Zeitschrift für Medienwissenschaft, Jg. 18, Heft 34 (1/2026), https://doi.org/10.14361/zfmw-2026-180108. Published by transcript Verlag. This work is licensed under the Creative Commons Attribution CC-BY-NC-ND 4.0 DE licence

YANNICK NEPOMUK FRITZ

AM HARTEN KERN

KI nach der ‹Hardware-Wende›

Im Frühjahr 2025 erlebte die US-amerikanische KI-Industrie einen vermeintlichen «Sputnik-Moment»: Mit der Veröffentlichung des neuesten Modells der chinesischen Firma DeepSeek sei, laut Investor Marc Andreessen und entsprechend diesem Masternarrativ des Kalten Krieges, die technologische Unterlegenheit des Westens zur Schau gestellt worden.[1] Ungeachtet dessen, dass die Entwicklung von KI-Modellen «grundsätzlich inkrementell» verläuft,[2] wie zahlreiche historisch argumentierende Arbeiten darlegen,[3] scheinen die Auswirkungen der Modelle stellenweise so gravierend, dass eine solche Rhetorik des radikalen Einschnitts bemüht wird. Auch in medienwissenschaftlichen Publikationen ist beispielsweise von einer «Zeit Vor-ChatGPT» die Rede, um die Zeit vor der Popularisierung generativer KI zu kennzeichnen.[4]

Im Feld der sich noch formierenden Critical AI Studies wurden indessen diverse Anliegen vorgebracht und so etwa Positionen zu Automatisierung, zu Mustererkennung und Diskriminierung durch KI oder zu diversen Intelligenzbegriffen formuliert.[5] Der ‹Moment› um DeepSeek lenkt die Aufmerksamkeit zunächst auf einen anderen Bereich: Wie zu zeigen ist, ist er auch ein Phänomen von Plattformisierungsdynamiken in der KI-Industrie und weitläufigeren geoökonomischen Intensivierungen. Letztere stehen dabei in Verbindung mit infolge der Neoliberalisierung geschaffenen globalen Wirtschaftszusammenhängen, die nun von staatlichen Akteur*innen nicht überformt, sondern zunehmend instrumentalisiert werden.[6] Eine Plattformisierung der KI-Industrie meint die zunehmende ‹Penetration› von Infrastrukturen, ökonomischen Prozessen und regulierenden Rahmenbedingungen durch Plattformen.[7] Plattformen sind dabei «(re-)programmierbare digitale Infrastrukturen», die Interaktionen zwischen Endnutzer*innen und Anbieter*innen von Komplementärsystemen mittels Daten organisieren, die systematisch gesammelt, algorithmisch verarbeitet, monetarisiert und zirkuliert werden.[8]

Dieser Beitrag plädiert im Anschluss an Ludovico Rella für eine kritische Betrachtung der materiellen politischen Ökonomie und Epistemologie der KI, um diese Zusammenhänge zu fassen. Damit soll die Frage beantwortet werden, was,

1 Vgl. Christiaan Hetzner: Marc Andreessen Warns Chinese ChatGPT rival DeepSeek Is ‹AI's Sputnik Moment›, *Fortune*, 27.1.2025, fortune.com/2025/01/27/marc-andreessen-deepseek-sputnik-ai-markets (15.9.2025).

2 Fabian Offert, Ranjodh Singh Dhaliwal: The Method of Critical AI Studies. A Propaedeutic, *arXiv*, 23.3.2025 (3. Fassung), 1–10, hier 5, doi.org/10.48550/arXiv.2411.18833 (15.9.2025), Übers. YNF.

3 Vgl. u. a. Jonathan Roberge, Michael Castelle (Hg.): *The Cultural Life of Machine Learning. An Incursion into Critical AI Studies*, Cham 2021; Matteo Pasquinelli: *The Eye of the Master. A Social History of Artificial Intelligence*, London 2023; Ranjodh Singh Dhaliwal, Théo Lepage-Richer, Lucy Suchman: *Neural Networks*, Minneapolis 2024.

4 Anna Tuschling, Andreas Sudmann, Bernhard J. Dotzler: Trialog über den Versuch, eine medienhistorische Passage zu dokumentieren, in: dies. (Hg.): *ChatGPT und andere «Quatschmaschinen». Gespräche mit Künstlicher Intelligenz*, Bielefeld 2023, 263–282, hier 265.

5 Vgl. für einen Überblick: Rita Raley, Jennifer Rhee: Critical AI. A Field in Formation, in: *American Literature*, Bd. 95, Nr. 2, 2023, 185–204, doi.org/10.1215/00029831-10575021.

6 Vgl. Milan Babić: *Geoökonomie. Anatomie der neuen Weltordnung*, Berlin 2025.

7 Vgl. Thomas Poell, David Nieborg, José van Dijck: Platformisation, in: *Internet Policy Review*, Bd. 8, Nr. 4, 2019, 1–13, hier 5 f., doi.org.10.14763/2019.4.1425.

8 Vgl. ebd., 3, Übers. YNF.

wenn sich eine «Zeit Vor-ChatGPT» konstatieren lässt, jene nach DeepSeek charakterisiert.[9] Zugleich soll gefragt werden, worin (in Zeiten der Diskussionen um eine immer größere, spekulative «KI-Blase»)[10] der ökonomische Wert generativer KI besteht. Konkret ist DeepSeek zum einen belastet durch Vorwürfe seiner US-amerikanischen Konkurrenz, es hätte seine KI-Modelle unerlaubt auf US-amerikanischen Modellen über deren *application programming interfaces* (APIs) trainiert, und zum anderen aufgrund von Handelsbeschränkungen weitestgehend von den US-amerikanischen Wertschöpfungsketten abgeschnitten, in denen KI-Hardware-Komponenten hergestellt werden.

Den Fragen von Wertschöpfung und Plattformisierung nähert sich dieser Beitrag über das an, was DeepSeeks CEO Liang Wengfeng in einem Interview von 2024 selbstbewusst «hardcore innovation» genannt hat.[11] Im Folgenden möchte ich mit der Untersuchung dieser *hardcore innovation* am buchstäblichen ‹harten Kern› beginnen und ausgehend von der Hardware zeigen, was eine KI-Industrie nach der ‹Hardware-Wende› auszeichnet.

Nach einer kurzen Einordnung des ‹DeepSeek-Moments› nehme ich mit der Hardware zunächst die materielle Dimension der politischen Ökonomie der KI in den Blick. Historisch wird nachverfolgt, inwieweit Hardware-Konfigurationen nicht nur Faktor bei der materiellen Implementierung, sondern auch bei der epistemologischen Konzeption von KI gewesen sind. Diese eng verflochtenen Entwicklungen von KI und entsprechender Hardware zeichne ich mit einem Fokus auf *graphics processing units* (GPUs) nach. Bereits an den GPUs, und darüber hinaus an den APIs sowie der Kontrolle der KI-Unternehmen über Protokolle, Standards, Entwickler*innen-Frameworks, Daten und Expert*innen, lässt sich dabei eine Plattformisierung der KI-Industrie erkennen. Die Innovationen von DeepSeek werden im Kontext dieser Plattformisierungsdynamiken und der US-amerikanischen Exportbeschränkungen verortet.

In einem ersten Schritt wird also der Zusammenhang von Hardware und Epistemologie herausgearbeitet. Als Hardware-bedingtes epistemologisches Instrument führen KI-Modelle zu spezifischen Repräsentationsregimen ihrer Trainingsdaten. Diese Funktionsweise zeichne ich in einem zweiten Schritt modellarchitekturspezifisch nach, um zu verstehen, welcher Wert bei einem Trainingsprozess von Modellen über APIs abgeschöpft werden könnte. Es wird gezeigt, dass sich die derzeit gängige Transformer-Architektur in eine Gesamtlogik zuvor analysierter Hardware-Komponenten und politischer Ökonomie einpasst. Diese Logik wird über eine medientheoretische Kritik des Kompressionsbegriffs erschlossen. Hier zeigt sich zum einen, wie Hardware-bedingte Epistemologien die Sicht durch KI-Modelle auf die Welt formen, und zum anderen, wie Plattformen und Interfaces der KI Schauplätze geoökonomischer Intensivierungen werden. Abschließend wird der ‹DeepSeek-Moment› als symptomatisch für einen von Luciano Floridi ausgerufenen *hardware turn* beschrieben. Wurde KI lange als ‹virtuell› und ‹immateriell› vermarktet und imaginiert, wird nun erkannt, dass sie über ihre Hardware-Bedingungen kontrollierbar

9 Vgl. Ludovico Rella: Close to the Metal. Towards a Material Political Economy of the Epistemology of Computation, in: *Social Studies of Science*, Bd. 54, Nr. 1, 2024, 3–29, doi.org/10.1177/03063127231185095.

