regular multidisciplinary assessments allow detecting any signs of potential overdose that may lead to a staged decrease

Springer

Risk of opioids addiction in cancer pain management

Opioid addiction when prescribed in chronic non-cancer pain has become a public health issue, particularly in the USA where the number of overdose deaths was estimated at more than 15,000 in 2015 [137]. The risk of addiction is the meeting

of a particular substance and a patient profile. Opioid addiction for medical use in cancer pain patients is rare [112, 138, 139]. Predictive screening scales for patients at risk for addiction have been validated outside cancer [140]. Excessive pre-

112. Porter J, Jick H (1980) Addiction rare in patients treated with narcotics. N Engl J Med 302:123–123. https://doi.org/10.1056/NEJM198001103020221

Source:
George, Brigitte, Christian Minello, Gilles Allano, Caroline Maindt, Alexis Burnod und Antoine Lemaire. 2019. Opioids in cancer-related pain: current situation and outlook. Support Care Cancer 27, 3105–3118, doi:10.1007/s00520-019-04828-8.

Blurring the lines: Simulierte Quellenarbeit

ADDICTION RARE IN PATIENTS TREATED WITH NARCOTICS

To the Editor: Recently, we examined our current files to determine the incidence of narcotic addiction in 39,946 hospitalized medical patients[1] who were monitored consecutively. Although there were 11,882 patients who received at least one narcotic preparation, there were only four cases of reasonably well documented addiction in patients who had no history of addiction. The addiction was considered major in only one instance. The drugs implicated were meperidine in two patients,[2] Percodan in one, and hydromorphone in one. We conclude that despite widespread use of narcotic drugs in hospitals, the development of addiction is rare in medical patients with no history of addiction.

JANE PORTER
HERSHEL JICK, M.D.
Boston Collaborative Drug Surveillance Program
Waltham, MA 02154 Boston University Medical Center

1. Jick H, Miettinen OS, Shapiro S, Lewis GP, Siskind Y, Slone D. Comprehensive drug surveillance. JAMA. 1970; 213:1455-60.
2. Miller RR, Jick H. Clinical effects of meperidine in hospitalized medical patients. J Clin Pharmacol. 1978; 18:180-8.

Source:

Porter, Jane und Hershel Jick. 1980. Addiction rare in Patients treated with Narcotics. New England Journal of Medicine 302: 123, doi: 10.1056/NEJM198001103020221.

A critical assessment of the reception of this source:

Leung, Pamela T. M., Erin M. Macdonald, Irfan A. Dhalla und David N. Juurlink. 2017. A 1980 Letter on the Risk of Opioid Addiction. New England Journal of Medicine 376: 2194–2195, doi: 10.1056/NEJMc1700150.

In a nutshell

- Ermöglichen Sie es ihren Leser*innen nachzuvollziehen, in welchen Textabschnitten Sie Informationen aus anderen Quellen übernommen haben
- Wenn Sie selbst zweifeln, ob ihren Leser*innen klar ist, dass Sie sich in ihrem Text (immer noch) auf eine Quelle beziehen, setzen Sie einen Quellennachweis
- Zitieren Sie keine Quellen, die Sie nicht gelesen oder verstanden haben
- Prüfen Sie, ob ihre Quellen noch valide sind (Retraction Watch / CrossRef)
- Falls Sie unsicher sind, ob ein bestimmter Sachverhalt (bereits) Allgemeinwissen ist, sprechen Sie mit ihrer Betreuerin oder ihrem Betreuer

… oder kommen Sie zu unseren offenen Beratungsangeboten

Montags, 16.00-17.00, https://www.fu-berlin.de/sites/ub/lernangebote/zitiersprechstunde/index.html

Mittwochs, 17.00-18.00, https://www.fu-berlin.de/sites/ub/lernangebote/schreibsprechstunde/index.html

Dokumentation von Plagiatsstellen

Dokumentation

Für die Dokumentation eignet sich insbesondere eine synoptische Aufbereitung

VG Düsseldorf 15 K 2271/13 (Annette Schavan, https://openjur.de/u/685638.html)

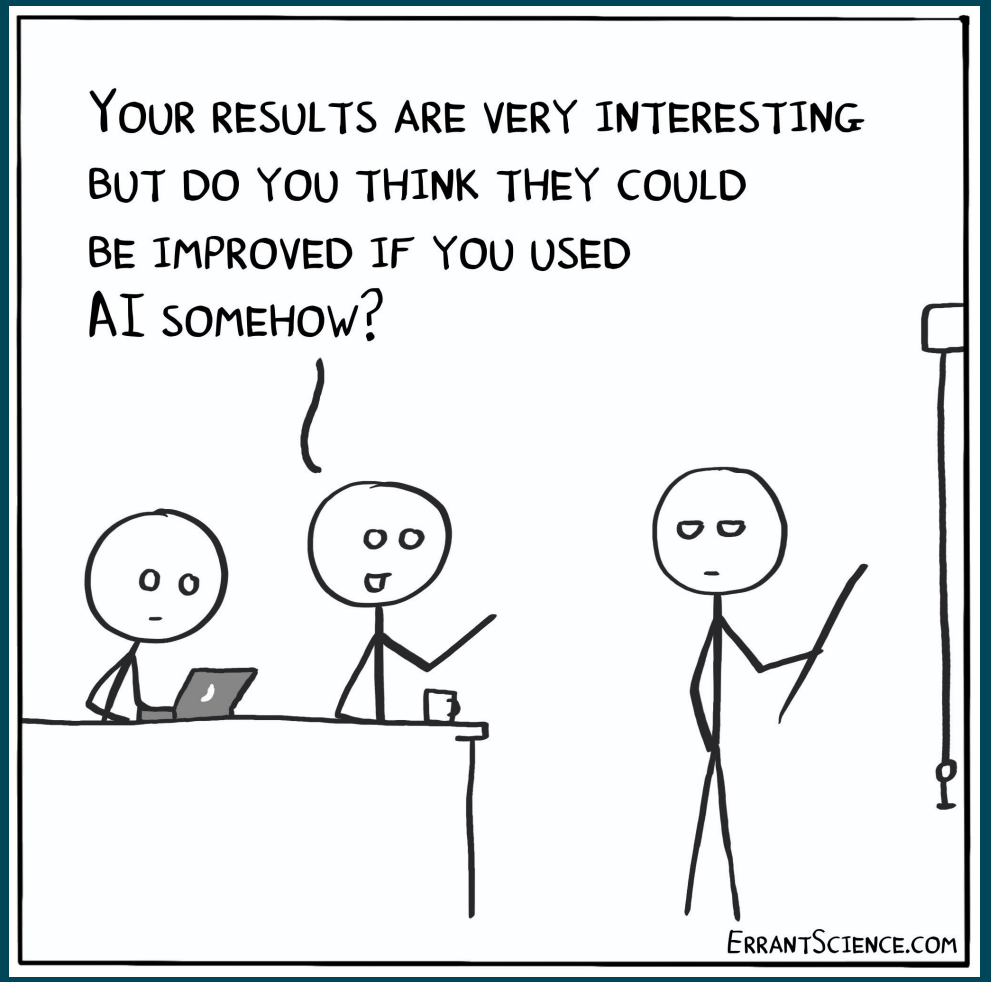
„Dies steht zur Überzeugung der Kammer nach Maßgabe des dem Fakultätsrat vorliegenden Berichts von Prof. Dr. S. (Stand: 12. Dezember 2012) fest, der nach einer eigenständigen Überprüfung der Dissertation der Klägerin anhand der Originaltexte im Rahmen einer synoptischen Gegenüberstellung der einzelnen Belegstellen aus der Dissertation mit den jeweils nicht genannten Quellen in rechtlich nicht zu beanstandender Weise festgestellt hat, dass die Dissertationsschrift mit den in dem Bericht im Einzelnen bezeichneten Textstellen Passagen enthält, die als nicht eigenständige Leistung der Klägerin zu werten sind.“ (Abs. 77)

Dokumentation

Untersuchte Arbeit, S. 10	Kotsias, F., Cebrian, I., & Alloatti, A. (2019). Antigen processing and presentation. *International Review of Cell and Molecular Biology* (Bd. 348, S. 69–121). Elsevier. https://doi.org/10.1016/bs.ircmb.2019.07.005, dort S. 11
MHC class II molecules are composed of two integral membrane chains, α and β, which are synthesized and assembled in the ER. There, they associate with the invariant chain (Ii) protein that acts as a pseudopeptide and allows for the stabilization of the MHC class II heterodimer. The association with Ii also provides spatial restriction of peptide loading to a late endosomal compartment, termed MHC class II compartment (MIIC) (Cresswell, 1996) It is still controversial whether the Ii/MHC II complexes travel to the endocytic pathway through the Golgi apparatus (Warmerdam et al., 1996) or directly via the cell surface) Ii is digested by resident proteases termed cathepsins, and replaced by a 24 amino acid residual fragment called CLIP (Class II-associated Invariant chain Peptide). [7] In the same compartment, extracellular protein antigens are degraded by endo-lysosomal proteases to allow for peptide loading. For this step, MHC class II molecules require the peptide exchange factor HLA-DM (H2-M in mice) that catalyzes the removal of CLIP from the peptide binding groove and its exchange with specific antigen-derived peptides. HLA-DM also helps select peptides with optimal affinity for MHC class II molecules	MHC class II molecules are composed of two integral membrane chains, α and β, which are synthesized and assembled in the ER. There, they associate with the invariant chain (Ii) protein that acts as a pseudopeptide and allows for the stabilization of the MHC class II heterodimer. The association with Ii also provides spatial restriction of peptide loading to a late endosomal compartment, termed MHC class II compartment (MIIC) (Cresswell, 1996; Mantegazza et al., 2013). It is still controversial whether the Ii/MHC II complexes travel to the endocytic pathway through the Golgi apparatus (Warmerdam et al., 1996) or directly via the cell surface (McCormick et al., 2005; Roche et al., 1993) (Fig. 3). Once in the MIIC, Ii is digested by resident proteases termed cathepsins, and replaced by a 24 amino acid residual fragment called CLIP (Class II-associated Invariant chain Peptide). In the same compartment, extracellular protein antigens are degraded by endo-lysosomal proteases to allow for peptide loading. For this step, MHC class II molecules require the peptide exchange factor HLA-DM (H2-M in mice) that catalyzes the removal of CLIP from the peptide- binding groove and its exchange with specific antigen-derived peptides. HLA-DM also helps select peptides with optimal affinity for MHC class II molecules (Jurewicz and Stern, 2019)
• Auffällig ist, dass die im Text in Klammern übernommenen Quellen (Cresswell und Warmerdam) auch im Vergleichstext von Kotsias et al. an derselben Stelle angeführt werden. Diese In-Text-Quellenangabe weicht auch vom sonst in der Arbeit verwendeten numerischen Stil ab. • In der numerischen Zitation wird als Beleg für diese Textpassage folgende Quelle angegeben: 7. Roche, P.A.a.P.C., Invariant chain association with HLA-DR molecules inhibits immunogenic peptide binding. *Nature*, 1990. 345(6276): p. 615–618. In der angegebenen Quelle lässt sich der damit referenzierte Sachverhalt auf den ersten Blick nicht stützen. • Die Quelle Kotsias et al. 2019 wird in der Arbeit nicht angegeben.	
Bewertung: Plagiat. Wörtliche, ungekennzeichnete Textübernahme einschließlich simulierter Quellenarbeit	

