Large Language Models

Katia Schwerzmann

1. Zur Verteidigung der universitären Autonomie

Dieser Text befasst sich mit der rasanten Akzeptanz der Verwendung von Large Language Models (LLMs) an deutschen Universitäten im Kontext von Lehre und Forschung.[1] In ihrer Einleitung zu diesem Band laden Patrizia Breil und Florian Sprenger dazu ein, über die Universität als einen Ort nachzudenken, der virtuell im Sinne von ›welterzeugend‹ ist – einen Ort also, an dem es möglich ist, ›über das Gegebene hinaus‹ zu denken. Ich frage in diesem Beitrag, unter welchen Bedingungen die Universität ihre Offenheit für das bewahren kann, was nicht ›gegeben‹ ist. Dabei denke ich zum Beispiel an mögliche Zukünfte, neue Wissensfelder, neue Formen des Studiums oder neue Beziehungen zu den Mitgliedern der Institution und der Gesellschaft.

Erste Hypothese: Die Universität kann ihre Offenheit gegenüber dem, was über das Gegebene hinausgeht, nur bewahren, wenn sie fähig bleibt, ihre eigenen Regeln zu formulieren. Autonomie als Fähigkeit, eigene Regeln zu bilden, ist nie absolut, sondern relational, das sie stets gegen die Heteronomie anderer Bereiche ausgehandelt wird. Im Fall der Universität sind diese anderen Bereiche gewöhnlich das Politische und das Wirtschaftliche. Auf ihre Autonomie zu verzichten, und sei es nur im Sinne eines regulativen Ideals, würde für die Universität bedeuten, sich dem ›Gegebenen‹ zu ergeben und das Virtuelle als das, was gerade ›jenseits des Gegebenen‹ liegt, aufzugeben.

Was passiert, wenn sich die Universität in ein von Heteronomie geprägtes Feld begibt, dessen ›Regeln‹ sich ständig verändern, implizit bleiben und sich ihrer Hinterfragung und aktiven Gestaltung entziehen? Das Feld, um das es hier geht, ist kein anderes

1 An der Ruhr-Universität Bochum werden umfangreiche Schulungen für Professor*innen und Dozent*innen zur Verwendung von LLMs und generativer KI in Lehre und Forschung angeboten (Ruhr Universität Bochum 2025). Die RUB bietet sogar eine eigene »datenschutzfreundliche« Version von GPT an (ebd.). Die Bewertung von studentischen Arbeiten mit Hilfe von KI ist sogar erlaubt, solange das Ergebnis der Bewertung »kritisch geprüft« wird, wobei die »Festlegung der Note […] ausschließlich durch den für die Bewertung zuständigen Menschen erfolgen« kann (ebd.). In den USA bietet OpenAI teure Versionen seiner Werkzeuge für Forschungszwecke an. Unter anderem hat die University of California eine Partnerschaft mit OpenAI geschlossen, um eine »bildungsspezifische Version« von GPT einzuführen (Kant 2025).

als das der generativen künstlichen Intelligenz, genauer gesagt, der LLMs. LLMs sind Machine-Learning-Modelle, die mithilfe von Big-Data trainiert werden und Outputs generieren, indem sie die Abfolge von Token (Wortteilen) vorhersagen, die statistisch gesehen die höchste Wahrscheinlichkeit aufweisen, auf die Token des Inputs zu folgen. Nachdem das Modell trainiert wurde, indem es statistische Muster aus dem Trainingsdatensatz erfasst hat, ist es in der Lage, Inputs einzukodieren und dabei auch die relative Wichtigkeit jedes Tokens in seinen Relationen zu den Nachbar-Token zu berücksichtigen.[2]

Der generative Charakter dieser Art von KI beruht auf der Konstruktion eines virtuellen *representation space*, der nicht nur eine Darstellung dessen ist, was in der Welt zählt (Amoore et al. 2024); der *representation space* ist zugleich auch *weltbildend* aufgrund seiner spezifischen Normativität, die dank des »künstlichen Naturalismus« von Machine Learning den Anschein von Wahrheit und Legitimität erhält (Campolo/Schwerzmann 2023: 2–3). Dieser künstliche Naturalismus entstammt der Überzeugung, dass die Welt – als statistische, der menschlichen Wahrnehmung entgehenden Struktur – sich im Big Data widerspiegelt und dass das Modell diese Struktur während des Trainings genau erfasst. Infolgedessen gewinnt der Output des Modells den Effekt der Selbstverständlichkeit bzw. Gegebenheit, der nur schwer zu hinterfragen ist.

