KI-Grundversorgung

für Hochschulen als Verbundaufgabe

Ein Beitrag aus NRW zur Diskussion über monozentrale vs. verteilte KI-Infrastruktur für Hochschulen

Das Papier greift den laufenden Diskurs um eine KI-Grundversorgung für Hochschulen auf und fokussiert die Frage nach digital souveräner Infrastruktur für Inferenz und Training. Ausgehend von vorliegenden Positionspapieren werden Anforderungen an eine KI-Grundversorgung für Hochschulen in Deutschland beschrieben. Im Mittelpunkt steht dann die Diskussion, ob eine solche Grundversorgung durch einen einzigen Anbieter (= monozentrale Infrastruktur) oder ein Netzwerk miteinander kooperierender Provider (= verteilte Infrastruktur) erbracht werden sollte. Als Ergebnis dieser Diskussion empfiehlt das Papier eine länderübergreifende Verbundlösung, da diese den heterogenen Anforderungen besser gerecht werden kann und zudem u.a. mit Blick auf die Ausfallsicherheit in diesem kritischen Infrastrukturbereich robuster ist. Abschließend werden nächste Schritte auf dem Weg zur Umsetzung einer verteilten Infrastruktur vorgeschlagen.

von:

Martin Bovermann, Patrick Holz, Dr. Martin Jungwirth, Jonas Leschke, Prof. Dr. Matthias S. Müller, Prof. Dr. Malte Persike, Dr. Peter Salden, Dr. Christian Terboven, Tim Trappen, Prof. Dr. Stefan Wesner

DIGITALE HOCHSCHULE NRW

KI: EDU. NRW

Ki:connect.nrw

OSKI.nrw

1. KI und digitale Souveränität an Hochschulen

Applikationen auf Basis generativer Künstlicher Intelligenz (KI) führen zu weitreichenden Veränderungen in der Art und Weise, wie an Hochschulen in Forschung, Lehre und Verwaltung gearbeitet wird. Ging es bei der Diskussion um KI seit 2022 zunächst vor allem um kurzfristige Konsequenzen wie beispielsweise die Veränderungen im Prüfungswesen, ist inzwischen die strategische Frage des Zugangs zu KI in den Blickpunkt gerückt. In der Lehre ist dies u.a. deswegen wichtig, um mit Studierenden den reflektierten Umgang mit KI einüben und prüfen zu können. In der Wissenschaft ist die Verfügbarkeit leistungsfähiger Infrastruktur für Inferenz und Training, aber auch die sichere und verlässliche Integration von Datenbeständen in den Trainings- und Inferenzprozess zunehmend ein Wettbewerbsfaktor – und das nicht nur in Forschung und Lehre. Auch in der Verwaltung ist die Nutzung von generativer KI kein Tabu mehr, sondern wird zunehmend als Chance zur Optimierung von Betriebsabläufen betrachtet, weshalb hochschulinterne KI-Systeme von Verwaltungsangehörigen besonders intensiv genutzt werden.

KI durchdringt somit in allen Bundesländern sämtliche Ebenen der Hochschulen. Noch immer geschieht dies häufig unter Rückgriff auf kommerzielle Dienstanbieter, was unterschiedliche kritische Implikationen hat, insbesondere: 1) Es kann zu einem Abfluss sensibler Daten kommen, 2) die Entstehung proprietärer KI-Modelle (und damit auch ihrer Ergebnisse) ist in der Regel intransparent, 3) die Nutzung kommerzieller KI Angebote macht von der Preispolitik der Anbieter abhängig und verursacht u.U. hohe Kosten, sowie 4) wenn KI-Dienste ausschließlich von außen bezogen werden, findet an den Hochschulen nur geringer eigener Expertiseaufbau statt, was wiederum die Abhängigkeit verschärft. All diese Aspekte gefährden die digitale Autonomie von Hochschulen und begründen den Handlungsbedarf, flankierend zu kommerziellen KI-Diensten auch von den Hochschulen selbst kontrollierte KI-Dienstleistungen anzubieten und so ein gewisses Maß an digitaler Autonomie zu wahren.

Digitale Autonomie beim Thema KI lässt sich als Konstrukt mit den drei Handlungsfeldern 1) Kompetenz, 2) Inferenz und Training sowie 3) Applikationen denken. *Kompetenz* meint die grundlegende Fähigkeit aller Beteiligten, KI-Tools richtig einzuschätzen und in Bezug auf ihren Einsatz reflektierte Entscheidungen treffen zu können. In dieser Dimension von digitaler Autonomie geht es also um Bildung und Weiterbildung. Hingegen sind die beiden anderen Dimensionen eher technischer Natur: *Inferenz und Training* beziehen sich auf die Bereitstellung von KI-Modellen für Interaktion mit den Modellen sowie das Training von KI-Modellen, bei *Applikationen* geht es um die Programmierung KI-basierter oder KI-bezogener Applikationen – insgesamt also um technische Infrastruktur (sog. AI Infrastructure as a Service) und konkrete KI-Anwendungen (sog. AI Software as a Service), um bestimmte Aufgaben ausführen zu können.