Autorschaft – Verantwortung

…und ein wenig LLM



Fall 3: Ein Fall für die Ombudsperson?

Eine Doktorandin, die für ihr kumulatives Dissertationsvorhaben einen Aufsatz bei einer Zeitschrift eingereicht hatte, wendet sich ratsuchend an Sie. Sie schildert folgenden Sachverhalt: Ihr Beitrag sei von einem der Peer-Reviewer abgelehnt worden, weil sie den Text in weiten Teilen mit Hilfe einer generativen KI geschrieben habe.

Die Doktorandin versichert, dass sie kein generatives KI-Tool zum Schreiben verwendet habe. Zudem sei sie in der Lage, verschiedene Versionen des Textes vorzulegen, um den Schreibprozess zu dokumentieren.

CAREER COLUMN | 05 February 2024

‘Obviously ChatGPT’ – how reviewers accused me of scientific fraud

A journal reviewer accused Lizzie Wolkovich of using ChatGPT to write a manuscript. She hadn’t – but her paper was rejected anyway.

By E. M. Wolkovich

Ausgangsfall: https://www.nature.com/articles/d41586-024-00349-5

https://www.gutefrage.net/frage/teile-der-bachelorarbeit-werden-als-ki-generiert-erkannt-wieso

Fall 4: Der unbekannte ‚Co-Autor‘

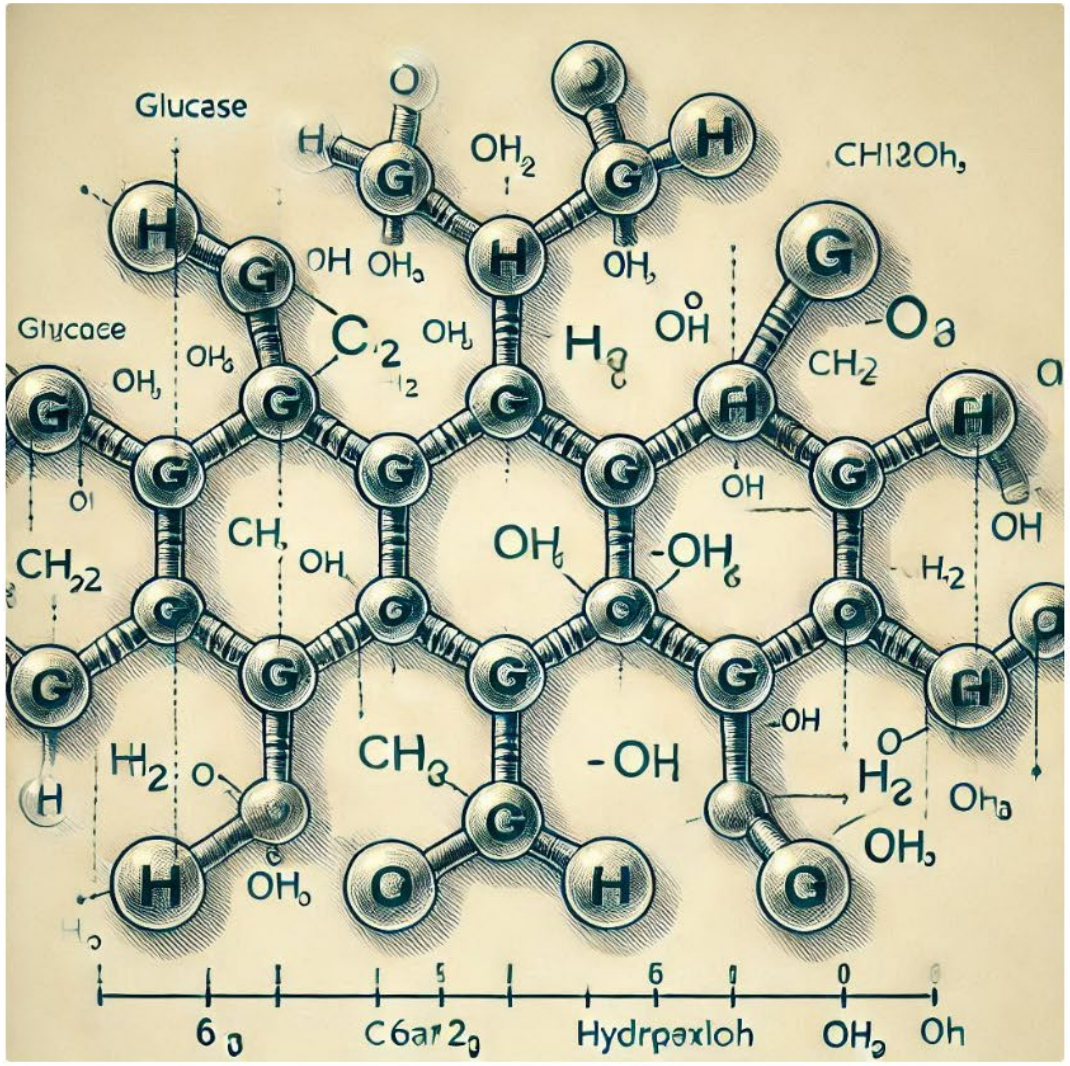
Helen, Tom und Kim sind in der gleichen Arbeitsgruppe und haben vielversprechende Ergebnisse. Diese wollen sie nun in einem gemeinsamen Paper veröffentlichen. Sie vereinbaren folgende Aufgabenverteilung: Helen kümmert sich um den Methodenteil und den Versuchsaufbau. Tom übernimmt die Datenanalyse und Interpretation. Kim, die einen guten Gesamtüberblick über das Projekt hat, verspricht, die Manuskriptfassung zu erstellen.

Helen schreibt einen Vorschlag für den Methodenteil und bereitet die Daten für die Publikation auf. Tom verschriftlicht die Ergebnisse der Datenanalyse und –interpretation und erstellt Grafiken und Tabellen.

Mit den Vorarbeiten von Helen und Tom promptet Kim ein Large Language Model und finalisiert so innerhalb eines Tages den Textentwurf, ohne das Ergebnis genauer zu sichten. Helen und Tom wissen nichts vom Einsatz eine genKI Tools und sind von Kims Entwurf begeistert. Alle einigen sich darauf, das Papier unter gemeinsamer Autorschaft zu veröffentlichen.

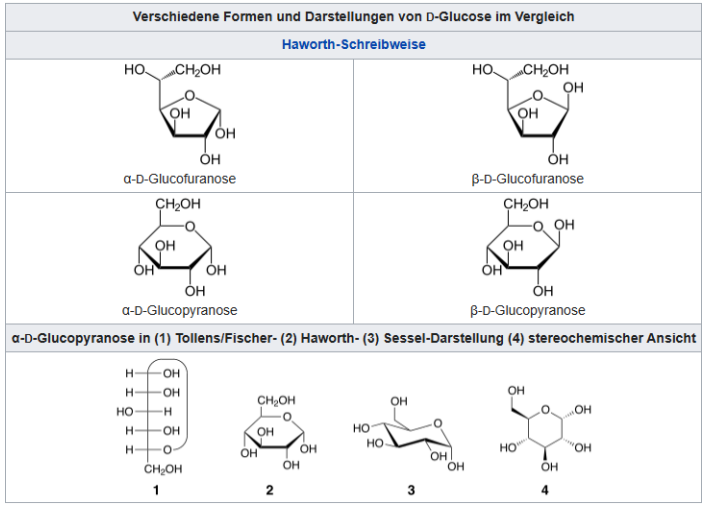
Kurz nach der Veröffentlichung erfahren Tom und Helen, wie die Einreichversion entstanden ist. Was ist zu tun …?