Was die generative KI kennzeichnet, möchte ich als *normative Rationalität* bezeichnen. Daraus folgt die zweite Hypothese: Diese normative Rationalität ist eine neue Art algorithmischer Rationalität, die sich deutlich von früheren Formen der regelbasierten KI unterscheidet (Symbolic AI, Expert System). In ihrem Buch *How Reason Almost Lost Its Mind* ordnen der Historiker Erickson und seine Kollegen diese früheren Formen der KI in eine »Rationalität des Kalten Krieges« ein – eine Rationalität, die menschliches Urteil und Interpretation so weit wie möglich ausschließt und daher als besonders geeignet gilt, Ungewissheit zu minimieren (Erickson et al. 2013; Daston 2022). Im Gegensatz dazu stützt sich die aktuelle, ML-basierte KI ausdrücklich auf die Automatisierung von moralischen Urteilen, Normen, Werten und Interpretationen.

Nach Foucault sind »Normen« sowohl Mittel zur Steuerung des Verhaltens (»la conduite des conduites«) als auch Vorstellungen dessen, was die Gesellschaft sein könnte oder sollte (Foucault 2021). Normen stabilisieren nicht nur soziale Ordnungen, sondern spielen auch eine entscheidende Rolle bei der Gestaltung neuer Ordnungen und Formen der Sozialität. Normen verkörpern Werte, die im Sinne Nietzsches qualitative »Schätzungen« der Welt sind, als etwas, das uns angeht (Nietzsche 1887: 190). Die normative Rationalität der KI präsentiert und generiert das Bild dessen, was sich die Entwickler von Modellen für die Gesellschaft wünschen.

Während die Outputs der ML-Modelle die dem Datensatz immanenten statistischen Regelmäßigkeiten widerspiegeln, ist eine solche Widerspiegelung aus mehreren Gründen nicht bloß deskriptiv, sondern auch normativ: Erstens verkörpern die Trainingsdaten Wertungen, die durch spezifische Machtgefüge innerhalb der Gesellschaft, aus der sie stammen, geprägt sind; zweitens werden große Datensätze sowohl *kuratiert* als

2 Dieser Prozess wird gemeinhin als *attention mechanism* bezeichnet (Vaswani et al. 2017) und ist einer der Gründe für den Erfolg von LLMs.

auch mithilfe von *Feature-Engineering* bearbeitet, um die Fähigkeit der Modelle zu optimieren, daraus Regelmäßigkeiten abzuleiten. Daten durchlaufen somit einen Prozess der ›Exemplifizierung‹ (Campolo/Schwerzmann 2023). Drittens machen generative Modelle eine zweite Trainingsphase durch, die *Fine-Tuning* genannt wird und deren Zweck es ist, die Verhaltensweise (*behavior*) der Modelle zu gestalten und zu regulieren (Ouyang et al. 2022; OpenAI 2025). *Fine-Tuning* besteht darin, die Modelle an einer Reihe von Werten auszurichten, die widerspiegeln, was die Entwickler*innen von Modellen (große Technologieunternehmen sowie nationale Akteure) als erwünschte Verhaltensweisen von Modellen und Menschen ansehen (Schwerzmann/Campolo 2025). Zu diesem Zweck wird in generative Modelle eine explizitere Art von Normativität eingeführt. Hierbei schreiben und bewerten menschliche *annotators* eine relativ niedrige Anzahl von Beispielen (Prompts und Antworten), die die als erwünscht festgelegten Werte verkörpern. Das Modell wird anhand dieser Beispiele erneut trainiert, was als *reinforcement learning* bezeichnet wird.