10 Vgl. Seth Fiegerman, Carmen Reinicke: Why Fears of a Trillion-Dollar AI Bubble Are Growing, *Bloomberg*, 24.11.2025, bloomberg.com/news/articles/2025-11-24/why-ai-bubble-concerns-loom-as-openai-microsoft-meta-ramp-up-spending (6.12.2025).

11 Liang Wengfeng zit. n. Jordan Schneider u. a.: Deepseek. The Quiet Giant Leading China's AI Race [Annotierte Interview-Übersetzung], *ChinaTalk*, 27.11.2024, chinatalk.media/p/deepseek-ceo-interview-with-chinas (2.9.2025).

ist.[12] Die Formulierung ‹*nach* der Hardware-Wende› soll dabei auch markieren, dass, im Zuge einer momentan gegebenen politisch-ökonomischen Zentrierung auf spezifische Hardware-Komponenten, der Horizont von KI durch diese Zentrierung auch spezifisch abgesteckt und gegebenenfalls verengt wird. Eine Einordnung dieser Dynamik will schließlich aktuelle Fluchtpunkte und kritische Betrachtungsweisen für die Weiterentwicklung der gegenwärtigen KI-Industrie aufzeigen.

Der ‹DeepSeek-Moment›

Die chinesische Firma DeepSeek, 2023 ausgegründet aus dem auf algorithmisches Trading spezialisierten Hedgefonds High-Flyer, veröffentlichte im Januar 2025 das KI-Modell R1, das in branchenüblichen Tests mit bisherigen Spitzenmodellen wie OpenAIs Modell o1 gleichzog.[13] Zwar gibt es mehrere Firmen, die OpenAI wirksam Konkurrenz machen, aber das Besondere an DeepSeeks Modell ist, dass das Trainieren von R1 nur 5,6 Millionen US-Dollar gekostet haben soll (bei OpenAIs o1 reichen Schätzungen bis 100 Millionen).[14] Und dass dies scheinbar mithilfe von GPUs geschehen ist, die aufgrund US-amerikanischer Exportbeschränkungen leistungstechnisch hinter denen der US-Konkurrenz zurückbleiben. Nvidia, die Firma, die mit einem Marktanteil von 92 Prozent (Stand März 2025) einen Großteil der für die KI-Industrie so wichtigen GPUs produziert,[15] verlor nach der Veröffentlichung von R1 an einem Tag 17 Prozent ihres Aktienwerts, ca. 600 Milliarden US-Dollar.[16] Viele der mit R1 demonstrierten technischen Errungenschaften von DeepSeek waren allerdings schon Teil der Vorgänger-Modelle V3 und V2 aus dem Vorjahr (und damit auch Teil der eigentlichen Entwicklungskosten bis hin zu R1).[17] Dies beinhaltet auch die gesteigerte Wirkmacht sogenannter *knowledge distillation*, jener Technik, bei der größere Modelle genutzt werden, um über deren Repräsentationen der Trainingsdaten kleinere Modelle zu trainieren.[18] OpenAI erhob schon kurz nach der Veröffentlichung von R1 den Vorwurf, dass DeepSeek sein Modell mithilfe der APIs und auf den Modellen von OpenAI trainiert habe.[19] *Knowledge distillation* ist dabei keine neue Technik und wird von Analyst*innen als gängige Praxis im Trainingsprozess von KI-Modellen bewertet:[20] Nicht nur würden KI-Unternehmen ihre eigenen Modelle häufig destillieren, dies passiere auch von außen über APIs, also über Spezifikationen und Protokolle, die den Austausch zwischen Software und Software regulieren, oder gar per Chat Interface, also per User Interface, das mittels gestalterischer Elemente Software für Nutzer*innen – oder bei Zweckentfremdung für Bots – zugänglich macht.[21] Die *terms of service* der meisten Modelle würden dies zwar verbieten, praktisch ließe sich dies aber nur über Zugangsbeschränkungen mittels einer Sperrung von IP-Adressen oder der Begrenzung von Netzwerkverkehr an eigenen Schnittstellen (*rate limiting*) umsetzen.[22] Ganz abgesehen davon, dass KI-Modelle maßgeblich auf kollektiver Arbeit beruhen, die nicht selten unerlaubt genutzt worden

12 Vgl. Luciano Floridi: The Hardware Turn in the Digital Discourse. An Analysis, Explanation, and Potential Risk, in: *Philosophy & Technology*, Bd. 37, ArtikelNr. 39, 2024, 1–7, doi.org/10.1007/s13347-024-00723-1.

13 Vgl. DeepSeek-AI u. a.: DeepSeek-R1. Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, *arXiv*, 22.1.2025, 1–22, doi.org/10.48550/arXiv.2501.12948.

14 Vgl. Louis Tompros, interviewt von Scott Young: DeepSeek, ChatGPT, and the Global Fight for Technological Supremacy, *Harvard Law Today*, 25.2.2025, hls.harvard.edu/today/deepseek-chatgpt-and-the-global-fight-for-technological-supremacy (26.9.2025).

15 Vgl. Mark Bergen: How the AI Boom Created the Most Valuable Monopolies in History, *Bloomberg*, 20.3.2025, bloomberg.com/news/features/2025-03-20/are-ai-monopolies-here-to-stay-nvidia-and-the-future-of-ai-chips (3.12.2025).

16 Vgl. Samantha Subin: Nvidia Sheds Almost $600 Billion in Market Cap, Biggest One-Day Loss in U.S. History, *CNBC*, 27.1.2025, cnbc.com/2025/01/27/nvidia-sheds-almost-600-billion-in-market-cap-biggest-drop-ever.html (15.9.2025).

17 Vgl. DeepSeek-AI u. a.: DeepSeek-V3 Technical Report, *arXiv*, 18.2.2025 (2. Fassung), 1–53, hier 5, doi.org/10.48550/arXiv.2412.19437 (15.9.2025).

18 Vgl. Ben Thompson: DeepSeek FAQ [Beitrag auf Unternehmensblog], *Stratechery*, 27.1.2025, stratechery.com/2025/deepseek-faq (14.9.2025).

19 Vgl. Eleanor Olcott, Cristina Criddle: OpenAI Says It Has Evidence China's DeepSeek Used Its Model to Train Competitor, *Financial Times*, 29.1.2025, ft.com/content/a0dfedd1-5255-4fa9-8ccc-1fe01de87ea6 (2.9.2025).

20 Vgl. Thompson: DeepSeek FAQ; Tompros, Young: DeepSeek, ChatGPT, and the Global Fight for Technological Supremacy.

21 Vgl. Matthew Fuller, Florian Cramer: Interface, in: Matthew Fuller (Hg.): *Software Studies. A Lexicon*, Cambridge (MA) 2008, 149–153, hier 149, doi.org/10.7551/mitpress/7725.003.0022.

22 Vgl. Thompson: DeepSeek FAQ.

ist, droht der bloße Fokus auf destillierten und daher vermeintlich ‹kopierten› Modellen durch DeepSeek (wie ihn z. B. OpenAI propagiert) eine Betrachtung fundamentalerer Dynamiken zu verstellen, die sich am Fall DeepSeek exemplifizieren lassen.