In der Diskussion um KI hat sich im Hochschulsektor schon früh die Erkenntnis durchgesetzt, dass die Herausforderungen in allen drei Handlungsfeldern von Hochschulen nicht jeweils individuell zu bewältigen sind. Da alle Hochschulen zeitgleich vor identischen Problemen stehen, ist es stattdessen sinnvoll, auf allen Ebenen Kooperationen einzugehen. Die Tendenz zu Verbundlösungen lässt sich auf regionaler, Landes- und Bundesebene bis hin in den europäischen Raum beobachten, so z.B. bei den landesübergreifend konzipierten Großrechenzentren der so genannten AI Factories (https://www.eurohpc-ju.europa.eu/ai-factories). Gerade mit Blick auf technische Infrastruktur setzen sich hier Ansätze fort, die im Bereich Forschungsdatenmanagement (NFDI) und im Bereich des Hoch- und Höchstleistungsrechnen (NHR, GCS) seit Jahren fest und erfolgreich etabliert sind.

Das vorliegende Papier beschäftigt sich mit der bundesweiten Kooperation im vorrangig technischen Bereich der Bereitstellung von digital souveräner Rechenkapazität für *die Inferenz und das*

Training von KI-Modellen. Es gibt einen kurzen Überblick über den Stand der Diskussion, formuliert die zwei aktuell am stärksten diskutierten Optionen für hochschulübergreifende Zusammenarbeit in Deutschland und positioniert sich dazu. Zweck des Papiers ist es, in den Diskussionen um eine KI-Grundversorgung eine wichtige Grundsatzentscheidung aus fachlicher Perspektive zu unterstützen und damit die baldige Umsetzung einer vertieften Kooperation in diesem Bereich herbeizuführen.

2. Strategievorschläge

Im Themenfeld digital souveräne Infrastruktur für KI-Inferenz an deutschen Hochschulen haben zwei Strategievorschläge zuletzt besonders viel Beachtung gefunden: (1) das Strategiepapier „Kompetenz, Inferenz, Apps“ (Autorinnen: NRW-Landesprojekte KI:edu.nrw, KI:connect.nrw und Open Source-KI.nrw) sowie (2) das Diskussionspapier für eine KI-Grundversorgung in Deutschland (Autorin: Gesellschaft für Wissenschaftliche Datenverarbeitung Göttingen, kurz GWDG).

1. **NRW-Papier**: Das im Juli 2025 veröffentlichte Strategiepapier aus Nordrhein-Westfalen skizziert zunächst, wie die 36 Hochschulen des Landes auf allen drei Ebenen digitaler Souveränität (Kompetenz, Inferenz/Training und Applikationen) zusammenarbeiten können. Das Papier gibt ein klares Bekenntnis für die hochschulübergreifende Kooperation ab und empfiehlt insbesondere für das Land Nordrhein-Westfalen das landeszentrale Hosting einer gemeinsam genutzten KI-Basisinfrastruktur. Diese soll den Betrieb von Open Source-Sprachmodellen ermöglichen, auf die dann auch konkrete KI-Dienste der Hochschulen aufgebaut werden können. Das Papier spricht sich zudem dafür aus, in allen diskutierten Bereichen über Bundeslandgrenzen hinweg deutschlandweite Zusammenarbeit einzugehen, um sowohl den Betrieb bestehender als auch das Training und Finetuning neuer Modelle zu ermöglichen. So wird ein NRW-zentraler Inferenzknoten als möglicher Baustein einer föderierten deutschlandweiten Infrastruktur mit mehreren Providern betrachtet.
2. **GWDG-Papier**: Das im September 2025 veröffentlichte Diskussionspapier der GWDG nimmt diesen Faden auf und akzentuiert ebenfalls die Notwendigkeit der Kooperation auf allen drei genannten Ebenen. Den Schwerpunkt des Papiers bildet die Diskussion um die KI-Grundversorgung, wobei es dabei aus GWDG-Perspektive vor allem um die bundesweite Sicherung von Inferenz für Hochschulen geht. Die GWDG schlägt eine dauerhafte KI-Grundversorgung auf Basis des bisher entwickelten eigenen Dienstleistungsportfolios rund um die Academic Cloud und die Nachnutzung via API vor. Hierbei sollen Open Source-Sprachmodelle den Kern eines grundfinanzierten KI-Ökosystems bilden. Durch Kooperationen mit weiteren Akteuren soll das Angebotsportfolio z.B. um weitere Applikationen erweitert werden.