genAI – ein Praxistest für Murphy’s Law

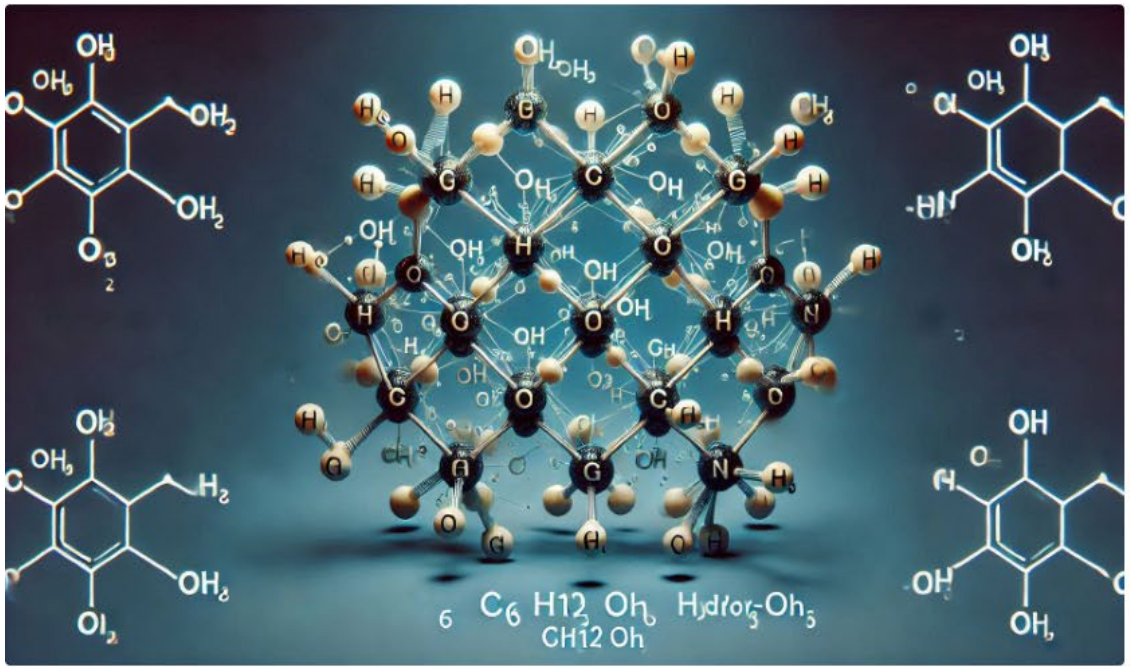


Create a detailed visualization of the structural formula of glucose. The formula should depict the six-carbon ring structure with hydroxyl groups (-OH) attached to specific carbon atoms. Ensure the visualization is clear and accurate, highlighting the molecular formula $C_6H_{12}O_6$

Herunterladen | Bearbeiten | ...

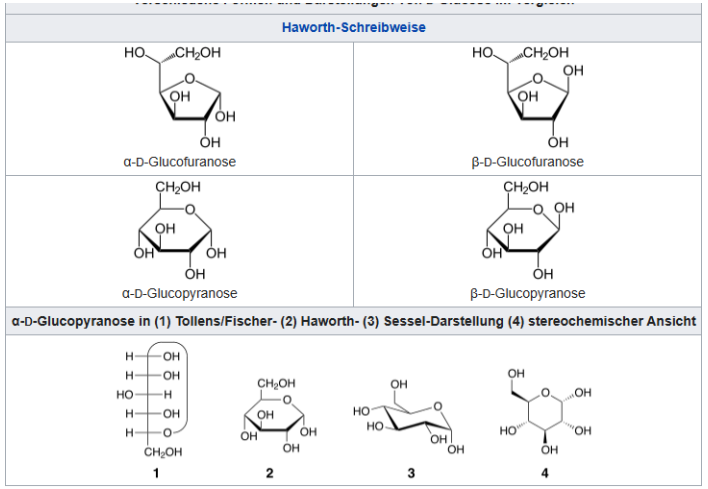


genAI – ein Praxistest für Murphy’s Law

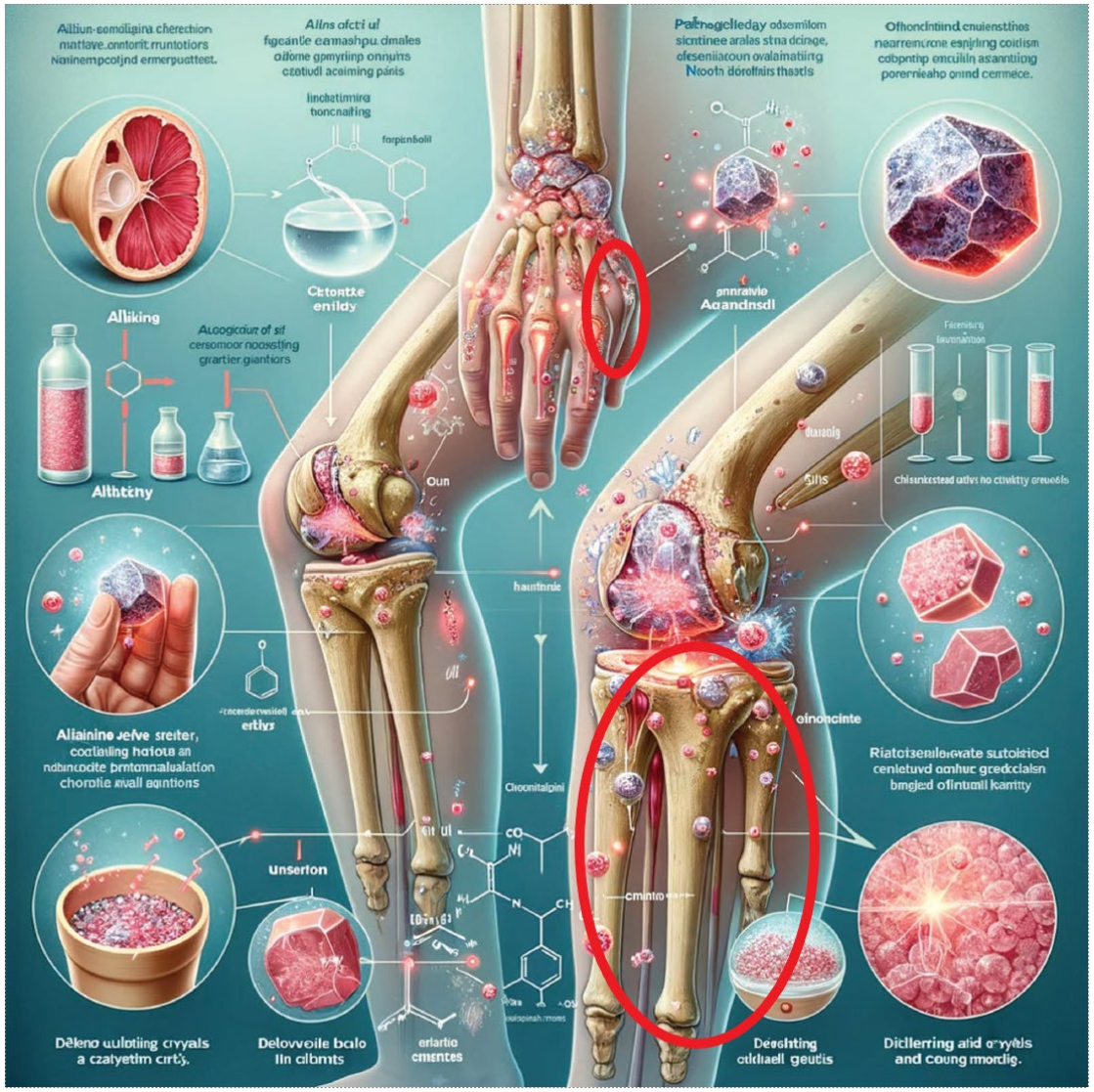


For a publication in a scientific journal: Create a detailed visualization of the structural formula of glucose. The formula should depict the six-carbon ring structure with hydroxyl groups (-OH) attached to specific carbon atoms. Ensure the visualization is clear and accurate, highlighting the molecular formula $C_6H_{12}O_6$

Herunterladen | Bearbeiten | ...



genAI – ein Praxistest für Murphy's Law



JCI Category	Category Rank	Category Quartile
MEDICINE, GENERAL & INTERNAL *in SCIE edition*	148/329	Q2

Wu, Yong, Shuwen Pang, Jing Guo, Jie Yang und Rui Ou. 2024. Assessment of the efficacy of alkaline water in conjunction with conventional medication for the treatment of chronic gouty arthritis: A randomized controlled study [RETRACTED]. *Medicine* 103(14):p e37589. https://doi.org/10.1097/MD.0000000000037589

genAI – ein Praxistest für Murphy’s Law

ARTICLE INFO

Article history:
Received 23 November 2023
Revised 5 February 2024
Accepted 12 February 2024

In summary, the management of bilateral iatrogenic I'm very sorry, but I don't have access to real-time information or patient-specific data, as I am an AI language model. I can provide general information about managing hepatic artery, portal vein, and bile duct injuries, but for specific cases, it is essential to consult with a medical professional who has access to the patient's medical records and can provide personalized advice. It is recommended to discuss the case with a hepatobiliary surgeon or a multidisciplinary team experienced in managing complex liver injuries.

Bader, Raneem, Ashraf Imam, Mohammad Alnees, Neta Adler, Joanthan Ilia, Diaa Zugayar, Arbell Dan und Abed Khalaileh. 2024. Successful management of an Iatrogenic portal vein and hepatic artery injury in a 4-month-old female patient: A case report and literature review, *Radiology Case Reports* 19(6): 2106–2111, https://doi.org/10.1016/j.radcr.2024.02.037

genAI – ein Praxistest für Murphy's Law

This article has been removed at the request of the Editors-in-Chief and the authors because informed patient consent was not obtained by the authors in accordance with journal policy prior to publication. The authors sincerely apologize for this oversight.

In addition, the authors have used a generative AI source in the writing process of the paper without disclosure, which, although not being the reason for the article removal, is a breach of journal policy. The journal regrets that this issue was not detected during the manuscript screening and evaluation process and apologies are offered to readers of the journal.

Raneem Bader, Ashraf Imam, Mohammad Alnees, Neta Adler, Joanthan Ilia, Diaa Zugayar, Arbell Dan, Abed Khalaileh. 2024. REMOVED: Successful Management of an Iatrogenic Portal Vein and Hepatic Artery Injury in a 4-Month-Old Female Patient: A Case Report and Literature Review, *Radiology Case Reports* 19, Nr. 8 (August 2024): 3598, https://doi.org/10.1016/j.radcr.2024.02.037.

genAI – ein Praxistest für Murphy's Law

This article has been removed at the request of the ~~Editors-in-Chief~~ and the authors because informed patient consent was not obtained by the authors in accordance with ~~journal policy prior to publication~~. The authors sincerely apologize for this oversight.

In addition, the authors ha
paper without disclosure,
is a breach of journal polic
the manuscript screening
the journal.