Zum einen sind das geoökonomische Zusammenhänge. Bezeichnend ist, dass R1 am 20. Januar 2025 veröffentlicht wurde, dem Tag, an dem Donald Trump in Anwesenheit von Tech-Industriegrößen wie Sam Altman, Mark Zuckerberg und Sundar Pichai als 47. Präsident der Vereinigten Staaten vereidigt wurde. Trotz bereits vorangegangener wesentlicher technologischer Innovation entfaltete DeepSeek vor diesem Hintergrund eine solche Wirkung und wurde maßgeblich im Kontext der US-Geoökonomie rezipiert. Zum anderen bedingte DeepSeeks beschränkter Zugang zu Hardware besagte *hardcore innovation*. Ihre kritische Beschreibung bedarf einer genaueren Charakterisierung im Rahmen der politischen Ökonomie der KI.

«Bigger is better»? Ausprägungen einer materiellen politischen Ökonomie der KI

Der Begriff der materiellen politischen Ökonomie, den Donald MacKenzie für seine Analyse des algorithmischen Hochfrequenzhandels verwendet[23] und den ich auf die KI-Industrie übertragen möchte, umfasst folgende Dimensionen: Sie ist *materiell*, weil nicht-menschliche Materialitäten eine gewisse politische Handlungsmacht in Bezug auf die Strukturen haben, auf die sie einwirken, und weil sie diese Beziehungen auf spezifische Weise verräumlichen; *politisch*, weil diese Handlungsmacht immer schon Einschränkungen des Handlungsspielraums anderer Akteur*innen bedeutet; und *ökonomisch*, weil diese materiell-politischen Zusammenhänge dazu dienen, Ressourcen und Gewinne zu extrahieren oder ihre Verteilung zu verändern. Rella fügt MacKenzies Begriff noch eine *epistemologische* Dimension hinzu, denn diese materielle politische Ökonomie beeinflusst zudem Weisen der Wissensproduktion.[24] Die Dynamiken dieser Ökonomie wie auch DeepSeeks Erfolg lassen sich an der Rolle der Hardware verdeutlichen.

Die heute erfolgreichen KI-Modelle funktionieren *konnektionistisch*: Künstliche neuronale Netze werden auf Daten trainiert, um daraus Muster und Gesetzmäßigkeiten abzuleiten.[25] Ein künstliches neuronales Netzwerk ist ein mehrschichtiges Ensemble mathematischer Funktionen, die als Netzwerk strukturiert sind, wobei jede Funktion (also jedes ‹Neuron›) eine Reihe von Werten als Eingabe erhält und diese auf bestimmte Weise transformiert. Diese Transformationen geschehen im Netzwerk *parallel* als eine Reihe von Matrix-Vektor-Multiplikationen.[26] Schon in der Geschichte des Konnektionismus wie auch in der von DeepSeek zeigt sich, dass Hardware kein starres Substrat abstrakter Kognition ist – gewissermaßen ein medientheoretischer Grundgedanke. Denn der konnektionistische Ansatz war bis in die 1980er Jahre zunächst

23 Vgl. Donald MacKenzie: *Trading at the Speed of Light. How Ultrafast Algorithms Are Transforming Financial Markets*, Princeton 2021.

24 Vgl. Rella: Close to the Metal, 7.

25 Vgl. für eine Historisierung des Konnektionismus: Dominique Cardon, Jean-Philippe Cointet, Antoine Mazières: Neurons Spike Back. The Invention of Inductive Machines and the Artificial Intelligence Controversy, in: *Réseaux*, Bd. 5, Nr. 211, 2018, 173–220, doi.org/10.3917/res.211.0173.

26 Vgl. Rella: Close to the Metal, 14.

erfolglos, die Datenmengen und die Rechenkapazitäten waren zu gering, als dass statistische Mustererkennung Wirkmacht hätte entfalten können. Der endgültige Durchbruch des konnektionistischen Paradigmas Anfang der 2010er Jahre war schließlich ganz direkt mit dem Aufkommen von Big Data und dem Einsatz des *parallel computing* verbunden. Letzteres wiederum wurde durch GPUs ermöglicht, die selbst aufgrund ihres Einsatzes in Videospielen und des Handels mit Kryptowährungen Massenware geworden waren[27] und die ebenjene Matrix-Vektor-Multiplikationen drastisch beschleunigten, wie Rella nachzeichnet.[28] Im Jahr 2004 wurde erstmals ein künstliches neuronales Netz auf einer GPU implementiert (und eine 20-fache Performancesteigerung erzielt).[29] Wie mittels GPUs Skalierungsprobleme sowohl bezüglich Modellgröße wie auch Trainingsdatenset-Größe adressiert werden können, wurde 2009 systematisiert vorgestellt. In diesem Zusammenhang wurde außerdem propagiert, dass GPUs die KI-Industrie «revolutionieren» würden.[30] Im Jahr 2012 wurde dann für den Einsatz von *deep learning*, einem Teilbereich konnektionistischer KI, mit dem künstlichen neuronalen Netz AlexNet die Verwendung von «web-scale data» und *parallel computing* über eine Vielzahl von GPUs vorgeschlagen und (mit großer Resonanz) demonstriert.[31] Zur Illustration des Performanceanstiegs: 2012 nutzten Forscher*innen bei Google 16.000 CPU-Kerne (*central-processing-unit*-Kerne), um Bilder von Katzen zu klassifizieren. Nur ein Jahr später wurde die gleiche Aufgabe mit 2 CPU-Kernen und 4 GPUs bewältigt.[32] Sara Hooker schreibt zu dieser Entwicklung:

> In the field of artificial intelligence research, […] it is our tooling which has played a disproportionate role in deciding what ideas succeed (and which fail). [...] [A] research idea wins because it is compatible with available software and hardware and not because the idea is superior to alternative research directions.[33]

Die unterschiedliche ‹Verfügbarkeit› jener Komponenten bedingt in den USA und in China unterschiedliche Entwicklungen. In den USA kam es zur Propagierung der Einsicht einer «bitter lesson» und zu einer «bigger is better»-Ideologie.[34] Die von Richard Sutton 2019 proklamierte «bitter lesson» der KI-Forschung besagt, dass im Zuge des stetigen Anstiegs der Rechenkapazität von Computerchips (die bereits Mitte der 1960er Jahre als Moore's Law bekannt geworden war) nicht bloß menschliches Wissen der treibende und entscheidende Faktor in der Weiterentwicklung von KI sei, sondern eben jene Methoden, die sich im Zuge jenes Kapazitätsanstiegs gleichsam skalieren ließen.[35] Daher auch der hohe Marktwert von Nvidia als GPU-Produzent: Steigende Investitionen in Chips bedeuteten bislang analog dazu eine Verbesserung der Leistung von KI-Modellen. Und mit mehr Chips konnten Modelle mit einer höheren Anzahl an Parametern trainiert werden.

Für China schienen buchstäblich andere Gesetze zu gelten: DeepSeek-CEO Liang propagierte die *hardcore innovation* seines Teams, weil es mutmaßlich nicht auf die zu diesem Zeitpunkt marktführenden Nvidia-H100-GPUs zugrei-

27 Vgl. ebd., 7–11.

28 Vgl. ebd., 14.

29 Vgl. Kyoung-Su Oh, Keechul Jung: GPU Implementation of Neural Networks, in: *Pattern Recognition*, Bd. 37, Nr. 6, 2004, 1311–1314, doi.org/10.1016/j.patcog.2004.01.013.

30 Vgl. Rajat Raina, Anand Madhavan, Andrew Y. Ng: Large-Scale Deep Unsupervised Learning Using Graphics Processors, in: *Proceedings of the 26th International Conference on Machine Learning*, Montreal, 14.6.2009, 873–880, doi.org/10.1145/1553374.1553486 (5.12.2025), Übers. YNF.

31 Vgl. Fabian Offert, Thao Phan: A Sign That Spells. Machinic Concepts and the Racial Politics of Generative AI, in: *Journal of Digital Social Research*, Bd. 6, Nr. 4, 2024, 49–59, hier 52, doi.org/10.33621/jdsr.v6i440462. Offert und Phan beziehen sich auf Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton: ImageNet Classification with Deep Convolutional Neural Networks, in: *Communications of the ACM*, Bd. 60, Nr. 6, 2017, 84–90, doi.org/10.1145/3065386.

