KI-Rechenzentren in Deutschland

Aktuelle Kapazität, zukünftiger Bedarf

21. September, 2025

Autoren: Philip Fox, Daniel Privitera, Monika Schnitzer

Kontakt: philip@kira.eu

Executive Summary

KI-Rechenkapazität ist eine Grundvoraussetzung, um in Zukunft wettbewerbsfähig und souverän zu sein. Allzweck-KI (*general purpose AI*) entwickelt sich rasant und durchdringt die Gesellschaft deutlich schneller als frühere Technologien wie der Computer oder das Internet. Führende KI-Nationen werden in den kommenden Jahren immer mehr Prozesse in Wirtschaft und Verwaltung teilweise oder vollständig auf KI umstellen. Voraussetzung dafür sind Rechenzentren mit hochmodernen KI-Chips für das Training und den Betrieb von KI-Modellen. Ohne eigene KI-Rechenkapazität verstärkt Deutschland seine Abhängigkeit von anderen Staaten und droht, wirtschaftlich zurückzufallen.

Deutschland verfügt im internationalen Vergleich über wenig KI-Rechenkapazität und baut diese weniger ambitioniert aus als andere Länder. Aktuell befinden sich weniger als 2% der globalen KI-Rechenkapazität in Deutschland. In den nächsten Jahren schaffen nicht nur KI-Schwergewichte wie die USA oder China deutlich mehr zusätzliche Kapazität als Deutschland. Auch Länder wie Frankreich, das Vereinigte Königreich, Indien, Saudi-Arabien oder die Vereinigten Arabischen Emirate bauen ihre KI-Rechenkapazität ambitionierter aus. Das gilt selbst dann, wenn in Deutschland wie geplant eine europäische KI-Gigafabrik mit 100.000 KI-Chips entsteht.

Je nach Ambitionslevel stehen der Bundesregierung verschiedene Strategien zum Ausbau von KI-Rechenkapazität zur Verfügung. Deutschlands Bedarf an KI-Rechenkapazität hängt davon ab, wie stark KI in Wirtschaft und Verwaltung integriert werden soll und ob Deutschland eigene Spitzenmodelle trainieren will. Dieser Bericht schlägt drei mögliche Strategien für den KI-Infrastrukturausbau bis Ende 2028 vor:

1. KI in Teilbereichen anwenden: benötigt 850k H100-Äquivalente bzw. 0,8 GW IT-Anschlussleistung
2. KI in allen Bereichen anwenden: benötigt 3,4m H100-Äquivalente bzw. 3,4 GW IT-Anschlussleistung
3. KI in allen Bereichen anwenden und zusätzlich eigene Spitzenmodelle trainieren: benötigt 6m H100-Äquivalente bzw. 5,9 GW IT-Anschlussleistung

Mit den richtigen Maßnahmen kann Deutschland zur KI-Nation werden. Laut Koalitionsvertrag soll Deutschland in den nächsten Jahren zu einer KI-Nation werden. Durch ein beherztes Vorgehen kann die Bundesregierung dieses Ziel erreichen. Dieser Bericht macht dazu sechs Empfehlungen:

1. Kapazität für KI-Training und -Inferenz gemäß strategischer Prioritäten ausbauen
2. Politische Voraussetzungen schaffen und Kompetenzen bündeln
3. Günstige, verlässliche Energie für KI-Rechenzentren bereitstellen
4. Planungs- und Genehmigungsprozesse beschleunigen
5. Sicherheit von KI-Rechenzentren erhöhen
6. KI-Ökosystem umfassend stärken

Autoren

Dr. Philip Fox ist Policy Specialist bei KIRA und Ko-Autor des International AI Safety Report. Er hat Philosophie und Volkswirtschaftslehre in Bayreuth, Oxford und Berlin studiert und wurde an der Humboldt-Universität zu Berlin in Philosophie promoviert.

Daniel Privitera ist Gründer und Executive Director bei KIRA. Er war Vice Chair für das Safety & Security Chapter des EU Code of Practice für Allzweck-KI und Lead Writer des International AI Safety Report.

Prof. Dr. Dr. h.c. Monika Schnitzer ist Professorin an der Ludwig-Maximilians-Universität in München und Vorsitzende des Sachverständigenrates zur Begutachtung der gesamtwirtschaftlichen Entwicklung.