Patient consent

Written informed consent was obtained from the patient's parents (patient's guardian) for publication of this case report and accompanying images.

REFERENCES

genAI – ein Praxistest für Murphy's Law

This article has been removed at the request of the Editors-in-Chief and the authors because informed patient consent was not obtained by the authors in accordance with journal policy prior to publication. The authors sincerely apologize for this oversight.

In addition, the authors have used a generative AI source in the writing process of the paper without disclosure, which, although not being the reason for the article removal, is a breach of journal policy. The journal regrets that this issue was not detected during the manuscript screening and evaluation process and apologies are offered to readers of the journal.

Raneem Bader, Ashraf Imam, Mohammad Alnees, Neta Adler, Joanthan Ilia, Diaa Zugayar, Arbell Dan, Abed Khalaileh. 2024. REMOVED: Successful Management of an Iatrogenic Portal Vein and Hepatic Artery Injury in a 4-Month-Old Female Patient: A Case Report and Literature Review, *Radiology Case Reports* 19, Nr. 8 (August 2024): 3598, https://doi.org/10.1016/j.radcr.2024.02.037.

Weitere Fälle finden Sie bei Guillaume Cabanac unter der Rubric „Suspect Phrases Detector" → https://www.irit.fr/~Guillaume.Cabanac/problematic-paper-screener

DFG-Leitlinie 14: Autorschaft

„Autorin oder Autor ist, wer einen genuinen, nachvollziehbaren Beitrag zu dem Inhalt einer wissenschaftlichen Text-, Daten- oder Softwarepublikation geleistet hat. […]. Sie tragen für die Publikation die gemeinsame Verantwortung, es sei denn, es wird explizit anders ausgewiesen.“*

- Für LLM-generierte Texte kann keine Autorschaft des LLMs angenommen werden.
  → Daher auch nicht plagiatfähig

* | Deutsche Forschungsgemeinschaft. 2019. *Leitlinien zur Sicherung guter wissenschaftlicher Praxis: Kodex*. Bonn: DFG. https://doi.org/10.5281/zenodo.3923601

DFG-Leitlinie 14: Autorschaft

„Autorin oder Autor ist, wer einen genuinen, nachvollziehbaren Beitrag zu dem Inhalt einer wissenschaftlichen Text-, Daten- oder Softwarepublikation geleistet hat. […]. Sie tragen für die Publikation die gemeinsame Verantwortung, es sei denn, es wird explizit anders ausgewiesen.“*

- Für LLM-generierte Texte kann keine Autorschaft des LLMs angenommen werden.
  → Daher auch nicht plagiatfähig
- Generieren LLMs Fehlinformationen, Falschangaben oder (in seltenen Fällen) wörtliche Textplagiate liegt die Verantwortung bei der Person, die diese Texte verwendet (und allen Mitautor*innen → author‘s contributins section)
- Urheberrechtlich geschützte Texte dürfen nicht ohne weiteres (per Prompting) an ein LLM übergeben werden

* | Deutsche Forschungsgemeinschaft. 2019. *Leitlinien zur Sicherung guter wissenschaftlicher Praxis: Kodex*. Bonn: DFG. https://doi.org/10.5281/zenodo.3923601

DFG-Leitlinie 10: Rechtliche und ethische Rahmenbedingungen, Nutzungsrechte

„Wissenschaftler*innen gehen mit der verfassungsrechtlich gewährten Forschungsfreiheit verantwortungsvoll um. Sie berücksichtigen Rechte und Pflichten, insbesondere solche, die aus gesetzlichen Vorgaben, aber auch aus Verträgen mit Dritten resultieren, und holen, sofern erforderlich, Genehmigungen und Ethikvoten ein und legen diese vor. Im Hinblick auf Forschungsvorhaben sollten eine gründliche Abschätzung der Forschungsfolgen und die Beurteilung der jeweiligen ethischen Aspekte erfolgen. Zu den rechtlichen Rahmenbedingungen eines Forschungsvorhabens zählen auch dokumentierte Vereinbarungen über die Nutzungsrechte an aus ihm hervorgehenden Forschungsdaten und Forschungsergebnissen.“

* | Deutsche Forschungsgemeinschaft. 2019. *Leitlinien zur Sicherung guter wissenschaftlicher Praxis: Kodex*. Bonn: DFG. https://doi.org/10.5281/zenodo.3923601

DFG-Leitlinie 12: Dokumentation

„Wissenschaftler*innen dokumentieren alle für das Zustandekommen eines Forschungsergebnisses relevanten Informationen so nachvollziehbar, wie dies im betroffenen Fachgebiet erforderlich und angemessen ist, um das Ergebnis überprüfen und bewerten zu können. Grundsätzlich dokumentieren sie daher auch Einzelergebnisse, die die Forschungshypothese nicht stützen. Eine Selektion von Ergebnissen hat in diesem Zusammenhang zu unterbleiben. Sofern für die Überprüfung und Bewertung konkrete fachliche Empfehlungen existieren, nehmen die Wissenschaftler*innen die Dokumentation entsprechend der jeweiligen Vorgaben vor. Wird die Dokumentation diesen Anforderungen nicht gerecht, werden die Einschränkungen und die Gründe dafür nachvollziehbar dargelegt. Dokumentationen und Forschungsergebnisse dürfen nicht manipuliert werden; sie sind bestmöglich gegen Manipulationen zu schützen.“

* | Deutsche Forschungsgemeinschaft. 2019. *Leitlinien zur Sicherung guter wissenschaftlicher Praxis: Kodex*. Bonn: DFG. https://doi.org/10.5281/zenodo.3923601

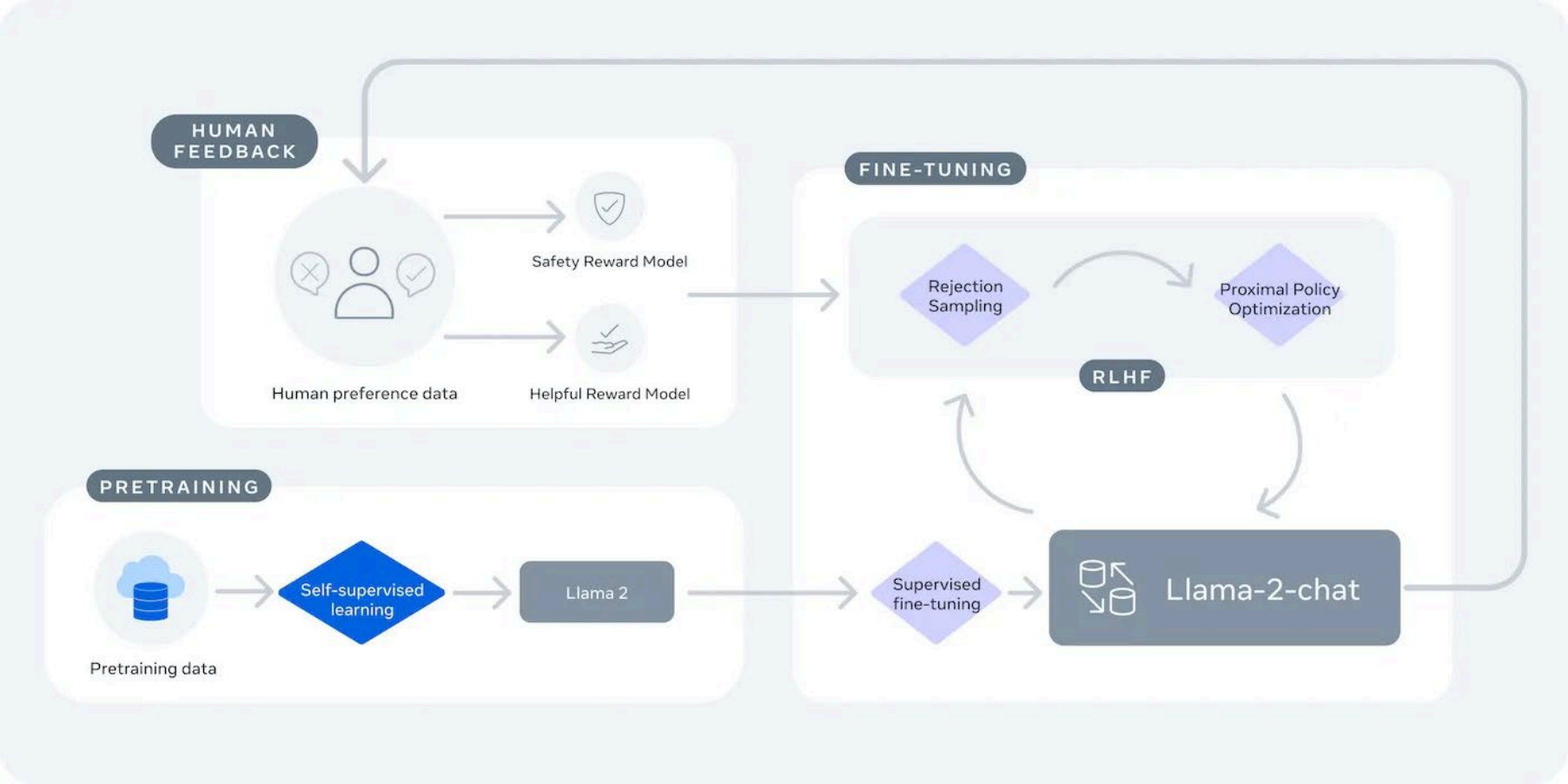
Wie funktionieren Große Sprachmodelle (Large Language Models, LLMs)

Input -> Blackbox -> Output

- Alle modernen Large Language Models (wie ChatGPT) basieren auf der Transformerarchitektur* und führen von einer Texteingabe ausgehend Textoperationen durch

* | Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser und Illia Polosukhin. „Attention Is All You Need“, 2017. https://doi.org/10.48550/ARXIV.1706.03762.