32 Vgl. Sara Hooker: The Hardware Lottery, *arXiv*, 21.9.2020 (2. Fassung), 1–19, hier 6, doi.org/10.48550/arXiv.2009.06489 (5.12.2025).

33 Ebd., 1.

34 Vgl. Yannick Fritz: Intelligenz als Machtkonzentration. Interview mit Meredith Whittaker, der Präsidentin von Signal, über KI und die Big-Tech-Konzerne, in: *springerin. Hefte für Gegenwartskunst*, Nr. 1: ArtGPT, 2024, 43–47, hier 46.

35 Vgl. Richard Sutton: The Bitter Lesson [Blog-Beitrag], *Incomplete Ideas*, 13.3.2019, incompleteideas.net/IncIdeas/BitterLesson.html (4.9.2025).

fen konnte (oder diese in begrenzter Zahl ins Land schmuggeln musste).[36] Die US-Administration hatte 2022 unter dem damaligen Präsidenten Joe Biden umfassende Exportembargos erlassen, nachdem sie inländische Investitionen in die Halbleiter-Industrie im Rahmen des CHIPS and Science Act erhöht hatte.[37] Unter diesen Hardware-Bedingungen hat DeepSeek unterschiedliche Innovationen auf Modellarchitekturebene vorangetrieben – Forschung, die die chinesische Konkurrenz weniger verfolgte und die auch weltweit für offen veröffentlichte Modelle weniger üblich war. So erreichte DeepSeek mit einer neuartigen *attention*-Architektur (*multi-head latent attention*) – einem Teil der Transformer-Architektur neuronaler Netze, auf die ich unten eingehe – eine Reduktion des Speicherverbrauchs von 5–13 Prozent. Auch wurden die Rechenkosten mittels einer sogenannten *mixture-of-experts*-Struktur (eine spezifische Anordnung von Sub-Netzwerken im Modell) minimiert, wodurch die Gesamtkosten ebenfalls gesenkt werden konnten.[38] Zudem wurden für das Training der Modelle *reinforcement-learning*-Algorithmen entwickelt, die keines menschlichen Feedbacks bedürfen und einen geringeren Verbrauch an Rechenkapazität haben.[39] Und schließlich hat das chinesische Unternehmen ihm zugängliche, weniger leistungsstarke Nvidia-Chips auf einem der Hardware noch näheren Instruktionslevel (PTX, *parallel thread execution*) programmiert, um die Nutzung der Chips zu optimieren.[40]

Plattformisierung «close to the metal»

Hier zeigt sich eine materielle Dimension der KI-Industrie ganz konkret. Nvidia ist der Entwickler von CUDA (Compute Unified Device Architecture), einer proprietären Software-Ebene samt API zur Datenverarbeitung auf GPUs und CPUs. Anders als CPUs können GPUs eine Vielzahl von Berechnungen simultan ausführen. CUDA wurde dabei ursprünglich extra für GPUs geschaffen, um dieses *parallel computing* auch für nicht-grafikbezogene Prozesse einsetzen zu können.[41] DeepSeeks Entwickler*innen umgingen CUDA durch die direkte Nutzung von PTX, eine Befehlssatzarchitektur, in die CUDA-Programme erst übersetzt werden müssen und die maschinenlesbar näher an der physischen GPU-Hardware ausgeführt wird. Die Entwickler*innen arbeiteten in diesem Sinne *closer to the metal* – also unter Beachtung der Affordanzen und Limitierungen der spezifischen Funktionsweise der GPUs.[42] Der analytische Begriff «close to the metal» wird von Rella im Zuge des von ihm vorgeschlagenen *metonymic turn* vorgebracht: Anstatt Abstraktionen als Repräsentationen genereller Konzepte zu verwenden, wird der Fokus auf die spezifischen Funktionsweisen von und die situativen Beziehungen zwischen den einzelnen Komponenten der *computation* gelegt. Diese beeinflussen, wie Repräsentationen gespeichert, verarbeitet und übertragen werden können.[43] Die für DeepSeek spezifischen Hardware-Bedingungen machten diese Hinwendung in der Praxis notwendig und bedingten schließlich eine wettbewerbsfähige Performance der verwendeten, andernfalls leistungsschwächeren GPUs.

36 Vgl. JS Tan: DeepSeeking the Truth, *The Baffler*, 21.5.2025, *thebaffler.com/latest/deepseeking-the-truth-tan* (4.9.2025).

37 Vgl. Stephen Nellis, Karen Freifeld, Alexandra Alper: U.S. Aims to Hobble China's Chip Industry with Sweeping New Export Rules, *Reuters*, 10.10.2022, *reuters.com/technology/us-aims-hobble-chinas-chip-industry-with-sweeping-new-export-rules-2022-10-07* (9.9.2025).

38 Vgl. Schneider, Shen, Zhang: Deepseek.

39 Vgl. Daya Guo u. a.: DeepSeek-R1 Incentivizes Reasoning in LLMs through Reinforcement Learning, in: *Nature*, Bd. 645, Nr. 8081, 2025, 633–638, *doi.org/10.1038/s41586-025-09422-z*.

40 Vgl. Thompson: DeepSeek FAQ.

41 Vgl. Rella: Close to the Metal, 9.

42 Vgl. ebd., 5.

43 Vgl. ebd.

Dahingehend zeichnet sich hier eine fortschreitende Plattformisierung der KI *close to the metal* ab. Wie oben bereits definiert, organisiert eine Plattform die Interaktionen zwischen Endnutzer*innen und Anbieter*innen von Komplementärsystemen mittels systematisch gesammelter Daten, die algorithmisch verarbeitet, monetarisiert und zirkuliert werden.[44] CUDA ist eine Plattform, weil sie als Ökosystem mit Compilern, Librarys und Toolkits den Zugang zu *high-performance computing* erleichtert, weil sie gleichzeitig Netzwerkeffekte hervorbringt, wenn Entwickler*innen Anwendungen eigens für die Plattform programmieren, die Benutzer*innen schließlich an sie binden, und weil sie das Training von KI-Modellen nur auf den ebenfalls proprietären Nvidia-GPUs ermöglicht.[45] Eine Plattformisierung – also jene ‹Penetration› von Infrastrukturen, ökonomischen Prozessen und regulierenden Rahmenbedingungen durch Plattformen – bewirkt zudem eine Neuordnung kultureller Praktiken und Vorstellungen in Bezug auf diese Plattformen.[46] Die «bigger is better»-Ideologie kann auch als Folge dieser Vorstellungen betrachtet werden – insbesondere als Folge der Vorstellung fortwährender Skalierung von Hardwarekomponenten, wie sie in der «bitter lesson» formuliert ist und worauf im Rahmen der Maxime «bigger is better» gesetzt wird. DeepSeeks Erfolg steht jener Plattformisierung zunächst entgegen und unterwandert damit diese Vorstellung, unterstreicht aber zugleich die zentrale Wirkmacht von Plattformen wie CUDA in der KI-Industrie. CUDA dabei als Plattform zu definieren, die auch ein API beinhaltet, verdeutlicht den dynamischen Charakter, den Interfaces im medialen Gefüge annehmen. Zudem lässt sich hieran ablesen, dass sich über das Modulieren von Interfaces eine Plattformisierung verstärken lässt.[47] Dies zeigt sich hier konkret im Fall DeepSeek und CUDA, an anderer Stelle lassen sich an OpenAIs Modell GPT-3 Plattformisierungsdynamiken im Kontext von Interfaces beobachten: Stand GPT-2 noch frei zum Download bereit, war GPT-3 nur noch *off-the-shelf* als ChatGPT (und damit per Chat User Interface) oder per API-Zugang nutzbar, die mit Bezahlschranken versehen wurden.[48]

Indessen ist CUDA also nur eine von vielen Dimensionen der Plattformisierung der KI – eine Dimension, die sich zudem durch die gemeinsame Open-Source-Initiative und CUDA-Alternative Triton von Microsoft, Meta, Google, Amazon und anderen oder die Entwicklung eigener KI-Chips, etwa TPUs (*tensor processing units*) durch Google, angegriffen sieht.[49] DeepSeeks Konkurrenten Meta und OpenAI, deren APIs DeepSeek zum Zwecke der *knowledge distillation* genutzt haben soll,[50] sind Konzerne, die ihre Vormachtstellung ansonsten auch im Modus «bigger is better» verteidigen: mit einem besonderen Zugang zu Rechenleistung (*compute*), zu den besagten GPUs, zu Daten und Expert*innen, im Besitz von Entwickler*innen-Frameworks und mit der Marktmacht, Arbeitsprozesse und Standards zu definieren.[51] Im Sinne der obigen ‹Penetration› von Infrastrukturen, ökonomischen Prozessen und regulierenden Rahmenbedingungen durch Plattformen kann dieser Modus selbst als Plattformisierung betrachtet werden, der sich DeepSeek (und andere Entwickler*innen jenseits

44 Vgl. Poell, Nieborg, van Dijck: Platformisation, 3.

45 Vgl. David Gray Widder, Sarah West, Meredith Whittaker: Open (For Business). Big Tech, Concentrated Power, and the Political Economy of Open AI, in: SSRN *Electronic Journal*, 17.8.2023, 1–27, hier 7, dx.doi.org/10.2139/ssrn.4543807.