2. Hinweis zum Virtuellen

Wie ist das Virtuelle hier zu verstehen? Das Virtuelle ist in seinen Wirkungen ebenso real wie das Aktuelle. In seiner Auseinandersetzung mit Bergson, Proust und Deleuze bezeichnet Rob Shields das Virtuelle als eine »reale Idealisierung« und beschreibt es als all das, was »fast so« ist (Shields 2005: 28). Wie das Reale hat das Virtuelle eine physische Wirkung auf die Welt. Um das Verhältnis zwischen dem Virtuellen und dem Aktuellen zu verdeutlichen, verstehe ich das Virtuelle als das Latente, das, was sich an der Peripherie des Aktuellen befindet und aus dem sich das Aktuelle kristallisieren kann. Ich schlage vor, dass das Verhältnis zwischen dem Virtuellen und dem Aktuellen *quantitativer* oder *qualitativer* Art sein kann. In quantitativer Hinsicht unterscheidet sich das Virtuelle vom Aktuellen, wie die Wahrscheinlichkeit, dass etwas geschieht, sich vom tatsächlichen Geschehen unterscheidet. So gesehen stellt das Virtuelle einen Raum für das Modellieren, Extrapolieren und Experimentieren mit dem Realen dar. Das quantitative Virtuelle haftet am ›Gegebenen‹ – im vorliegenden Fall an den Datensätzen, die durch einen indexikalischen Charakter gekennzeichnet sind (Kitchin/McArdle 2016). In deren quantitativen Verhältnis fungiert das Aktuelle als Norm oder Maßstab für das Virtuelle. Qualitativ unterscheidet sich das Virtuelle vom Realen auf quasi-ontologische Art und ist auf das Reale nicht reduzibel; (Science-)Fiction, Spekulation, Spiele u.a. eröffnen alle eine Vielzahl von Welten, die nicht notwendigerweise an das ›Gegebene‹ gebunden sind.[3] Das qualitative Virtuelle erlaubt uns, das ›Gegebene‹ zu hinterfragen, indem es uns ermöglicht, uns vorzustellen, wie die Welt ganz anders *sein könnte*.

Ausgangspunkt dieses Beitrags war die Idee der Virtuellen Universität als Ort, an dem eine Pluralität von künftigen Welten eröffnet wird; ein Ort, der sich weigert, sich unkritisch der Heteronomie benachbarter Felder und dem ›Gegebenen‹ unterzuordnen.

3 Diese Unterscheidung zwischen einem quantitativen und einem qualitativen Virtuellen entspricht der Unterscheidung zwischen Extrapolation und (Science-)Fiction, auf der Ursula K. Le Guin in ihrer Einleitung zu *The Left Hand of Darkness* (1969/2019) insistiert.

Dennoch hat die Universität LLMs und mit ihnen deren *representation spaces* sehr schnell und unkritisch übernommen.

3. Das Virtuelle der generativen KI

Wie andere generative Modelle funktionieren LLMs auf der Grundlage eines hochdimensionalen Raums, der aus *representations* besteht, die das Modell aus den Trainingsdaten konstruiert.[4] Dieser Vektorraum ermöglicht und beschränkt zugleich, was das Modell generiert. Der oben vorgeschlagenen Typologie zufolge ist der *representation space* eines generativen Modells quantitativ virtueller Art: Das Modell erfasst die Wahrscheinlichkeitsverteilung des input space, der aus den ›aktuellen‹, d.h. Trainings-Daten besteht, indem es sie im niedrig dimensionalen *representation space* komprimiert (Bengio/Courville/Vincent 2013).

Um Outputs zu generieren, aktualisiert das Modell einen Bereich des *representation space* – ein Verfahren, das als Interpolation bezeichnet wird (Arvanitidis/Hansen/Hauberg 2021; Chollet 2021). Dass generative Modelle interpolieren und nicht extrapolieren, ist indes von Bedeutung. Denn bislang erzeugen generative Modelle keine Outputs, die über die Wahrscheinlichkeitsverteilung der Trainingsdaten hinausgehen (was als *Extrapolation* zu verstehen wäre), sondern ›*sampeln*‹ den *representation space zwischen* den gelernten Repräsentationen (*Interpolation*). Die Fähigkeit der Modelle, noch nicht ›gesehene‹ Daten zu verarbeiten, geschieht daher innerhalb der Grenzen der Daten, die sie bereits ›gesehen‹ haben (die Trainingsdaten).

Einen weiteren bedeutsamen Punkt gilt es zu beachten: Bereiche des *representation space* werden als ›dicht‹ bezeichnet, wenn sie eine Vielzahl an Repräsentationen bzw. Vektoren enthalten, die aus dem Training des Modells entstanden sind. In dichten Bereichen ist die Gewissheit des generierten Outputs höher als in vergleichbar weniger dichten Bereichen (Arvanitidis/Hansen/Hauberg 2021: 4). Wenn das Modell außerhalb dieser dichten Zonen zum Output aufgefordert wird, neigt es zu »Halluzinationen« und liefert ungenaue Ergebnisse (Ji et al. 2023). Zweck des ›Prompt-Engineering‹ ist es demzufolge, die ›Aufmerksamkeit‹ (*attention*) des Modells auf die dichteren Bereiche des *representation space* zu lenken, was eine ›bessere‹ Verarbeitung des Prompts ermöglicht.

Die Virtualität des *representation space* der LLMs ist quantitativer Art, da all das, was das Modell generiert, an die aktuellen, indexikalischen Daten gebunden bleibt, die die Norm darstellen, auf die synthetische Daten bei deren Generierung gerichtet sind.