Als unstrittig kann im Lichte beider Papiere (sowie auch weiterer Stellungnahmen, z.B. Stellungnahme AIDARE-Netzwerk, Expertenanhörungen des Hochschulforum Digitalisierung in 2024 und 2025) gelten, dass Hochschulen im KI-Zeitalter Zugang zu KI-Rechenleistung benötigen und diese zumindest teilweise auch selbst kontrollieren sollten. Ebenfalls weitgehend unstrittig erscheint, dass nicht jede Hochschule individuell KI-Rechenleistung für eine größere Nutzendenzahl anbieten kann oder sollte, sondern dass eine professionell betriebene, hochschulübergreifende KI-Grundversorgung aus unterschiedlichen Gründen sinnvoll ist, wobei als solche Gründe hervorzuheben sind:

- *Eignung der Infrastruktur*: Herkömmliche IT-Systeme an Hochschulen sind regelhaft nicht für die rechenintensiven und spezialisierten Anforderungen von KI-Anwendungen ausgelegt.
- *Kosten*: Die Anschaffung und der Betrieb der benötigten Systeme (z.B. spezialisierte Grafikkarten, GPUs) sind für einzelne Institute, Fachbereiche oder Hochschulen in der Regel zu ressourcenintensiv.
- *Datenschutz und Datensicherheit*: Eine spezialisierte Infrastruktur kann besser sicherstellen, dass die Daten in einer DSGVO-konformen Umgebung verbleiben.

Abgesehen von diesen Gemeinsamkeiten sind die beiden Papiere teils verkürzt als konkurrierende Ansätze gelesen worden. So ist das NRW-Papier in dem Sinne gedeutet worden, dass es sich für bis zu 16 bundeslandspezifische Lösungen starkmache. Für das GWDG-Papier wiederum ist vermutet worden, dass die GWDG damit einen Alleinanspruch auf die bundesweite Versorgung mit KI-Inferenz erhebe. Beide Deutungen lassen sich aus den Papieren und auch aus Gesprächen mit den Akteuren nicht ableiten. Allerdings erscheint es produktiv im Sinne des Diskurses, die Impulse beider Papiere aufzunehmen und die Reflexion darüber zu vertiefen, wie hochschulübergreifende Lösungen zukünftig aussehen können – und zwar insbesondere mit Blick auf die Frage, ob Infrastruktur für Inferenz und Training eher *monozentral* (= von einem einzigen Anbieter) oder eher *verteilt* (= in einem Verbund mehrerer Anbieter) gestaltet werden sollte.

Um diese Frage diskutieren zu können, wird zunächst skizziert, welche Merkmale ein hochschulübergreifendes Angebot überhaupt haben sollte. Danach werden monozentrale und verteilte Ansätze verglichen und abschließend diskutiert.

3. Merkmale eines hochschulübergreifenden Angebots

Eine KI-Grundversorgung auf Bundesland- oder sogar Bundesebene macht in jedem Fall eine zentrale Dienstleistungsplattform erforderlich, die einer größeren Zahl von Hochschulen einen einfachen und skalierbaren Zugang zu Rechenleistung für die Anwendung und das Training von KI-Modellen bietet. Eine solche Plattform würde verschiedene KI-Modelle als „Models-as-a-Service“ bereitstellen, die direkt für Lehre, Forschung und Verwaltung genutzt werden können.

Ein solcher hochschulübergreifender Dienst, der in einem Verbund auf Bundesland- oder Bundesebene angeboten wird, sollte das folgende Spektrum von technischen Anforderungen erfüllen:

- **Modellvielfalt**: Zentrales Merkmal ist die Bereitstellung eines breiten Spektrums unterschiedlicher KI-Modelle. Dieses Spektrum reicht von unterschiedlichen Sprachmodellen (LLMs) bis hin zu solchen, die für spezifische wissenschaftliche Domänen trainiert wurden. Die Fähigkeit, verschiedene Modellarchitekturen und Software-Frameworks wie TensorFlow und PyTorch effizient zu betreiben, ist eine Grundvoraussetzung des Angebots. Dies erfordert eine flexible Hardware-Ausstattung, die neben Allzweck-GPUs auch spezialisierte Beschleuniger umfassen sollte.
- **Geringe Latenz und Echtzeitanforderung:** Benötigt werden interaktive Dienste, bei denen eine unmittelbare Reaktion für die Nutzererfahrung entscheidend ist. Beispiele hierfür sind Chatbots, KI-gestützte Programmierassistenten oder Echtzeit-Textbearbeitungs- und Übersetzungsdienste. Bei diesen Anwendungen ist die Minimierung der Verzögerung (Latenz) für eine hohe Akzeptanz durch die Nutzenden das primäre Leistungsziel. Dies be-

trifft z.B. Anwendungen aus dem Bereich Lehre (als Teil von LMS-Systemen), administrative Unterstützungsprozesse (z.B. als Teil eines Dokumentenmanagementsystems) oder Anwendungen zur Auswertung von Datenverkehr wie Gebäudeleittechnik-Informationen.