46 Vgl. Poell, Nieborg, van Dijck: Platformisation, 6.

47 Vgl. Sarah Burkhardt, Bernhard Rieder: Foundation Models Are Platform Models. Prompting and the Political Economy of AI, in: *Big Data & Society*, Bd. 11, Nr. 2, 2024, 1–15, hier 2, doi.org/10.1177/20539517241247839.

48 Vgl. Dieuwertje Luitse, Wiebke Denkena: The Great Transformer. Examining the Role of Large Language Models in the Political Economy of AI, in: *Big Data & Society*, Bd. 8, Nr. 2, 2021, 1–14, hier 9, doi.org/10.1177/20539517211047734.

49 Vgl. Tim Bradshaw: Nvidia's Rivals Take Aim at Its Software Dominance, *Financial Times*, 21.5.2024, ft.com/content/320f35de-9a6c-4dbf-b42f-9cdaf35e45bb (3.12.2025).

50 Die Verwendung von Metas Open-Source-Modell LLaMA3 hat DeepSeek selbst ausgewiesen. Vgl. DeepSeek-AI: DeepSeek-R1.

51 Vgl. Fritz: Intelligenz als Machtkonzentration, 46.

der US-amerikanischen Big-Tech-Konzerne) gegenübersah. Das chinesische Unternehmen vertreibt seine Modelle *open weight*, d. h. die Modelle können frei heruntergeladen und weiterverwendet werden, Trainingsdaten und der gesamte Quellcode jedoch nicht.[52] Das zeichnet es allerdings nicht bloß als antagonistischen Verfechter von mehr Offenheit aus. Ähnlich Metas Llama-Modellen versprechen DeepSeeks offene, Hedgefonds-finanzierte Modelle, Knotenpunkt für externe Innovationen zu werden[53] und außerdem eine schnellere Adaption durch industrielle Nutzer*innen zu ermöglichen.[54] In gewisser Weise steht DeepSeek also bestehenden Plattformisierungsdynamiken entgegen, versucht sich diese aber zugleich zunutze zu machen.

«The difference between originality and imitation»

Wenn Liang behauptet: «[T]he real gap is the difference between originality and imitation»,[55] dann geht es ihm nicht um die Unterschiede zwischen einzelnen KI-Modellen, sondern um die Rolle Chinas und chinesischer Unternehmen in der weltweiten KI-Industrie. DeepSeeks Innovationen bei der Modellarchitektur und *close to the metal* lassen OpenAIs Anschuldigungen als primär strategisches Narrativ erscheinen: Es bedient sich historisch gewachsener Stereotype von China als Skalierer und Optimierer der Kostenstrukturen von Innovationen (hier KI), die ursprünglich im Westen auf den Weg gebracht wurden. Liang bezieht sich auf ebendieses nach wie vor wirkmächtige Stereotyp, bei dem chinesische Imitation westlicher Originalität gegenüberstellt wird. So wird etwa der republikanische US-Senator Josh Hawley im *Wall Street Journal* im März 2025 zur chinesischen Tech-Industrie folgendermaßen zitiert: «I don't think that they can do much innovation on their own, but they will if we keep sharing all this tech with them».[56] Wenn OpenAI in einem der US-Regierung vorgelegten Maßnahmenvorschlag DeepSeek als «state-subsidized» und «state-controlled» charakterisiert,[57] sind damit auch die geoökonomischen Dimensionen der US-Industriepolitik angesprochen. Denn die Beschwörung staatlich kontrollierter chinesischer Konkurrenz liefert Argumente für mehr staatliche Investition und weniger Regulierung von Unternehmen in den USA.[58] Hier zeichnet sich geoökonomisch einerseits die Überholung eines US-amerikanischen Konsens ab, nach dem neoliberaler Freihandel zugleich den Interessen kapitalistischer Unternehmen und dem US-amerikanischen Staat dienten, und andererseits wird hier dem Umstand Rechnung getragen, dass sich der chinesische Fokus hinsichtlich des eigenen Tech-Sektors von (Weiter-)Entwicklungsbemühungen zu Unabhängigkeitsbestrebungen verschoben hat.[59] Dieses Streben nach «originality» zeigt sich bei DeepSeek nicht nur darin, dass das Unternehmen ausschließlich junge und lokale Mitarbeiter*innen rekrutiert hat,[60] sondern auch daran, dass China nun, ungeachtet US-amerikanischer Exportbeschränkungen, selbst dazu übergegangen ist, US-amerikanische GPU-Importe als Sicherheitsrisiko zu diskreditieren oder vorerst ganz

52 Vgl. o. A.: Bring Us Your LLMs. Why Peer Review Is Good for AI Models, in: *Nature*, Bd. 645, Nr. 8081, 2025, 559, *doi.org/10.1038/d41586-025-02979-9*.

53 Vgl. Nick Srnicek: *Silicon Empires. The Fight for the Future of AI*, Cambridge (UK) 2025, 65.

54 Vgl. ebd., 138.

55 Liang Wengfeng zit. n. Schneider, Shen, Zhang: Deepseek.

56 Josh Hawley zit. n. Amrith Ramkumar: A Peter Thiel Protégé Is Leading Trump's AI Strategy Against China, *The Wall Street Journal*, 30.3.2025, *wsj.com/politics/policy/trump-ai-michael-kratsios-peter-thiel-protege-1457e276* (6.12.2025).

57 Christopher Lehane: o. T. [Konsultationspapier, geschickt an Faisal D'Souza, NCO at the Office of Science and Technology Policy], OpenAI, 13.3.2025, 1–15, hier 3, *cdn.openai.com/global-affairs/ostp-rfi/ec680b75-d539-4653-b297-8bcf6e5f7686/openai-response-ostp-nsf-rfi-notice-request-for-information-on-the-development-of-an-artificial-intelligence-ai-action-plan.pdf* (15.9.2025).

58 Vgl. Tan: DeepSeeking the Truth.

59 Vgl. Srnicek: *Silicon Empires*, 73–126.

60 Vgl. Tan: DeepSeeking the Truth.

abzulehnen.[61] Nach eigenen Aussagen im August 2025 ist DeepSeek mit chinesischen Chip-Produzenten im Gespräch, um die inländische KI-Hardware-Produktion in enger Absprache voranzutreiben.[62] In diesem Sinne ist auch die Zuspitzung von DeepSeeks R1-Veröffentlichung auf einen (an Sputnik erinnernden) «Moment» das Ergebnis einer westlichen Perspektive. Wie ein Blick auf die zurückliegende Entwicklung des chinesischen KI-Markts zeigt, verfolgte DeepSeek z. B. seine *open-weight*-Strategie auch, um damit größere inländische Marktkonkurrenten wie Alibaba, ByteDance, Tencent oder Baidu auf der Preisebene anzugreifen. Im Unterschied zu diesen hatte es DeepSeek schließlich geschafft, durch die angesprochenen Innovationen profitabel zu werden.[63] Zusätzlich begünstigte diese Strategie anschließend eine industrielle Adaption der Modelle. Innerhalb eines Monats wurden diese in Haushaltsgeräte, Krankenhäuser, Autos, staatliche Betriebe und private Unternehmen integriert; Liang nahm an Treffen mit Präsident Xi Jinping und Ministerpräsident Li Qiang teil.[64]

Droht der bloße Destillationsvorwurf gegen DeepSeek den Blick auf diese Aspekte also zu verstellen, lässt sich indessen durch seine weitere Einordnung das zentrale Zusammenspiel materieller politischer Ökonomie und Architektur-bedingter Repräsentationsregime explizieren.