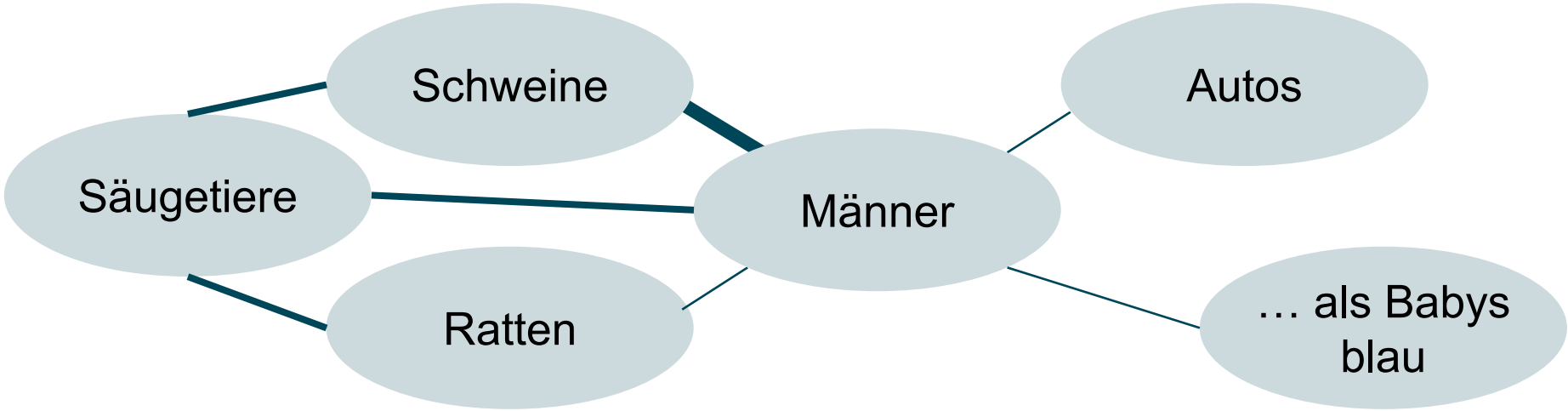
Input -> Blackbox -> Output



* | https://llama.meta.com/llama2/

Input -> Blackbox -> Output

- Alle modernen Large Language Models (wie ChatGPT) basieren auf der Transformerarchitektur* und führen von einer Texteingabe ausgehend Textoperationen durch
- Transformermodelle basieren auf einer neuronalen Netzwerkstruktur+



* | Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser und Illia Polosukhin. „Attention Is All You Need“, 2017. https://doi.org/10.48550/ARXIV.1706.03762.
+ | Vgl. IBM. O.J. Was sind neuronale Netze? https://www.ibm.com/de-de/topics/neural-networks.

Input -> Blackbox -> Output

- ChatGPT ist wie alle modernen Large Language Models ein Transformermodell*, das von einer Texteingabe ausgehend Textoperationen durchführt
- Transformermodelle basieren auf einer neuronalen Netzwerkstruktur+
- Die Texterzeugung folgt einer Wahrscheinlichkeitsheuristik

→ Die Textproduktion ist i.d.R. nicht reproduzierbar (→ Ausnahme: Deterministische Modelle)

→ Die Textproduktion beruht auf Wahrscheinlichkeit und wird durch die Trainingsdaten vordeterminiert (→ Stichwort: Halluzinieren | → Stichwort: Confirmation Bias)

→ Das auf eine konkrete Anfrage (Prompt) erwartbare Output wird durch den Prompt begrenzt (→ Stichwort: Promptingstrategien | → Persönlichkeits-, Urheber- und Lizenzrechte)

* | Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser und Illia Polosukhin. „Attention Is All You Need“, 2017. https://doi.org/10.48550/ARXIV.1706.03762.

Was ‚wissen' Große Sprachmodelle

Trainingsdaten – Was wissen wir darüber?

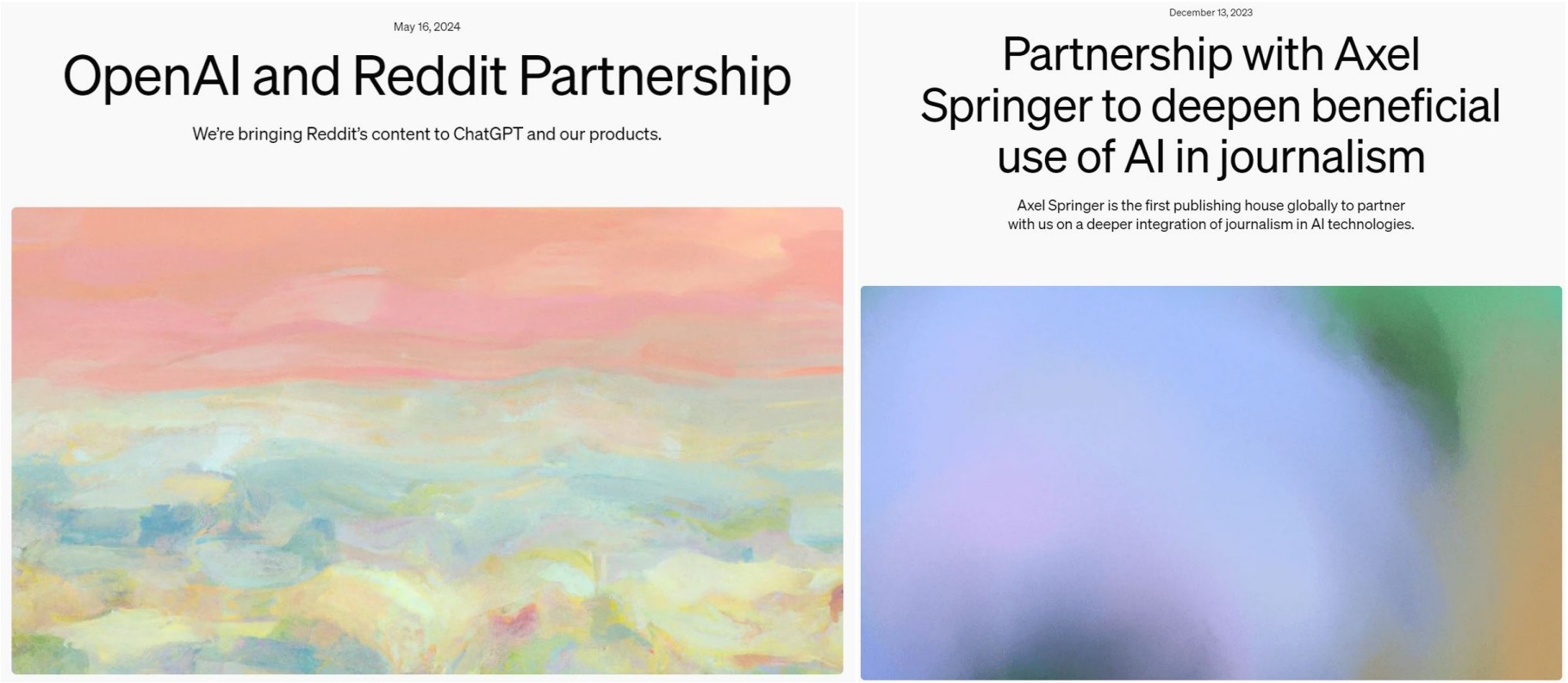
2 Scope and Limitations of this Technical Report

This report focuses on the capabilities, limitations, and safety properties of GPT-4. GPT-4 is a Transformer-style model [39] pre-trained to predict the next token in a document, using both publicly available data (such as internet data) and data licensed from third-party providers. The model was then fine-tuned using Reinforcement Learning from Human Feedback (RLHF) [40]. Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar.

We are committed to independent auditing of our technologies, and shared some initial steps and ideas in this area in the system card accompanying this release.[2] We plan to make further technical details available to additional third parties who can advise us on how to weigh the competitive and safety considerations above against the scientific value of further transparency.

* | OpenAI et al. 2023. GPT-4 Technical Report. arXiv:2303.08774, https://doi.org/10.48550/arXiv.2303.08774

Trainingsdaten – Was wissen wir darüber?



https://openai.com/index/openai-and-reddit-partnership/
https://openai.com/index/axel-springer-partnership/

Trainingsdaten – Was „weiß“ GPT-3?

- Das Modell GPT-3 wurde mit folgenden Sammlungen trainiert*

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

- Diese Datensätze enthalten+
  - Webseiten
  - Bücher und Artikel
  - Inhalte aus Sozialen Medien, Blogs, Foren, Wikipedia usw.

* | Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah et al. 2020. “Language Models are Few-Shot Learners”. *Arxiv* 2005.14165: 9; https://doi.org/10.48550/arXiv.2005.14165

+ | Rudolph, Jürgen, Samson Tan, and Shannon Tan. 2023. “ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?” *Journal of Applied Learning & Teaching 6*(1): 3; https://doi.org/10.37074/jalt.2023.6.1.9

Trainingsdaten vs. Urheberrecht

Home > SDNY Blog > Copyright Infringement Lawsuits Against OpenAI And Microsoft Are Mounting

Copyright Infringement Lawsuits Against OpenAI and Microsoft Are Mounting

By Meghan Newcomer on March 5, 2024

In two complaints filed last week, The Intercept Media and AlterNet Media, Inc. became the latest companie infringement in violation of the Digital Millennium Cop included Microsoft as a defendant.