- **Hoher Durchsatz (Batch-Verarbeitung):** Für die wissenschaftliche Forschung, aber auch für eine Reihe weiterer Anwendungen (z.B. Anomalie-Erkennung im Datenverkehr) müssen oft große Datenmengen verarbeitet werden. Dies kann die Klassifizierung von (medizinischen) Bildern, die Analyse von Simulationsdaten oder die Verarbeitung umfangreicher Texte umfassen. Hier ist der Gesamtdurchsatz entscheidend, also die Fähigkeit, eine große Anzahl von Inferenzen in einem bestimmten Zeitraum abzuschließen. In diesem Anwendungsbereich ist es sinnvoll, Systeme für KI-Inferenz eng an HPC-Systeme und ebenfalls an die in Deutschland verteilten Datenbestände anzubinden.
- **Loadbalancing:** Die Auslastung der aufzubauenden Hardware durch Inferenz, Training und angrenzende Computing-Aufgaben ist regional zu koordinieren, um die Bedarfe der Hochschulen in Deutschland abzubilden und eine ausgeglichene Nutzung der Infrastruktur zu ermöglichen.
- **Sicherheit**: Bei nahezu allen Einsatzzwecken von KI im Hochschulsektor muss von der Verarbeitung sensibler Daten ausgegangen werden. Informationssicherheit ist daher von höchster Bedeutung. Um diese gewährleisten zu können, bedarf es spezialisierter Systeme und geschulten Personals. Hinzu kommen sehr hohe Anforderungen an die Ausfallsicherheit. Gerade wenn ein Dienst mehrere Hochschulen versorgt und diese Hochschulen sich infrastrukturell auf den Dienstleister verlassen, muss das Angebot mit höchstmöglicher Verlässlichkeit zur Verfügung stehen.

Ein derartiges Angebot kann entweder als monolithisches System mit einem einzigen Einstiegspunkt oder aber als lose gekoppeltes System mit unterschiedlichen Einstiegspunkten gedacht werden. Dies leitet zur Frage nach der Zahl der Provider über, die im nächsten Kapitel diskutiert wird.

4. Monozentrale vs. verteilte Lösung für Inferenz und Training: Überblick

Diskutabel erscheint die Frage, ob die zentrale Bereitstellung eines solchen Angebots von *einem einzigen* bundesweiten Provider („monozentrale Infrastruktur“) oder über eine föderierte Infrastruktur *mehrerer* über das Bundesgebiet verteilter und miteinander vernetzter Provider („verteilte Infrastruktur“) geleistet werden sollte. Diese Diskussion ist bisher kaum ausdrücklich geführt worden.

Wir stellen im Folgenden beide Ansätze tabellarisch gegenüber und skizzieren die jeweiligen Merkmale der Ansätze mit Blick auf die wichtigsten Kriterien der Bereitstellung (wobei eine grüne Färbung jeweils die tendenziell anzustrebende Variante markiert).

Kriterium	Monozentrale Infrastruktur	Verteilte Infrastruktur
(1) Verantwortung	eindeutig zugewiesen; klare Leitungs- und Entscheidungsstruktur	geteilt; kollektive Entscheidungsverantwortung
(2) Entscheidungsgeschwindigkeit	kurze Entscheidungswege, schnelle Umsetzung	umfangreichere Abstimmungsprozesse, komplexe Umsetzung

Kriterium	Monozentrale Infrastruktur	Verteilte Infrastruktur
(3) Koordinationsaufwand	einfache Kommunikation und begrenzte Abstimmungserfordernisse	regelmäßige Sitzungen, komplexe Abstimmungen
(4) Stakeholder-Einbindung	wenige Kontaktpunkte, top-down gesteuert; Distanz zu lokalen Stakeholdern	viele Kontaktpunkte, lateral gesteuert; Nähe zu lokalen Stakeholdern
(5) Resilienz bei Ausfall	gering und abhängig von einer einzigen Institution	hoch durch verteilte Infrastruktur sowie dezentrales Wissen & Verantwortungsmanagement
(6) Angebotsbreite/ Flexibilität	begrenzt auf den Fokus bzw. die Erweiterungsmöglichkeiten des monopolisierten Standorts und von diesem abhängig	hoch durch Möglichkeit zur Profilbildung, Ausdifferenzierung und Skalierung an allen beteiligten Standorten
(7) Komplexitätsbewältigung	begrenzt durch das Angebots- und Expertiseprofil an nur einem Standort	breit aufgestellt durch das Pooling von Kompetenzen mehrerer Standorte und Berücksichtigung spezieller regionaler Erfordernisse
(8) Know-How-Aufbau und -Erhalt, Personalgewinnung.	lokaler Know-How-Aufbau mit wenigen Beteiligten, Gefahr des Brain-Drain in anderen Regionen, Problem der Personalgewinnung am gewählten Standort	regionaler Kompetenzaufbau und -erhalt an mehreren Standorten, verteiltes Risiko bei der Personalgewinnung
(9) Kosten	Kostenpositionen fallen nur an einem Standort an; durch Solitär evtl. schwache Rabattierungsposition bei Beschaffungen	teils redundante Kostenpositionen unter den Verbundpartnern (z.B. Wachdienst); bei Beschaffungen evtl. gute Rabattierungsposition durch Verbundauftritt
(10) Daten-/Informationssicherheit	hoher Standard möglich; Sicherheitsvorfall führt im worst case zum Ausfall der gesamten Infrastruktur	hoher Standard möglich, muss an jedem Standort sowie für den Verbund konzipiert werden; gegenseitige Absicherung möglich

5. Diskussion

Gemäß dieser Betrachtung haben sowohl monozentrale als auch verteilte Infrastrukturen jeweils eine Reihe von Stärken und Schwächen. Eine monozentrale Infrastruktur spielt ihre Stärken bei Aspekten wie klarer Verantwortung, schlanken Prozessen sowie dem schnellen initialen Aufbau von Know-How und Lösungen aus. Eine verteilte Infrastruktur hat dagegen Stärken durch ein hohes Maß an Resilienz (z.B. Ausfallsicherheit), fachliche Kompetenz in der Breite und die Möglichkeit, technische Ansätze auch mit Blick auf heterogene fachliche und/oder regionale Anforderungen auszurichten.