4. Die Normativität des representation space

Indem die Universität LLMs akzeptiert, bevor sie einen kritischen Rahmen für ihre Verwendung in pädagogischen und forschungsbezogenen Kontexten entwickelt hat, lässt

4 Bei LLMs wird dieser *representation space* »vector space« genannt, während er bei Image Generation Models als »latent space« bezeichnet wird. Für eine kritische Analyse des *latent space* von GANs siehe Offert 2021.

sie sich durch volatile und oft implizite Normen leiten, die den virtuellen *representation space* der generativen Modelle ausmachen. Diese Normen könnten in absehbarer Zeit den Ideologien der Trump-affinen Tech-Unternehmer im Silicon-Valley entsprechen und haben sich im Fall von DeepSeek bereits an der chinesischen Staatsdoktrin ausgerichtet. Mit ihrem Anspruch, »neutral and balanced« (ChatGPT) zu sein, verschleiern LLMs nicht nur systematisch die ethisch-technische Arbeit, die nötig ist, um diesen Anspruch auf normative Neutralität zu produzieren; sie neutralisieren auch ihre Positionalität, indem sie weniger den Blick von oben beanspruchen als einen Blick von überall, der aus der Wirklichkeit selbst entstehen und sie in ihrer Totalität repräsentieren würde (Schwerzmann 2025).

Gewährt die Universität der normativen Rationalität generativer Modelle freie Verbreitung, besteht das Lernen und Forschen zunehmend darin, die Ergebnisse der Modelle zu bewerten und zu korrigieren. Anstatt den Studierenden beizubringen, aktiv zur Produktion von Sinn, Analysen und Argumenten aus ihrer singulären Perspektive beizutragen, lernen sie, synthetische Ergebnisse zu bewerten, zu modulieren und zu verbessern. Eine solche Pädagogik geht dann von der normativen Rationalität generativer Modelle aus und bleibt daher in ihren Werten und ihrer Optimierungslogik gefangen. Würde Forschung zunehmend als evaluierendes *Supplement* des Outputs dieser Modelle fungieren, könnte die idiosynkratische und kreative Dimension der Wissensproduktion verloren gehen.

Am Anfang habe ich das ›Gegebene‹ als den gegenwärtigen und zugleich künftig erwarteten Zustand der Welt definiert – ein Zustand, der durch Technologie maßgeblich bestimmt wird. Diese Definition muss aber erweitert werden: Das ›Gegebene‹ ist *Effekt* und *Affekt* der aktuellen KI und ihrer normativen Rationalität. Will die Universität sich der Kapitulation vor dem ›Gegebenen‹ in Form des *representation space* generativer Modelle widersetzen, um ein Ort des Experimentierens zu bleiben, sollte das Virtuelle dort nicht nur quantitativ-*inter*polierend sein, sondern sowohl quantitativ-*extra*polierend als auch qualitativ bleiben. Medienwissenschaft und Technikphilosophie sollten dazu beitragen, die Logik der Technologie und ihre naturalisierenden, normgebenden Operationen kritisch zu untersuchen, um den Raum für das Nicht-Gegebene offen zu halten.

Literatur

Amoore, Louise/Campolo, Alexander/Jacobsen, Benjamin/Rella, Ludovico (2024): »A World Model: On the Political Logics of Generative AI«, in: Political Geography 113, S. 1–9. https://doi.org/10.1016/j.polgeo.2024.103134.

Arvanitidis, Georgios/Hansen, Lars Kai/Hauberg, Søren (2021): »Latent Space Oddity: On the Curvature of Deep Generative Models«, in: arxiv.org (13.12.2021). Online unter: https://arxiv.org/abs/1710.11379v3 (letzter Zugriff: 07.06.2025). https://doi.org/10.48550/arXiv.1710.11379.

Bengio, Yoshua/Courville, Aaron/Vincent, Pascal (2013): »Representation Learning: A Review and New Perspectives«, in: IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (8), S. 1798–1828. https://doi.org/10.1109/TPAMI.2013.50.

Campolo, Alexander/Schwerzmann, Katia (2023): »From Rules to Examples. Machine Learning's Type of Authority«, in: Big Data & Society 10 (2), S. 1–13. https://doi.org/10.1177/20539517231188725.

Chollet, François (2021): Deep Learning with Python, Shelter Island, NY: Manning Publications Co.

Daston, Lorraine (2022): Rules: A Short History of What We Live By, Princeton: Princeton University Press. https://doi.org/10.2307/j.ctv27qzsfn.