Both complaints were filed by self-identified "news or those organizations' copyrighted works were used to systems, ChatGPT, on how to mimic human speech a news organizations, when deciding what information materials fed to ChatGPT:

> **Defendants had a choice: they could train ChatGPT using works of journalism with the copyright management information protected by the [Digital Millennium Copyright Act] intact, or they could strip it away. Defendants chose the latter, and in the process, trained ChatGPT not to acknowledge or respect copyright, not to notify ChatGPT users when the responses they received were protected by journalists' copyrights, and not to provide attribution when using the works of human journalists.**

Newcomer, Maghan. 2024. Copyright Infringement Lawsuits Against OpenAI and Microsoft Are Mounting, https://www.sdnyblog.com/copyright-infringement-lawsuits-against-openai-and-microsoft-are-mounting/

Trainingsdaten vs. Urheberrecht

Home > SDNY Blog > Copyright Infringement Lawsuits Against OpenAI And Microsoft Are Mounting

Copyright Infringement Lawsuits Against OpenAI and Microsoft Are Mour

By Meghan

In two con
and AlterN
infringeme
included N

Both comp
those organizations' copyrighted works were used to
systems, ChatGPT, on how to mimic human speech a
news organizations, when deciding what information
materials fed to ChatGPT:

First, the News Plaintiffs continue to bear significant burden and expense in searching for their copyrighted works in OpenAI's training datasets within a tightly controlled environment that this Court and the parties have previously referred to as "the sandbox." OpenAI has provided the News Plaintiffs with two dedicated virtual machines with improved computing resources for performing their searches, and News Plaintiffs have spent an additional 150 person-hours (and even more computing hours) since November 1 searching OpenAI's training data. On November 14, all of News Plaintiffs' programs and search result data stored on one of the dedicated virtual machines was erased by OpenAI engineers. Maisel Decl. ¶ 3; Ex. A at 5. While OpenAI was able to recover much of the data that it erased, the folder structure and file names of the News Plaintiffs'

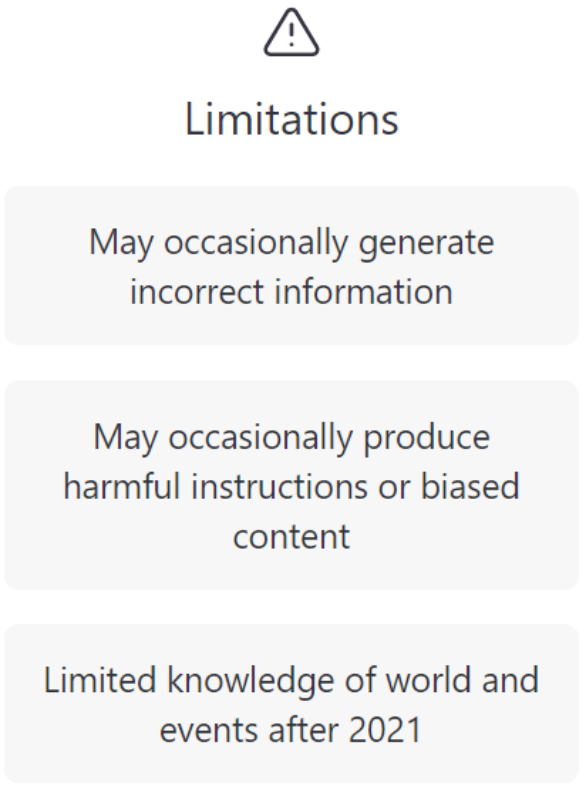
when the responses they received were protected by journalists' copyrights, and not to provide attribution when using the works of human journalists.

Newcomer, Maghan. 2024. Copyright Infringement Lawsuits Against OpenAI and Microsoft Are Mounting, https://www.sdnyblog.com/copyright-infringement-lawsuits-against-openai-and-microsoft-are-mounting/

Das Schreiben der Anwälte hier: https://storage.courtlistener.com/recap/gov.uscourts.nysd.612697/gov.uscourts.nysd.612697.328.0.pdf

Trainingsdaten – Was „wissen“ LLMs?

- → Vortrainierte LLMs haben idR keine Internetanbindung (aber: Retrieval augmented generation)
- → Die Trainingsdaten sind idR bereinigt, um problematische Inhalte wie Gewalt, Vorurteile, Hate Speech etc. auszuschließen*
  - → Die Trainingsdaten enthalten ein umfangreiches Spektrum unterschiedlicher menschlicher Sprache
  - → Die Trainingsdaten allgemeiner LLMs haben *keinen spezifischen wissenschaftlichen Zuschnitt*
  - → Die Trainingsdaten können *Fehler, Verzerrungen, Biases und Mißrepräsentationen* enthalten (und tun dies auch)
  - → Die Auswahl der *Trainingsdaten* und die Kriterien ihrer Bereinigung liegen *in der ausschließlichen Hoheit der jeweiligen Anbieter*



* | Perrigo, Billy. 2023. “The $2 Per Hour Workers Who Made ChatGPT Safer”. *Time*, 18.01.2023; https://time.com/6247678/openai-chatgpt-kenya-workers/

On Bullshit

“Bullshit is unavoidable whenever circumstances require someone to talk without knowing what he is talking about. Thus the production of bullshit is stimulated whenever a person’s obligations or opportunities to speak about some topic exceed his knowledge of the facts that are relevant to that topic.”*

- LLMs haben kein Textverständnis
- LLMs haben keine Kenntnis oder ein Bewusstsein über die Welt
- LLMs sind Sprach- nicht Wissensmodelle (das ‚Wissen‘ entsteht eher beiläufig -> Probabilistik)
- Alle derzeit verfügbaren LLMs sind nicht spezifisch wissenschaftlich vortrainiert
- LLMs halluzinieren und erfinden Sachzusammenhänge, Informationen und Quellen
- LLM-generierte Texte sind keine wissenschaftlichen Quellen

→ Ungerechtfertigtes Vertrauen (Es ‚menschelt‘)

* | Frankfurt, Harry G. 2005. *On Bullshit*. Princeton University Press, S. 63. https://doi.org/10.1515/9781400826537

Rules for tools – genKI und GwP

Zur Situation an der Freien Universität Berlin

- Eckpunkte zum Umgang mit KI-basierten Systemen und Tools in Studium und Lehre vom 10.05.2023
  - Über die grds. Zulässigkeit der Verwendung als „zugelassenes Hilfsmittel“ in Prüfungen entscheidet der jeweilige Prüfungsausschuss
    - Daraus resultiert, dass die Verwendung solcher Hilfsmittel a) unter Vorbehalt steht und b) zwingend offengelegt werden muss
  - In diesem Rahmen liegt die Entscheidung darüber, ob und wenn ja, welche Tools verwendet werden dürfen bei der Person, die die Prüfungsleistung abnimmt
    - Wenn Sie beabsichtigen, KI-basierte Tools bei der Erstellung einer schriftlichen Arbeit zu verwenden, besprechen Sie dies mit dem/der Betreuer*in
      - Welche Tools wollen Sie verwenden?
      - Wozu wollen Sie diese verwenden?
      - Wie wird die Verwendung der Tools dokumentiert?

Basics

- Bevor Sie genKI-Tools zur Erstellung eines Textes verwenden, klären Sie, ob dies zulässig ist und in welcher Form die Nutzung dokumentiert werden muss
  - Die meisten Verlage haben bereits Richtlinien (z.B.: https://www.nature.com/nature-portfolio/editorial-policies/ai)
- Wenn Sie genKI-Tools bei der Erstellung von Texten verwenden, sollten Sie die Verwendung für Ihre eigenen Unterlagen vollständig dokumentieren
  - Prompt
  - Output
  - Verwendung des Outputs
  - Hersteller des LLMs
  - Name des LLMs
  - Version des LLMs

Weiterführende Ressourcen

- Chicago, APA und MLA haben jeweils Vorschläge vorgelegt, wie KI-generierte Texte zitiert werden können
  - https://www.chicagomanualofstyle.org/qanda/data/faq/topics/Documentation.html
  - https://apastyle.apa.org/blog/how-to-cite-chatgpt
  - https://style.mla.org/citing-generative-ai/
- VG München, Beschluss v. 28.11.2023 – M 3 E 23.4371 (Zulassung zum Masterstudium wg. mutmaßlicher Nutzung eines LLMs verweigert)
  - https://www.gesetze-bayern.de/Content/Document/Y-300-Z-BECKRS-B-2023-N-42327
- DFG zum Umgang mit generativen KI-Modellen
  - https://www.dfg.de/de/service/presse/pressemitteilungen/2023/pressemitteilung-nr-39

Urheberrechtlich Aspekte bei der Nutzung von LLMs

Eigentum an den generierten Inhalten

- Autorschaft ist ein personenbezogenes Konzept, daher kommen LLMs oder LMMs (Large Multimodal Models) nicht als Autoren infrage
  - Generierte Inhalte sind somit urheberrechtsfrei (-> keine aus dem Urheberrecht begründete Kennzeichnungspflicht)
- In der Regel (-> AGB, Terms of Use) können Nutzende die generierten Inhalte frei – selbst kommerziell – verwenden
- Fraglich ist, ob für generierte Inhalte die Urheberschaft übernommen werden kann:
  - USA: LLM-generierte Texte sind weder die Texte einer dritten Person noch als eigenes Werk urheberrechtlich geschützt (so die US Copyright Authority 2023 im Fall Zarya of the Dawn; diskutierbar)

Eigentum an den generierten Inhalten

The Office has completed its review of the Work's original registration application and deposit copy, as well as the relevant correspondence in the administrative record.[1] We conclude that Ms. Kashtanova is the author of the Work's text as well as the selection, coordination, and arrangement of the Work's written and visual elements. That authorship is protected by copyright. However, as discussed below, the images in the Work that were generated by the Midjourney technology are not the product of human authorship. Because the current registration for the Work does not disclaim its Midjourney-generated content, we intend to cancel the original certificate issued to Ms. Kashtanova and issue a new one covering only the expressive material that she created.

https://www.copyright.gov/docs/zarya-of-the-dawn.pdf

Prompting: Übertragung von Nutzungsrechten

OpenAI (ChatGPT)

Content

Your content. You may provide input to the Services ("**Input**"), and receive output from the Services based on the Input ("**Output**"). Input and Output are collectively "**Content**". You are responsible for Content, including ensuring that it does not violate any applicable law or these Terms. You represent and warrant that you have all rights, licences, and permissions needed to provide Input to our Services.

Ownership of content. As between you and OpenAI, and to the extent permitted by applicable law, you (a) retain your ownership rights in Input and (b) own the Output. We hereby assign to you all our right, title, and interest, if any, in and to Output.

Similarity of content. Due to the nature of our Services and artificial intelligence generally, Output may not be unique and other users may receive similar output from our Services. Our assignment above does not extend to other users' output or any Third Party Output.

Our use of content. We can use your Content worldwide to provide, maintain, develop, and improve our Services, comply with applicable law, enforce our terms and policies and keep our Services safe.