Diskutiert man die Vor- und Nachteile vertieft vor dem Hintergrund der Idee einer KI-Grundversorgung für Hochschulen in Deutschland, können einige der in der Tabelle separat dargestellten Aspekte gruppiert werden, was zu folgenden besonders wichtigen Argumenten führt:

- **Schnelle Handlungsfähigkeit**: In einem sich schnell entwickelnden Feld wie KI ist Geschwindigkeit als wichtiger Faktor zu verstehen – dies erscheint als das stärkste Argument für eine monozentrale Infrastruktur. Denn in ihr scheint es besonders einfach zu sein, kurzfristig Entscheidungen zu treffen und ihre Umsetzung einzuleiten (Punkte 1-3).
- **Bewältigung vielfältiger Anforderungen**: Die im vorigen Punkt genannte schnelle Handlungsfähigkeit im monozentralen Szenario kann allerdings an ihre Grenzen kommen, wenn Komplexität zu bewältigen ist. Denn im Falle eines einzigen Dienstleisters sind zwangsläufig dessen Kapazitäten zur Reaktion auf heterogene Anforderungen begrenzt. Zudem entstehen typische Monopolisierungsrisiken, die Fragen von Stakeholder-Einbindung, Angebotsbreite, Innovationsdynamik und Komplexitätsbewältigung noch stärker in den Fokus stellen (Punkte 4, 6, 7). Ein bundesweit gedachtes System der KI-Grundversorgung muss die Bedürfnisse von mehreren hundert Hochschulen decken, die sich nicht nur als Hochschultypen grundlegend unterscheiden (von der Technischen Universität bis zur Kunst- und Musikhochschule), sondern auch jeweils fachlich binnendifferenziert sind wie auch z.B. in Hinblick auf IT-Spezifika (z.B. durch die gemeinsame Nutzung regionaler Dienste) extrem diverse Anforderungen haben. Dass eine solche Vielfalt der Anforderungen von einem einzigen deutschlandweiten Provider abgedeckt werden könnte, erscheint unrealistisch, wohingegen ein verteiltes System u.a. dank der Möglichkeit zur lokalen Profilbildung und Spezialisierung hier leistungsfähiger erscheint.
- **Sicherheit und Souveränität**: Bedenkt man die hohe Relevanz, die KI-Anwendungen schon jetzt für den Wissenschaftsbereich zugeschrieben wird, verdienen auch sicherheitsbezogene Argumente (Punkte 5, 8, 10) besondere Beachtung. Bauen Hochschulen verstärkt Dienste auf KI auf, ist es riskant, sich bundesweit auf nur einen einzigen Anbieter zu stützen. Die KI-Grundversorgung ist im Gegenteil als Teil der kritischen und souveränen Infrastruktur von Hochschulen zu verstehen, die nicht von einem einzelnen Provider abhängen sollte – denn würde ein solcher ausfallen oder in seiner Leistungsfähigkeit eingeschränkt werden, wäre die Versorgung des gesamten Hochschulraums beeinträchtigt oder sogar lahmgelegt.

Bewertet man diese drei aggregierten Punkte in der Zusammenschau, so scheint es, dass eine verteilte Lösung insgesamt den Anforderungen einer bedarfsgerechten und stabilen Grundversorgung besser gerecht werden kann. Dem Aspekt der schnellen Handlungsfähigkeit muss für die Konzeption einer verteilten Lösung allerdings besondere Aufmerksamkeit gelten, d.h. dies ist als Anforderung an die Governance eines verteilten Ansatzes zu verstehen (und zugleich als dessen Risiko).

6. Fazit und Ausblick

Zusammenfassend und in der Schlussfolgerung sollten nach unserer Einschätzung folgende Aspekte für kommende Diskussionen über KI-Infrastruktur an deutschen Hochschulen als Grundkonsens dienen:

- **Zugang zu KI-Infrastruktur**: Hochschulen benötigen Zugang zu Infrastruktur für KI-Inferenz und -Training. Vor allem die KI-Angebote in der Inferenz müssen eine Vielzahl von interaktiven Diensten für Forschung, Lehre und Verwaltung umfassen und die hochdynamische Entwicklung im Bereich KI berücksichtigen und unterstützen.
- **Betrieb souveräner Infrastruktur**: Hochschulen sollten sich für solche Zugänge nicht allein auf kommerzielle Dienstleister stützen, sondern auch Zugriff auf digital souveräne Infrastruktur haben und diese entweder selbst oder in Kooperation mit öffentlichen Partnereinrichtungen betreiben.
- **Verteilte Infrastruktur**: Hochschulen sollten nicht jede für sich KI-Infrastruktur aufbauen, sondern gemeinsame, vernetzte Angebote schaffen bzw. nutzen. Eine solche gemeinsame KI-Infrastruktur sollte insbesondere auf Bundesebene nicht von einem monozentralen Anbieter, sondern als sicheres, robustes, verteiltes System in einem bundesweiten Verbund betrieben werden.