Architektur und Kompression

Um diese wechselwirksame Entwicklung besser zu fassen, bedarf es der Integration einer Kritik KI-gestützter Wissensproduktion. Hierzu lässt sich aus medientheoretischer Perspektive fragen, wie konnektionistische KI Wissen über die Welt unter spezifischen Hardware-Bedingungen produziert. Jene Repräsentationsregime führen zum Anfang dieses Beitrags zurück: Worin besteht der neuartige Wert, den DeepSeek vermeintlich abgeschöpft hat, und wie fügt sich dieser in die materielle politische Ökonomie der KI ein? Wie gezeigt werden soll, stehen Architektur, Hardware und politische Ökonomie der heutigen KI-Industrie hier in einem sich wechselseitig validierenden Zusammenhang.

Dazu bedarf es eines Nachvollzugs jenes Übersetzungsprozesses, der von den Trainingsdaten zum Modell führt. Diese Kompression der Daten wird oft als Reduktion verstanden. Dabei bedeutet Kompression in Bezug auf KI zunächst die Umwandlung von Trainingsdaten in Vektoren im sogenannten *latent space*, dem multi-dimensionalen Repräsentationsraum des Modells. Damit hat Kompression eine ganz pragmatische Funktion und ist gleichsam ein Indikator für die Performance von KI: Ein Trainingsdatensatz der Größe x ist Grundlage für ein Modell der Größe <x, das die Logik der im Datensatz enthaltenen Informationen repräsentieren (und nachfolgend wiedererkennen und generieren können) soll.[65] In dieser Kompression lässt sich dann jedoch nicht bloß eine verlustreiche Reduktion sehen, sondern gerade die «verheißungsvolle Verlockung» der KI: Nach Louise Amoore u. a. wird der *latent space* als

61 Vgl. Raffaele Huang: Chinese Officials Urge Firms to Shun Nvidia AI Chip, *The Wall Street Journal*, 17.9.2025, wsj.com/tech/ai/china-nvidia-ai-chips-fbdbc30a (19.9.2025); Zijing Wu: China Set to Limit Access to Nvidia's H200 Chips Despite Trump Export Approval, *Financial Times*, 9.12.2025, ft.com/content/c4e81a67-cd5b-48b4-9749-92ecf116313d (13.12.2025).

62 Vgl. Dylan Butts: DeepSeek Hints Latest Model Will Be Compatible with China's ‹Next Generation› Homegrown AI Chips, MSN, 22.8.2025, msn.com/en-us/technology/hardware-and-devices/deepseek-hints-latest-model-will-be-supported-by-chinas-next-generation-homegrown-ai-chips/ar-AA1LoWom (5.9.2025).

63 Vgl. Schneider, Shen, Zhang: Deepseek.

64 Vgl. Srnicek: *Silicon Empires*, 138.

65 Vgl. Matteo Pasquinelli, Vladan Joler: The Nooscope Manifested. AI as Instrument of Knowledge Extractivism, in: *AI & Society*, Bd. 36, Nr. 4, 2021, 1263–1280, hier 1269, doi.org/10.1007/s00146-020-01097-6.

«höchst generativer Möglichkeitsraum» postuliert, der die Welt des Modells mit all seinen Limitierungen und Potenzialen konstituiert.[66] Medientheoretisch konturiert Jonathan Sterne den Begriff der Kompression noch schärfer. Er schreibt:

> [Media] have no existence apart from their containers and from their movements – or the possibility thereof. Compression makes infrastructures more valuable, capable of carrying or holding materials they otherwise would or could not, even as compression also transforms those materials to make them available to the infrastructure.[67]

Und weiter: «[C]ompression accommodates signals to infrastructures. But it also transforms infrastructures by enabling them to carry different kinds of signals.»[68] Die Geschichte der (konnektionistischen) KI ist also auch eine Geschichte der Kompression. Dies zeigt sich bereits an einer Untersuchung der Hardware. Doch ist, auch im Sinne Sternes, die heutige KI historisch nicht schlicht das Ergebnis dessen, dass nun für zuvor formulierte Ziele die passenden Hardware-Konfigurationen zur Verfügung stehen.[69] Heutige konnektionistische KI ist auch das Ergebnis eines epistemologischen Wandels, der mittlerweile ökonomische Effekte zeitigt. Vor diesem Wandel war KI gemeinhin noch als regelbasierte, logische und schrittweise Manipulation von Symbolen gedacht worden. Es bedurfte daher weniger paralleler Berechnungen als einer Serie sequenzieller Rechenschritte. *Parallel computing* wird also epistemologisch dann interessant, wenn KI entsprechend konzeptualisiert wird. Und ohne passende Hardware, die einen solchen Parallelismus ermöglicht, wären entsprechende Formen der Wissensproduktion wiederum nicht möglich.

Der Fall DeepSeek verdeutlicht in diesem Zusammenhang insbesondere zwei Dynamiken: Zum einen fällt er in eine Zeit ‹Nach-ChatGPT›, d. h. auch nach der breiten Adaption der Transformer-Architektur (das ‹T› in GPT) und damit ihrem Aufstieg zur Standardarchitektur gegenwärtiger KI-Modelle.[70] Vorgänger-Architekturen künstlicher neuronaler Netze hatten bis dahin stets ein *memory*-Problem, konnten also z. B. schlecht lange Eingabesequenzen verarbeiten.[71] Der von Google im Jahr 2017 entwickelte Transformer bedeutet die Überführung von Eingabesequenzen (wie z. B. Sätzen, aber auch Bildern) in parallel zu prozessierende Matrizen. Der Transformer ist somit auch erst nach dem erfolgreichen Einsatz von GPUs sinnvoll implementierbar und fügt sich zudem nahtlos in eine «bigger is better»-Ideologie ein: Parallelisierbarkeit ist besonders für diejenigen interessant, die bereits über die entsprechende Hardware verfügen oder diese mieten oder kaufen können. Parallelisierbarkeit ist damit nicht nur Teil einer epistemologischen Wende, sondern zentrales Merkmal einer Architektur wie der des Transformers, das diese kompatibel mit einer bereits vorhandenen Infrastruktur aus Datenzentren und Cloud-Computing macht.[72]

Doch machen nicht bloß die Komponenten und Betriebskosten der Datenzentren den Wert eines KI-Modells aus. Der Fall DeepSeek zeigt auch, dass

66 Vgl. Louise Amoore u. a.: A World Model. On the Political Logics of Generative AI, in: *Political Geography*, Bd. 113, Artikelnr. 103134, 2024, 1–9, hier 4, doi.org/10.1016/j.polgeo.2024.103134, Übers. YNF.

67 Vgl. Jonathan Sterne: Compression. A Loose History, in: Lisa Parks, Nicole Starosielski (Hg.): *Signal Traffic. Critical Studies of Media Infrastructures*, Urbana 2015, 31–52, hier 36.

68 Ebd., 34.

69 Vgl. Rella: Close to the Metal, 16.

70 Vgl. Ashish Vaswani u. a.: Attention Is All You Need, *arXiv*, 2.8.2023 (7. Fassung), 1–15, doi.org/10.48550/arXiv.1706.03762 (22.1.2025).

71 Vgl. Yoshua Bengio, Patrice Simard, Paolo Frasconi: Learning Long-Term Dependencies with Gradient Descent Is Difficult, in: *IEEE Transactions on Neural Networks*, Bd. 5, Nr. 2, 1994, 157–166, doi.org/10.1109/72.279181.