Opt out. If you do not want us to use your Content to train our models, you have the option to opt out by updating your account settings. Further information can be found in this Help Center article. Please note that in some cases this may limit the ability of our Services to better address your specific use case.

Anthropic (Claude)

We will not use your Inputs or Outputs to train our models, unless: (1) your conversations are flagged for Trust & Safety review (in which case we may use or analyze them to improve our ability to detect and enforce our Usage Policy, including training models for use by our Trust and Safety team, consistent with Anthropic's safety mission), or (2) you've explicitly reported the materials to us (for example via our feedback mechanisms), or (3) by otherwise explicitly opting in to training.

Our Privacy Policy explains your rights regarding your personal data, including with respect to our training activities. This includes your right to request a copy of your personal data, and to object to our processing of your personal data or request that it is deleted. We make every effort to respond to such requests. However, please be aware that these rights are limited, and that the process by which we may need to action your requests regarding our training dataset are complex.

Datenschutzrechtliche Aspekte bei der Nutzung von LLMs

Eigene und personenbezogene Daten Dritter

- Die meisten LLMs werden auf Servern außerhalb der EU gehostet und sind daher nicht DSGVO-konform
- Datensparsamkeit mit Blick auf die eigenen Daten
- Personenbezogene Daten Dritter dürfen beim Prompten nicht verwendet werden, wenn die Nutzung dieser Daten durch den Anbieter des LLMs nicht zweifelsfrei ausgeschlossen werden kann oder die ausdrückliche Zustimmung der betroffenen Person(en) eingeholt wurde

Ethische Aspekte

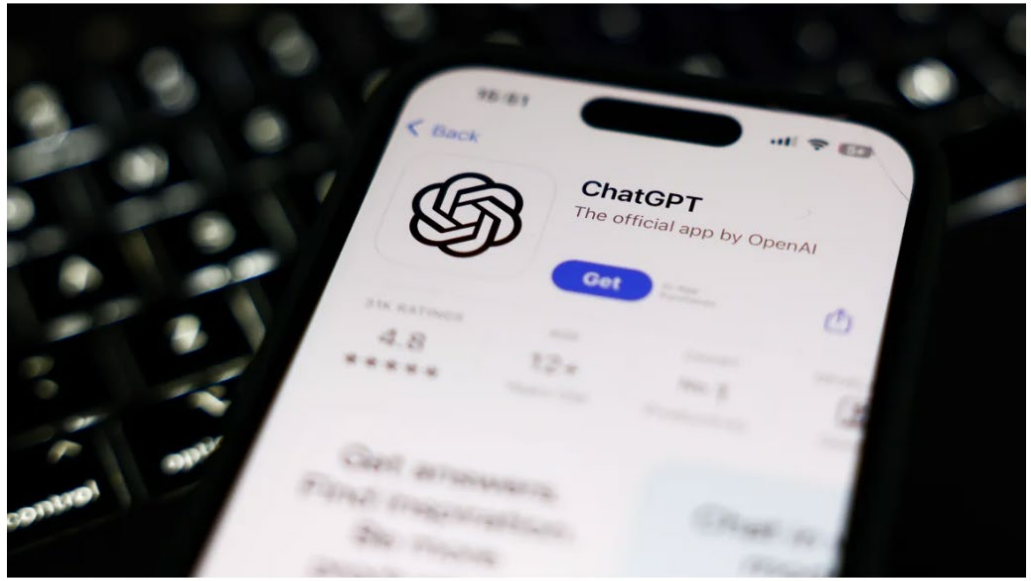
Ethische und weitere Problemfelder

- Trainingsdaten
  - Zusammensetzung
    - Biases
    - Fehlinformationen
  - Herkunft (Urheberrecht, Nutzungsrechte) -> Was bedeutet das z.B. für Open Access / Open Science?
  - Bereinigung
    - Nach welchen Kriterien?
    - Wer definiert diese? –> Überwiegend proprietäre Anbieter
    - Auslagerung der Klickarbeit -> Neokolonialismus?
- Faktisch fehlerhafte Inhalte, erfundene Inhalte, Biases und Stereotype
- Zugang zu LLMs als Exklusionfaktor
- Nachhaltigkeit (Energie, Ressource, Hardware)

Robust? Deep Fakes und problematische Inhalte

ChatGPT can be tricked into telling people how to commit crimes, a tech firm finds

By Anna Cooban, CNN
3 minute read · Published 11:06 AM EDT, Wed October 23, 2024



Generative AI chatbot ChatGPT can produce lists of methods to help businesses evade Western sanctions against Russia and commit other crimes, a tech firm found. Jakub Porzycki/NurPhoto/Getty Images

URTEIL IN BOLTON

Brite hat Missbrauchsfotos mit KI erstellt – 18 Jahre Haft

28.10.2024, 14:51 Lesezeit: 1 Min.

Der 27-Jährige hat echte Kinderfotos mithilfe von KI-Programmen verändert und sexuellen Missbrauch animiert. Nach eigener Aussage erhielt er die Fotos von seinen Auftraggebern: Vätern, Onkeln und Familienfreunden.

- | https://edition.cnn.com/2024/10/23/business/chatgpt-tricked-commit-crimes/index.html
- | https://www.faz.net/aktuell/gesellschaft/kriminalitaet/missbrauchsfotos-mit-ki-erstellt-18-jahre-haft-fuer-briten-110074854.html

Biases und Stereotype

Article

AI generates covertly racist decisions about people based on their dialect

https://doi.org/10.1038/s41586-024-07856-5

Received: 8 February 2024

Accepted: 19 July 2024

Published online: 28 August 2024

Open access

Check for updates

Valentin Hofmann[1,2,3]✉, Pratyusha Ria Kalluri[4], Dan Jurafsky[4] & Sharese King[5]✉

Hundreds of millions of people now interact with language models, with uses ranging from help with writing[1,2] to informing hiring decisions[3]. However, these language models are known to perpetuate systematic racial prejudices, making their judgements biased in problematic ways about groups such as African Americans[4–7]. Although previous research has focused on overt racism in language models, social

Hofmann, V., Kalluri, P.R., Jurafsky, D. *et al.* AI generates covertly racist decisions about people based on their dialect. *Nature* **633**, 147–154 (2024). https://doi.org/10.1038/s41586-024-07856-5

Biases und Stereotype

Press release ›

Generative AI: UNESCO study reveals alarming evidence of regressive gender stereotypes

Ahead of the International Women's Day, a UNESCO study revealed worrying tendencies in Large Language models (LLM) to produce gender bias, as well as homophobia and racial stereotyping. Women were described as working in domestic roles far more often than men ¬– four times as often by one model – and were frequently associated with words like “home”, “family” and “children”, while male names were linked to “business”, “executive”, “salary”, and “career”.

UNESCO, IRCAI (2024). “Challenging systematic prejudices: an Investigation into Gender Bias in Large Language Models”. https://unesdoc.unesco.org/ark:/48223/pf0000388971.locale=en

https://www.unesco.org/en/articles/generative-ai-unesco-study-reveals-alarming-evidence-regressive-gender-stereotypes

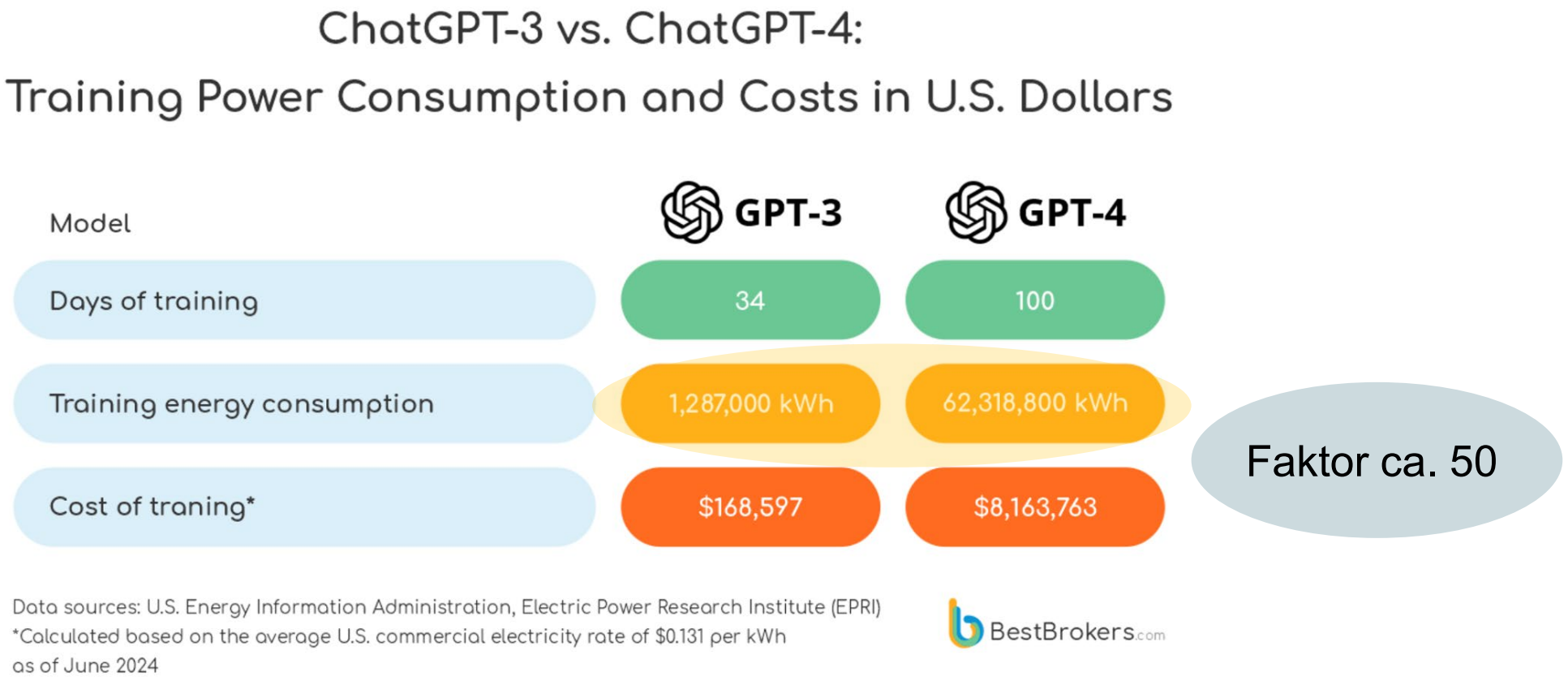
Nachhaltigkeit

„… one assessment suggests that ChatGPT, the chatbot created by OpenAI in San Francisco, California, is already consuming the energy of 33,000 homes. It's estimated that a search driven by generative AI uses four to five times the energy of a conventional web search. Within years, large AI systems are likely to need as much energy as entire nations.