Für einen solchen Verbund schlagen wir den Zusammenschluss einer begrenzten Zahl geeigneter Rechenzentren vor. Auf die Dienste des Verbunds sollten Hochschulen aller Bundesländer Zugriff haben, wenngleich nicht alle Bundesländer ein eigenes Rechenzentrum für den Verbund benennen müssten. Stattdessen sollte im Sinne nächster Schritte diskutiert werden:

- Welche Bundesländer möchten Rechenleistung in eine deutschlandweite KI-Grundversorgung für Hochschulen einbringen?
- Wie soll ein Verbund für KI-Grundversorgung organisiert werden? Wie verhält sich das vernetzte Angebot einer KI-Grundversorgung zu bestehenden Einrichtungen und Aktivitäten (z.B. im Kontext des Nationalen Hochleistungsrechnens)?
- Wie wird ein Angebot der KI-Grundversorgung finanziert?

Hinzu kommen die Fragen, inwieweit aufbauend auf einem Angebot für Inferenz und Training auch bei der Programmierung KI-basierter Applikationen sowie beim Aufbau von KI-Kompetenz bundeslandübergreifend zusammengearbeitet werden kann. Bei diesen Fragen sind allerdings teils auch andere Akteurinnen und Akteure als Rechenzentren einzubeziehen, wie z.B. das Netzwerk der Landeseinrichtungen für digitale Hochschullehre und auf Kompetenzerwerb gerichtete Vorhaben wie KI:edu.nrw.

Gelingt die Lösung dieser Fragen, so kann die Bereitstellung einer leistungsstarken, skalierbaren und zugänglichen Infrastruktur für KI-Inferenz und -Training zu einem wichtigen Grundstein der Wettbewerbsfähigkeit von Forschung, Lehre und Verwaltung an deutschen Hochschulen im KI-Zeitalter werden.

ANHANG: Monozentrale vs. verteilte Infrastruktur für Inferenz und Training: vertiefte Überlegungen

Die aus unserer Perspektive wichtigsten Argumente in der Diskussion um monozentrale oder verteilte Infrastruktur sind im Haupttext des vorliegenden Dokuments ausgeführt. Die Kriterien der dort dargestellten Tabelle und ihre Bewertung werden im Folgenden detaillierter betrachtet. Wir sind uns dessen bewusst, dass es zu jedem Aspekt der Tabelle eine Vielzahl von Argumenten gibt und auch die unten angestellten Überlegungen hinterfragt werden können. Die Überlegungen sind daher als Explikation von Argumenten zu verstehen, zu deren Diskussion herzlich eingeladen sei.

Grundsätzlich leistet in einer monozentralen Infrastruktur ein einzelnes Rechenzentrum das gesamte Angebot, vom Hosting und Betrieb der Systeme bis zur Bereitstellung von Diensten. In einer verteilten Infrastruktur leisten mehrere Rechenzentren das Hosting und den Betrieb der Systeme, ggf. auch partnerschaftlich. Ein Grundangebot von KI-Diensten wird dabei von allen Rechenzentren erbracht, eine darauf aufbauende Spezialisierung und die Verteilung von Kunden und Daten ist möglich. Die Dienste sind über einem zentralen oder mehrere gleichwertige Zugangspunkte erreichbar. Auf Basis dieser Architekturen einer monozentralen und verteilten Infrastruktur lässt sich eine Bewertung der im Hauptdokument formulierten Kriterien (1) bis (10) ableiten.

(1) Verantwortung & (2) Entscheidungsgeschwindigkeit: Auf den ersten Blick scheinen beide Kriterien in einem monozentralen Ansatz eindeutig effizienter. Dies gilt allerdings insbesondere dann, wenn primär die strategisch-administrative Ebene in Betracht gezogen wird.

Auf operativer Ebene fordert die hohe Entwicklungsdynamik im KI-Bereich eine Vielzahl von kurzfristig zu treffenden Entscheidungen, deren Ausgestaltung sich maßgeblich aus den vorhandenen Ressourcen ergibt (Hardware, Personal). Zudem ist durch die wachsende Nutzung von KI auch mit einer zunehmenden Breite an Anforderungen zu rechnen, welche eine fachliche Spezialisierung nicht nur auf Anwendungsebene, sondern auch auf Ebene der Infrastruktur notwendig macht. Abhängig vom Grad der regionalen Autonomie können solche Entscheidungen in einem verteilten Ansatz besser und schneller entwickelt, bewertet, erprobt und in die Breite vermittelt werden.