72 Vgl. Luitse, Denkena: The Great Transformer, 7.

die Proliferation von Techniken zur Manipulation multidimensionaler Vektorräume die Privatisierung kollektiver Arbeit und kollektiven Wissens in einigen wenigen KI-Modellen vorangetrieben hat. Wenn Hardware, wie bereits diskutiert, schon immer auch epistemologisch und eingebettet in eine politische Ökonomie ist, wie lässt sich dann das per KI modellierte Wissen ökonomisch fassen? Erste theoretische Einordungsversuche lassen sich über eine Verbindung des Kompressionsbegriffs mit der materiellen politischen Ökonomie der KI situieren.

Kommensurabilität und Verwertbarkeit

Knowledge distillation ist eine Technik, die dazu dient, in KI-Modelle übersetztes Wissen konzentriert zu extrahieren. Die Plausibilität dieser Begrifflichkeit fußt auf den Repräsentationen dieses Wissens im *latent space*, die Ergebnis von Vektorisierungs- bzw. Kompressionsprozessen sind. In ihren Ausführungen zum repräsentationellen Charakter von KI-Modellen betont M. Beatrice Fazi, dass Kompression immer einen gewissen Grad an Vereinheitlichung von Informationen mit sich bringt.[73] Eine Kompression der Trainingsdaten bedeute eine «Strukturierung», also eine Modellierung von Wissen. Es gehe darum, jedwede Datenpunkte in einem Raum aufeinander beziehen zu können. Fazi nennt das eine «togetherness of distributed representations».[74]

Diese vermeintlich «universelle Kommensurabilität», so beschreiben es Fabian Offert und Leonardo Impett, liegt in der doppelten Natur, die sämtliche Medien im Zuge generativer KI annehmen: Hier sind Medien z. B. Text, Bild oder Audio und vektorisierte *embeddings* im *latent space* des Modells.[75] Offenkundig hat diese Idee universeller Kommensurabilität ein ideologisches Moment: Jene Übersetzungsprozesse unterliegen schließlich spezifischen Medialitäten, die sich über die verwendeten Daten, Interfaces und Algorithmen bis zu den Modellen spannen. So ist die Idee ‹originaler› Trainingsdaten ideologisch motiviert, sind diese in Form von Datenbanken doch immer schon soziales Konstrukt,[76] und verschiedene Modelle bedeuten verschiedene Vektor- und damit Repräsentationsräume. Dennoch verdeutlichen die diskutierten Zugriffsbeschränkungen (nicht nur im Hinblick auf Hardware-Komponenten sondern auch auf Modelle reguliert über ihre Interfaces), dass auch diese Sphäre geoökonomischen Strategien unterworfen werden könnte. Diese Sphäre ist nicht die der Hardware, sondern des sich in den Modellen befindlichen neuartigen Kapitals – dieses besteht in der Relationierung von Trainingsdaten und Modellparametern, von kollektivem Wissen, das bereits überführt wurde in «Maschinenwissen».[77]

Vor dem Hintergrund dieser geoökonomischen Zuspitzungen zeichnet sich die zentrale Rolle der Kontrolle über Plattformen und Interfaces für die Fortentwicklung von KI ab. Abschließend möchte ich zu jenem Moment zurückkehren, an dem ebendiese Kontrollierbarkeit über die Hardware erkannt wurde.

73 Vgl. M. Beatrice Fazi: The Computational Search for Unity. Synthesis in Generative AI, in: *Journal of Continental Philosophy*, Bd. 5, Nr. 1, 2024, 31–56, hier 35, doi.org/10.5840/jcp202411652.

74 Vgl. ebd., 47.

75 Vgl. Fabian Offert: «Vector Media». Towards a Materialist Epistemology of Artificial Intelligence [Video vom Vortrag, gehalten am 13.3.2025 am Berkeley Department of German], hochgeladen von @Sunrise Lecture on Media and Technology am 17.3.2025 auf *YouTube*, youtu.be/xNf7Y5d7tow (7.12.2025), TC 00:05:34–00:05:43, Übers. YNF.

76 Vgl. Jonathan Roberge, Tom Lebrun: Parrots All the Way Down. Controversies within AI's Conquest of Language, in: Richard Groß, Rita Jordan (Hg.): *KI-Realitäten. Modelle, Praktiken und Topologien maschinellen Lernens*, Bielefeld 2023, 39–65, hier 46, doi.org/10.14361/9783839466605-003.

77 Matteo Pasquinelli: Vectors for Workers. Models of Automation and Autonomy in the Long AI Century [Vortrag gehalten am 7.11.2025 an der SOAS University of London, Pre-Print], 7.11.2025, 1–20, hier 12, hdl.handle.net/10278/5099627 (25.11.2025), Übers. YNF. Offert und Impett sprechen von «neural exchange value». Vgl. Offert: «Vector Media», TC 00:34:19–00:35:14.

KI und die ‹Hardware-Wende›

Die in vielfacher Hinsicht kritische Rolle, die die Hardware von KI-Systemen heute einnimmt, lässt sich mit Floridi als Teil einer rezenten ‹Hardware-Wende› beschreiben. Entscheidend ist hierbei, dass es Floridi nicht bloß um eine Betrachtung der materiellen Zusammensetzung von Technologie geht, die schließlich schon immer vorhanden war, sondern um die Tatsache, dass diese materielle Grundlage erst kürzlich als Kern der Diskussionen um «digitale Souveränität» in den Fokus gerückt ist (wie etwa durch den CHIPS and Science Act).[78] Unter digitaler Souveränität versteht Floridi dabei die Kontrolle über ‹das Digitale›, worunter er Daten und Software (z. B. KI), Protokolle und Standards (z. B. Domain-Namen), Prozesse (z. B. Cloud-Computing), Hardware (z. B. GPUs), Services (z. B. Social Media) und Infrastrukturen (z. B. Kabel oder Satelliten) versteht.[79]

Für Floridi ist der Zusammenhang diverser Beziehungsgeflechte zentral. Er sucht mit der ‹Hardware-Wende› nicht bloß jedwede «immaterielle Erfahrung» im Kontext des Digitalen in ihrer Hardware-Bedingtheit herauszustellen, sondern auch hervorzuheben, dass das Materielle eine Frage von Beziehungen ist, physisch wie sozial, inklusive Fragen von Souveränität, Kontrolle und Macht.[80] Laut Floridi wurde die Debatte um die «digitale Revolution» in der Vergangenheit als Debatte um «immaterielle Erfahrung» geführt.[81] Die historischen Gründe dafür finden sich schon bei Claude Shannons entkörperlichter theoretischer Konzeption digitaler Signalverarbeitung oder bei Alan Turings formalistischer, symbollogischer Auffassung von (Künstlicher) Intelligenz.[82] Die akademische Beschäftigung mit materiellen Aspekten sei, so Floridi, dabei kein neues Phänomen.[83] Gegenwärtig würde ‹das Digitale› allerdings nicht nur unsere Ontologie, also unsere Konzeption von Realität, als zunehmend vernetzt verändern (wie etwa durch die Idee universeller Kommensurabilität). Hatte ein populärer Diskurs des ‹Immateriellen› lange wirtschaftliche Vorzüge, würde nun auf politischer Ebene auch die Kontrollierbarkeit seiner materiellen Bedingtheit erkannt werden.[84] Obschon die ‹digitale Revolution› ein ideologisch geprägter Begriff ist,[85] lassen sich Floridis Ideen dennoch auf die technisch neuartige Entwicklungslinie gegenwärtiger KI seit den 2010er Jahren beziehen. Der Fall DeepSeek demonstriert die entschiedene Hinwendung zur Hardware exemplarisch.

Nach der ‹Hardware-Wende›

Microsoft-CEO Satya Nadella beschwichtigte Investor*innen angesichts der Infragestellung ebenjener US-Souveränität mit einem Post auf X. Am 27. Januar 2025 referierte er auf das Jevons-Paradoxon: Wenn technischer Fortschritt eine effektivere Nutzung einer Ressource erlaube – wie bei KI durch DeepSeeks optimierte Nutzung von GPUs –, dann führe das nicht zu einer geringeren Nutzung dieser Ressource, sondern zu einer breiteren Anwendung und daher

78 Vgl. Floridi: The Hardware Turn in the Digital Discourse, 4, Übers. YNF. Srnicek zeichnet diese Entwicklung zwischen China und den USA detaillierter nach. Vgl. Srnicek: *Silicon Empires*, 73–126.