And it's not just energy. Generative AI systems need enormous amounts of fresh water to cool their processors and generate electricity. In West Des Moines, Iowa, a giant data-centre cluster serves OpenAI's most advanced model, GPT-4. A lawsuit by local residents revealed that in July 2022, the month before OpenAI finished training the model, the cluster used about 6% of the district's water. As Google and Microsoft prepared their Bard and Bing large language models, both had major spikes in water use — increases of 20% and 34%, respectively, in one year, according to the companies' environmental reports."

Crawford, Kate. 2024. Generative AI's environmental costs are soaring — and mostly secret. *Nature* 626: 693. https://doi.org/10.1038/d41586-024-00478-x

Nachhaltigkeit



https://www.bestbrokers.com/forex-brokers/ais-power-demand-calculating-chatgpts-electricity-consumption-for-handling-over-78-billion-user-queries-every-year/

Nachhaltigkeit

https://www.appypie.com/blog/hardware-requirements-for-llm-training

Nachhaltigkeit

ChatGPT's Energy Consumption for Responding to Prompts and Its Cost in the U.S.

Time Period	Number of prompts	Energy Consumption	Cost
Day	214,285,714	621,429 kWh	$81,407
Week	1,500,000,000	4,350,000 kWh	$569,850
Month	6,517,860,000	18,901,794 kWh	$2,476,135
Year	78,214,350,000	226,821,615 kWh	$29,713,632

Data sources: U.S. Energy Information Administration, Electric Power Research Institute (EPRI)
Calculations based on: 100 million weekly users, 15 weekly prompts per user, 0.0029 kWh of energy consumption per prompt, average U.S. commercial electricity rate of $0.131/kWh as of June 2024

BestBrokers.com

"Apparently, each time you ask ChatGPT a question, it uses about 0.0029 kilowatt-hours of electricity. This is nearly ten times more than the energy needed for a typical Google search, which consumes about 0.0003 kilowatt-hours per query, according to The Electric Power Research Institute (EPRI)."

https://www.bestbrokers.com/forex-brokers/ais-power-demand-calculating-chatgpts-electricity-consumption-for-handling-over-78-billion-user-queries-every-year/

Vielen Dank für Ihr Interesse

“The author of an ‘artificially intelligent’ program is […] clearly setting out to fool some observers for some time. His success can be measured by the percentage of the exposed observers who have been fooled multiplied by the length of time they have failed to catch on. Programs which become so complex (either by themselves, e.g. learning programs, or by virtue of the author’s poor documentation and debugging habits) that the author himself loses track, obviously have the highest IQ’s.”

Joseph Weizenbaum. 1962. How to make a computer program appear intelligent. *Datamation* 8(2): 24-26 [24]

Anhang: Was motiviert zu wissenschaftlichem Fehlverhalten?

Prävalenz, Gründe und Ursachen

Studentisches Fehlverhalten – Prävalenz

- Mindestens einmal im Erhebungszeitraum
  - Plagiiert → 18 % (Studenten: 19,3 %, Studentinnen: 17 %)
  - Daten gefälscht oder manipuliert → 24 % (Studenten: 25,8 %, Studentinnen: 23,3 %)
- Art und Häufigkeit des Fehlverhaltens korreliert mit Prüfungsformaten
- Niedrige (2–7) und höhere (14+) Semester stärker betroffen

Vier konsekutive, halbjährliche Erhebungswellen beginnend mit SoSe 2010 (n_1=5822, n_2=3486, n_3=2466, n_4=1852 und n_{x+1} jeweils Teilmenge von n_x).

Quelle

Sattler, Sebastian und Martin Diewald. 2013. „FAIRUSE - Fehlverhalten und Betrug bei der Erbringung von Studienleistungen: Individuelle und organisatorisch-strukturelle Bedingungen.“ Projektbericht. Doi: 10.2314/GBV:773897283

Prävalenz, Gründe und Ursachen

Studentisches Fehlverhalten – Gründe und Ursachen

- Die Neigung zu Fehlverhalten sinkt,
    - je höher die eigene Fachkompetenz eingeschätzt wird
    - je höher die eigene Methodenkompetenz eingeschätzt wird
    - je höher die intrinsische Motivation eingeschätzt wird
    - wenn Fehlverhalten moralisch negativ bewertet wird
    - bei Lehrformaten, die auf Verständnis zielen
    - bei fairen, wertschätzenden Dozent:innen
    - Mit steigendem Entdeckungsrisiko und Sanktionsdrohung

Quelle

Sattler, Sebastian und Martin Diewald. 2013. „FAIRUSE - Fehlverhalten und Betrug bei der Erbringung von Studienleistungen: Individuelle und organisatorisch-strukturelle Bedingungen.“ Projektbericht. Doi: 10.2314/GBV:773897283

Prävalenz, Gründe und Ursachen

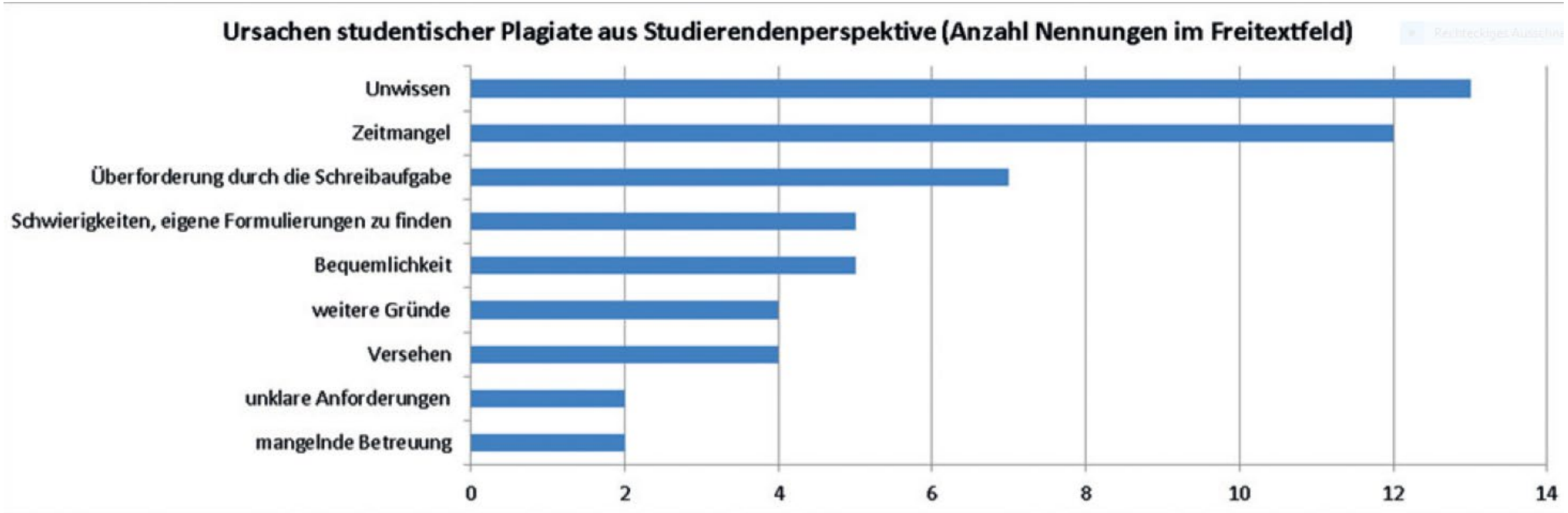
Studentisches Fehlverhalten – Gründe und Ursachen

- Die Neigung zu Fehlverhalten steigt,
    - Je höher der Konkurrenzdruck bewertet wird
    - je höher die Prüfungsangst eingeschätzt wird
    - je höher der Stress eingeschätzt wird
    - Mit der Tendenz zur Prokrastination

Quelle:

Sattler, Sebastian und Martin Diewald. 2013. „FAIRUSE - Fehlverhalten und Betrug bei der Erbringung von Studienleistungen: Individuelle und organisatorisch-strukturelle Bedingungen.“ Projektbericht. Doi: 10.2314/GBV:773897283

Prävalenz, Gründe und Ursachen



Quelle:
Hoffmann, Nora. 2014. Vermittlung wissenschaftlicher Schreibkompetenz zur Förderung akademischer Integrität. *Information. Wissenschaft & Praxis* 65(1): 51–62 [52].

Gründe und Ursachen

Fehlverhalten in der Wissenschaft

- Die Neigung zu Fehlverhalten steigt,
    - je höher der Publikationsdruck bewertet wird
    - bei hohem Druck, Finanzierung einzuwerben
    - in einem als instrumentell empfundenen Arbeitsklima
    - in einem durch Misstrauen geprägten Arbeitsklima

Schlechtes Arbeitsklima erklärt 22 % Abweichung hinsichtlich der Häufigkeit berichteten Fehlverhaltens

Publikationsdruck erklärt 12 % Abweichung hinsichtlich der Häufigkeit berichteten Fehlverhaltens

Quelle:

Haven, Tamarinde, Joeri Tijdink, Brian Martinson, Lex Bouter und Frans Oort. 2021. „Explaining Variance in Perceived Research Misbehavior: Results from a Survey Among Academic Researchers in Amsterdam.“ *Research Integrity and Peer Review* 6 (1): 7. https://doi.org/10.1186/s41073-021-00110-w.

Gründe und Ursachen