Dafür sind geeignete Governance-Strukturen zu entwickeln. Verantwortung muss an den jeweiligen Standort übertragen und mit klarer Perspektive auf den Verbund eingelöst werden. Es müssen effiziente Kommunikationswege und Prozesse eingeführt sowie Entscheidungsprozesse durch (i) eine fachliche Spezialisierung und (ii) das Clustering von Kunden mit ähnlichen Anforderungen an Infrastruktur und Dienste zusammengeführt werden.

(3) Koordinationsaufwand & (4) Stakeholder-Einbindung: Ähnlich der Entscheidungsgeschwindigkeit spricht mindestens das erste Kriterium zunächst für eine monozentrale Infrastruktur. Der Koordinationsaufwand einer verteilten Infrastruktur ist sowohl auf administrativer wie operativer Ebene durch die aufzuwendende Abstimmung und gemeinsame Umsetzung deutlich höher.

Dem höheren Koordinationsaufwand steht eine engere Stakeholder-Einbindung (im Sinne von Kollaboration) gegenüber. Regionale Bedarfe und Standortspezifika können durch kürzere Kontaktwege und die Möglichkeit der stärkeren Binnendifferenzierung besser adressiert werden. Relevant wird dies insbesondere im Bereich von Problemlösung und Prävention, wo Bedarfe von Nutzenden vielschichtig eingebunden werden können.

(5) Resilienz bei Ausfall & Transparenz: Der Ausfall eines zentralen Anbieters stellt unwiderruflich einen Ausfall für sämtliche Stakeholder dar. Dabei kann ein Ausfall technisch, aber auch personell bedingt sein und unterschiedlich starke Auswirkungen nach sich ziehen.

Auch eine verteilte Infrastruktur kann durch einen systemweiten Ausfall, beispielsweise eines zentralen Zugangsknotens, betroffen sein. Allerdings lässt sich die Resilienz durch regionale bzw.

separate Zugänge zu einzelnen Standorten erhöhen. Durch breiter verteilte Kompetenzen auf personeller Ebene sind Ausfälle besser kompensierbar und haben wenig Einfluss auf die Funktionsfähigkeit im Verbund. Ferner ermöglicht eine verteilte Infrastruktur nach Absprache unter den Partnern auch Wartungsarbeiten ohne Unterbrechung des Dienstangebotes, da bei geplanten temporären Unterbrechungen des Betriebs die Zugangspunkte flexibel delegiert werden können.

Das Maß an Transparenz im Szenario eines Ausfalls, wie auch der allgemeinen administrativen und operativen Ebene, ist durch die Notwendigkeit sauberer Protokollierung und Kommunikation in einem verteilten Ansatz höher zu bewerten, bedingt durch die gegenseitige Supervision der Akteure.

(6) Angebotsbreite/Flexibilität & (7) Komplexitätsbewältigung: Hier muss zwischen dem generalistisch anzulegenden Aufbau einer zentralen Zugriffsstelle und den spezifischen Angeboten verteilter Standorte unterschieden werden. Während Angebote in einer verteilten Infrastruktur als wahrscheinlich breiter und flexibler angelegt sind und zeitgleich die Komplexität der Umsetzung durch das dezentral ansässige Wissen überschaubar bleiben dürfte, bedarf die Entwicklung der zentralen Zugriffsstelle eines koordinierten Einverständnisses mit Blick auf standardisierte Prozesse, Technologien und Zuständigkeiten. Zudem besteht bei mangelnder Abstimmung die Gefahr, dass multiple Angebote für dieselben Anforderungen entstehen. Dieser Aufwand bzw. diese Gefahr entfällt bei einem monozentralen Ansatz größtenteils, was nicht nur die operative Belastung senkt, sondern auch mögliche Reibungsflächen mit Anforderungen und Systemen der Nutzenden minimiert.

Entgegen der als agil zu begreifenden standortinternen Arbeit ist der gemeinsame Teil der verteilten Infrastruktur - die Abstimmung von Umsetzungsvorhaben - daher als größte Herausforderung zu benennen. Risiken sind dabei personeller wie technischer, nicht zuletzt aber administrativer Natur. Gelingt die Schaffung einer zentralen Zugriffsstelle auf verteilte Angebote legt dies den Grundstein für eine Zusammenarbeit über Länder- wie Bundesebene hinaus und kann bis auf EU-Ebene ausgeweitet werden. Wird ein solches Szenario von einem zentralisierten Ansatz aus angestrebt, muss eine solche Struktur direkt von der höchsten Ebene (bspw. europäisch) aus „top-down“ konzipiert, koordiniert und umgesetzt werden; abhängig von den Erfahrungswerten der Stakeholder kann es zu unvorhergesehenen Problemen in der Vernetzung und Skalierung kommen, die in einem dezentralen „bottom-up“ Ansatz gegebenenfalls bereits aufgetreten sind und bewältigt werden mussten.