79 Vgl. Luciano Floridi: The Fight for Digital Sovereignty. What It Is, and Why It Matters, Especially for the EU, in: *Philosophy & Technology*, Bd. 33, Nr. 3, 2020, 369–378, hier 370 f., doi.org/10.1007/s13347-020-00423-6.

80 Vgl. ders.: The Hardware Turn in the Digital Discourse, 5, Übers. YNF.

81 Vgl. ebd., 4, Übers. YNF.

82 Vgl. N. Katherine Hayles: *How We Became Posthuman. Virtual Bodies in Cybernetics, Literature, and Informatics*, Chicago 1999.

83 Neben der von Floridi zitierten Literatur, die vor allem politische, ökonomische und ethische Perspektiven abbildet, sind medienwissenschaftlich besonderes jüngere öko-materielle Arbeiten und frühere Medientheorien zu erwähnen, z. B. Jennifer Gabrys: *Digital Rubbish. A Natural History of Electronics*, Ann Arbor 2011; Jussi Parikka: *A Geology of Media*, Minneapolis 2015; Harold A. Innis: *Empire and Communications*, Toronto 1975; Marshall McLuhan: *Die magischen Kanäle. «Understanding Media»*, Düsseldorf u. a. 1992; Friedrich A. Kittler: Es gibt keine Software, in: ders.: *Draculas Vermächtnis. Technische Schriften*, Leipzig 1993, 225–242.

84 Vgl. Floridi: The Hardware Turn in the Digital Discourse, 3.

85 Vgl. Gabriele Balbi: *The Digital Revolution. A Short History of an Ideology*, Oxford 2023.

gesteigerten Nachfrage, «turning it into a commodity we just can't get enough of».[86] Eine gesteigerte Gesamtnachfrage würde so zum Vorteil aller (und nicht zuletzt zum Vorteil von Microsoft samt der Rechenkapazität bereitstellenden Cloud-Computing-Plattform Azure). Nadella erwähnt dabei nicht die geoökonomische Relevanz der Kontrolle über Wertschöpfungsketten und Ressourcen und damit auch nicht, dass eine US-amerikanische Plattformisierung von KI erst ein Anstoß für jenen technischen Fortschritt bei DeepSeek gewesen sein mag. Auch wenn DeepSeek mit dem US-amerikanischen Tech-Sektor in Konkurrenz steht – neun Tage nach Veröffentlichung stellte Microsoft DeepSeeks Modell auf Azure zur Nutzung bereit.[87]

So eröffnet sich für den Ausgangspunkt dieses Beitrags ein vorsichtiger Ausblick. Der ‹DeepSeek-Moment› ist nicht nur das Ergebnis veralteter Stereotype, er ist auch Symptom einer Industrie, in der sich KI-Forschung (wie etwa zu Hardware oder Modellarchitekturen), geoökonomische Strategisierungen wirtschaftlicher Verflechtungen (etwa mittels Handelszöllen) und Plattformisierungsdynamiken (von proprietären Software-Ebenen bis hin zu ‹offenem› Modellvertrieb) gegenseitig bedingen und dabei KI einer Dynamik unterwerfen, der aufgrund dieser Wechselwirkungen schwer zu entkommen ist. Wie dieser Beitrag demonstriert hat, war die Entwicklung gegenwärtiger KI außer von der Software maßgeblich von der Nutzbarmachung vorhandener Hardware abhängig, was die konsequente Optimierung dieser Komponenten auf den Anwendungsbereich der KI (wie z. B. bei Nvidia oder bei Googles TPUs) zur Folge hatte. Das heißt perspektivisch jedoch auch, dass neuartige KI-Forschungsansätze ökonomisch weit weniger tragfähig sind, sollten sie nicht zu vorhandenen Hardware- und Software-Komponenten passen – eine Ausgangslage, der sich konnektionistische KI vor den 2010ern selbst gegenübersah. Oder, wie es Hooker fragend formuliert: «What are the failures we still don't have the hardware and software to see as a success?»[88] Und darüber hinaus: Wenn also unter anderem die Hardware einen so großen Einfluss auf den Erfolg von Forschung hat, ist eine offene Frage, inwieweit staatliche Kontrollbestrebungen über Hardware diese Forschung beeinflussen wird. Genauso ist offen, ob aus diesen staatlichen Bestrebungen eine beständige Zusammenarbeit mit privatwirtschaftlichen Firmen erwachsen wird.

Indessen deutet sich in Nadellas Bezeichnung von KI als «commodity» eine spezifische Plattformisierung der KI als KI *as a service* an: KI-Modelle, die nurmehr als fertige Produkte genutzt und gehandelt werden können, die aber zunehmend nicht mehr offen vertrieben oder erforscht werden. Anbieter*innen entsprechender Cloud-Infrastruktur wären hier die Profiteur*innen – eine Marktentwicklung, die sich auch darin abzeichnet, dass die eigene Erzeugung von oder der Handel mit Energie durch KI-Unternehmen mittlerweile als Wettbewerbsvorteil verstanden wird.[89]

Sind diese Dynamiken maßgeblich für die gegenwärtige KI-Industrie, findet die vorliegende Perspektive ihre Grenzen bei den Folgen der Verwendung

86 Satya Nadella @satyanadella: Jevons Paradox Strikes Again! As AI Gets More Efficient and Accessible, We Will See Its Use Skyrocket, Turning It into a Commodity We Just Can't Get Enough of. en.m.wikipedia.org/wiki/Jevons_paradox, X, 27.1.2025, x.com/satyanadella/status/1883753899255046301 (4.8.2025).

87 Reuters: Microsoft Rolls out DeepSeek's AI Model on Azure, *Reuters*, 29.1.2025, reuters.com/technology/artificial-intelligence/microsoft-rolls-out-deepseeks-ai-model-azure-2025-01-29 (6.12.2025).

88 Hooker: The Hardware Lottery, 7.

89 Vgl. z. B. Josh Saul, Riley Griffin, Naureen S. Malik: Meta Looks to Power Trading to Support Its AI Energy Needs, *Bloomberg*, 21.11.2025, bloomberg.com/news/articles/2025-11-21/meta-enters-power-trading-to-support-ai-data-centers (28.11.2025).

der KI-Modelle selbst. Dies meint nicht nur die Untersuchung KI-generierter Outputs. Während dieser Beitrag sich überwiegend auf *räumliche* Dimensionen bezogen hat (etwa als Analyse der Hardware und globaler Wertschöpfungsketten und im Sinne eines *metonymic turn*), deutet der Fall DeepSeek auch auf neuartige *Zeitskalen* hin. Im so oft beschworenen ‹KI-Wettrennen› scheint es nicht mehr nur darum zu gehen, wann die Trainingsdaten erstellt wurden, wie schnell ein Modell entwickelt und trainiert werden kann oder wie viel Zeit zwischen Prompt und Output und bei KI-Agenten zwischen Aktion und Reaktion vergeht.[90] Zunehmend zentral wird, wie schnell semiotische Prozesse jedweder Art in neue Kontexte der menschlichen oder maschinellen Wissensproduktion eingefügt werden können, was spezifischen Medialitäten unterliegt.[91] Diese Medialitäten stehen wiederum in direktem Zusammenhang mit den hier besprochenen Interface-Dynamiken, Hardware-Konfigurationen und skizzierten Verwertungslogiken, also der materiellen politischen Ökonomie der KI.

90 Vgl. Paul Kockelman: *Last Words. Large Language Models and the AI Apocalypse*, Chicago 2024, 98.
91 Vgl. ebd., 104.

Für das Durchsehen früherer Versionen dieses Textes und wertvolle Hinweise möchte ich mich bei Hannes Bajohr, Moritz Konrad, Lars Pinkwart und Magnus Rust bedanken.