Erickson, Paul/Klein, Judy L./Daston, Lorraine/Lemov, Rebecca/Sturm, Thomas/Gordin, Michael D. (2013): How Reason Almost Lost Its Mind: The Strange Career of Cold War Rationality, Chicago/London: The University of Chicago Press.

Foucault, Michel (2021): Die Strafgesellschaft: Vorlesungen am Collège de France 1972–1973, Berlin: Suhrkamp.

Ji, Ziwei/Lee, Nayeon/Rita Frieske/Yu, Tiezheng/Su, Dan/Xu, Yan/Ishii, Etsuko/Bang, Ye Jin/Madotto, Andrea/Fung, Pascale (2023): »Survey of Hallucination in Natural Language Generation«, in: ACM Computing Surveys 55 (12), S. 1–38. https://doi.org/10.1145/3571730.

Kant, Rishi (2025): »OpenAI Targets Higher Education in the U.S. with ChatGPT Rollout at California State University«, in: reuters.com (04.02.2025). Online unter: https://www.reuters.com/technology/openai-targets-higher-education-us-with-chatgpt-rollout-california-state-2025-02-04/ (letzter Zugriff: 07.06.25).

Kitchin, Rob/McArdle, Gavin (2016): »What Makes Big Data, Big Data? Exploring the Ontological Characteristics of 26 Datasets«, in: Big Data & Society 3 (1), S. 1–10. https://doi.org/10.1177/2053951716631130.

Le Guin, Ursula K. (1969/2019): The Left Hand of Darkness, New York, NY: Ace Books.

Nietzsche, Friedrich Wilhelm (1887): »Die fröhliche Wissenschaft; Wir Furchtlosen: Neue Ausgabe«, in: Ders. (2013), Friedrich Nietzsche. Philosophische Werke in Sechs Bänden, Band 5. Hg. von Claus-Artur Scheier, Hamburg: Meiner.

Offert, Fabian (2021): »Latent Deep Space: Generative Adversarial Networks (GANs) in the Sciences«, in: Media+Environment 3 (2), S. 1–20. https://doi.org/10.1525/001c.29905.

OpenAI (2025): »Sharing the Latest Model Spec«, in: openai.com (12.02.2025). Online unter: https://openai.com/index/sharing-the-latest-model-spec/ (letzter Zugriff: 07.06.2025).

Ouyang, Long/Wu, Jeff/Jiang, Xu/Almeida, Diogo/Wainwright, Carroll L./Mishkin, Pamela/Zhang, Chong/Agarwal, Sandhini/Slama, Katarina/Ray, Alex/Schulman, John/Hilton, Jacob/Kelton, Fraser/Miller, Luke/Simens, Maddie/Askell, Amanda/Welinder, Peter/Christiano, Paul/Leike, Jan/Lowe, Ryan (2022): »Training Language Models to Follow Instructions with Human Feedback«, in: arxiv.org (04.03.2021). Online unter: https://arxiv.org/abs/2203.02155 (letzter Zugriff: 07.06.2025). https://doi.org/10.48550/arXiv.2203.02155.

Ruhr Universität Bochum, Zentrum für Wissenschaftsdidaktik (2025): »Künstliche Intelligenz in Studium & Lehre«, in: zfw.rub.de (2025). Online unter: https://zfw.rub.de/lehrende/lehre-gestalten/kuenstliche-intelligenz-in-studium-und-lehre/ (letzter Zugriff: 07.06.2025).

Schwerzmann, Katia (2025): »From Enclosure to Foreclosure and Beyond: Opening AI's Totalizing Logic«, in: AI & Society (im Erscheinen). Preprint in: philpapers.org. Online unter: https://philpapers.org/rec/SCHFET-6 (letzter Zugriff: 07.06.2025).

Schwerzmann, Katia/Campolo, Alexander (2025): »›Desired Behaviors‹: Alignment and the Emergence of a Machine Learning Ethics«, in: AI & Society, o.S. https://doi.org/10.1007/s00146-025-02272-3.

Shields, Rob (2005): The Virtual, Hoboken: Taylor and Francis.

Vaswani, Ashish/Shazeer, Noam/Parmar, Niki/Uszkoreit, Jakob/Jones, Llion/Gomez, Aidan N./Kaiser, Lukasz/Polosukhin, Illia (2017): »Attention Is All You Need«, in: arxiv.org (12.06.2017). Online unter: http://arxiv.org/abs/1706.03762 (letzter Zugriff: 07.06.2025).