Hingegen bringt ein Misserfolg des dezentralen Ansatzes die Gefahr einer Zersplitterung der Akteure mit sich, bei dem zwar Kompetenzen weiterhin standortgebunden gesichert wären, sich schlussendlich aber erneut einzelne Leuchttürme entwickeln und somit mehrfache zentrale Insellösungen entstehen würden. Dieses Szenario ist für den monozentralen Ansatz unerheblich, wenngleich die Konsequenzen beim Scheitern desselbigen, beispielsweise durch geringe Akzeptanz des Angebotes, ähnlich zu bewerten sind.

(8) Know How-Aufbau und Erhalt, Personalgewinnung & (9) Kosten für Beschaffung und Betrieb: Die Bereitstellung von KI-Modellen, insbesondere „at scale“, erfordert Personal mit einem hohen Maß an technischem Wissen. Zentralisierung könnte den notwendigen Personalschlüssel kleiner halten, wenn eine Beschränkung auf Basisdienste erfolgt, was auch die Personalgewinnung zunächst vereinfachen könnte. Die gilt jedoch nur, sofern das Einzugsgebiet des zentralen Standortes die gesuchten Fachkräfte bietet. Mehrere Standorte erschließen hingegen automatisch ein größeres Einzugsgebiet für die Personalgewinnung. Hier ließe sich flexibler auf die Gegebenheiten und aktuelle Marktlagen reagieren und prinzipiell Ressourcen zwischen Institutionen verschieben

Darüber hinaus ist eine Differenzierung der Angebotspalette mit Blick auf verschiedene Betriebs-, Architektur- wie Infrastrukturmodelle in einem monozentralisierten Ansatz nur eingeschränkt zu gewährleisten und wird u.a. durch homogene Hardwareausstattung erschwert. Durch einen

„Vendor-Lock-In“ könnten, etwa im Falle einer Quasi-Monopolbildung, nicht vorhersehbare Kosten entstehen. Im Rahmen verteilter und in bestimmter Hinsicht heterogener Systeme lassen sich Alternativen schneller erproben.

Ebenfalls bedacht werden sollte hierbei, dass das Vorhandensein von technisch hochspezialisiertem Personal an verschiedenen Hochschulen starke Synergieeffekte mit den ansässigen Lehrstühlen und somit für die Forschung haben kann, was die KI-Forschung auf der einen, die Fortbildung des Personals auf der anderen Seite nachhaltig stärkt.

(10) Informationssicherheit: Ausgehend von weiteren Fortschritten im Bereich KI ist die Informationssicherheit, zu der hier ausdrücklich die Teilaspekte von Datensicherheit und Datenschutz hervorgehoben werden sollen, von höchster Bedeutung. Hochschulen sind beliebte Ziele von Cyberangriffen geworden. Als Konsequenz solcher Angriffe können Universitätsnetzwerke und darin enthaltene Dienste, aber auch sensible Daten über Hochschulangehörige und Verwaltungsvorgänge kompromittiert werden.

Es ist davon auszugehen, dass KI-Angebote, in denen neben unspezifischen textbasierten Prompts auch allgemeine sowie sensible personenbezogene- (z.B. medizinische) und anderweitige sensible Daten (z.B. aus der Forschung) verarbeitet werden, ebenfalls Ziele von Angriffen werden. KI-Infrastruktur muss somit höchsten Sicherheitsansprüchen genügen.

Die hierfür zu erlangenden Zertifizierungen sind bei einer monozentralisierten Infrastruktur naturgemäß mit einem einmaligen Aufwand zu erlangen, während sich dieser bei einer verteilten Infrastruktur multipliziert. Zudem dürften nicht alle Zertifizierungen, die ein einzelner Standort erlangen kann, aufgrund unterschiedlicher standortspezifischer Voraussetzungen flächendeckend an allen Standorten nachweisbar sein.

Kommt es allerdings bei einem zentralisierten Ansatz zu einem erfolgreichen Angriff, sind die Konsequenzen für Betrieb und Datenschutz verheerend. Eine verteilte Infrastruktur hingegen kann ein solches Szenario besser abfedern und den Ausfall eines betroffenen Standortes durch Übernahme von Diensten kompensieren, sodass mindestens ein Basisangebot von KI-Angeboten ausfallsicher angeboten werden kann. Mit Bezug auf möglichen Datenabfluss ist die Menge der Daten außerdem auf den jeweiligen Standort begrenzt.

Zusammenfassend ergeben sich aus dieser Darstellung ergänzend zu der im vorherigen Kapitel vorgestellten Tabelle detailliertere Vor- und Nachteile einer monozentralen wie auch einer verteilten Infrastruktur für Inferenz und Training. Weiterhin gilt: Keiner der Ansätze hat ausschließlich Vorteile und keiner der Ansätze hat ausschließlich Nachteile. Es gibt Aspekte, die bei der Findung und Entscheidung zu gewichten und zu bewerten sind. Ausschlaggebend ist letztlich eine Betrachtung der maßgeblichen Anforderungen an eine KI-Grundversorgung. Hierfür sind im Folgenden einige Aspekte eines möglichen Angebotsspektrums benannt.