Inhalt

Executive Summary..... 2

1. Status quo..... 5

2. Strategien..... 14

3. Empfehlungen..... 21

4. Fazit..... 26

Technischer Appendix..... 27

Literaturverzeichnis..... 31

Über diesen Bericht..... 36

1. Status quo

Die Fähigkeiten von Allzweck-KI (fortan "KI") entwickeln sich rasant und werden sämtliche Bereiche in Wirtschaft und Verwaltung durchdringen. Die Geschwindigkeit, mit der Künstliche Intelligenz (KI) stetig leistungsfähiger wird, überrascht selbst Fachleute. Besonders bemerkenswert ist die Entwicklung der sogenannten Allzweck-KI (*general-purpose AI*). Dazu gehören vielseitig einsetzbare, mit riesigen Datenmengen trainierte KI-Systeme wie ChatGPT (OpenAI), Gemini (Google) oder Claude (Anthropic) sowie zahlreiche darauf aufbauende Anwendungen. ***Wegen der überragenden Bedeutung von Allzweck-KI beschränkt sich dieser Report ausschließlich darauf; wo in diesem Report von "KI" die Rede ist, ist Allzweck-KI gemeint.*** Während solche KI-Systeme noch vor fünf Jahren kaum einen grammatikalisch korrekten Satz formulieren konnten, lösen sie heute in wenigen Sekunden komplexe Probleme aus den Naturwissenschaften oder der Informatik – teilweise auf einem ähnlichen Niveau wie Menschen mit Doktortitel. Gleichzeitig arbeiten die besten Forschungsteams mit immensen Budgets an der Lösung bestehender Probleme wie Halluzinationen oder der mangelnden Verlässlichkeit von KI-Agenten. Setzt sich der Trend der vergangenen Jahre fort – oder beschleunigt er sich gar, wie manche Fachleute erwarten – werden führende KI-Nationen absehbar immer mehr Prozesse in Wirtschaft und öffentlicher Verwaltung teilweise oder ganz auf KI umstellen. Wer nicht über die nötige Rechenkapazität für diesen Wandel verfügt, droht abgehängt zu werden.

KI durchdringt die Gesellschaft schneller als frühere Technologien. Wenige Jahre nach der Einführung liegt die Adoptionsrate von Allzweck-KI deutlich höher als bei Computern oder dem Internet zu einem vergleichbaren Zeitpunkt (Abbildung 1). Die schnelle Verbreitung ergibt sich unter anderem aus der einfachen individuellen Nutzung: Moderne KI ist über leicht zu bedienende, oft kostenlose Chatbots in einer vertrauten digitalen Umgebung verfügbar. Einige Fachleute gehen davon aus, dass bis 2030 praktisch alle Unternehmen kommerzielle KI nutzen werden (Abbildung 2). Aufgrund der schnellen Verbreitung und der langfristigen strategischen Bedeutung sollte Deutschland deshalb so schnell wie möglich die Voraussetzungen für einen starken KI-Standort schaffen.

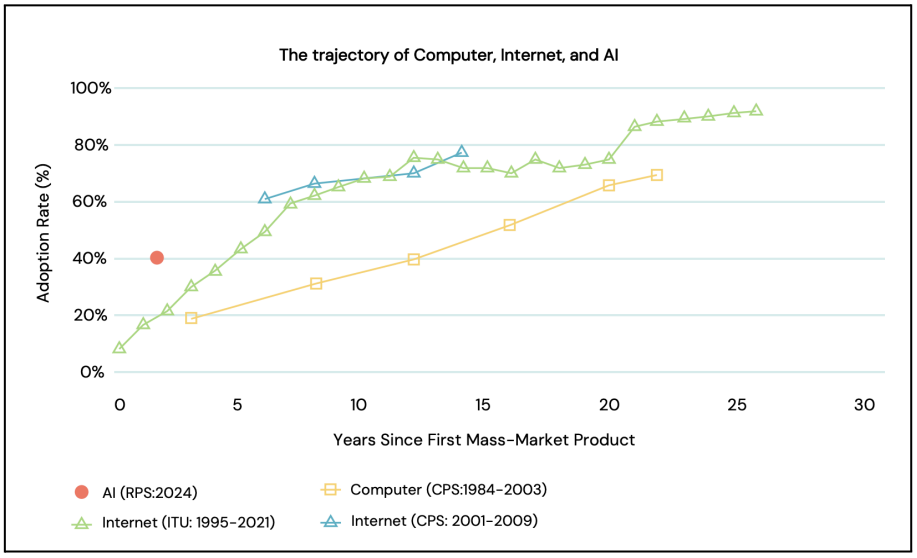

Abbildung 1: Daten zur beruflichen Nutzung weisen auf eine Adoptionsrate von generativer KI hin, die deutlich schneller wächst als bei früheren Technologien wie dem Computer oder Internet.
Quelle: Bick et al. (2024)

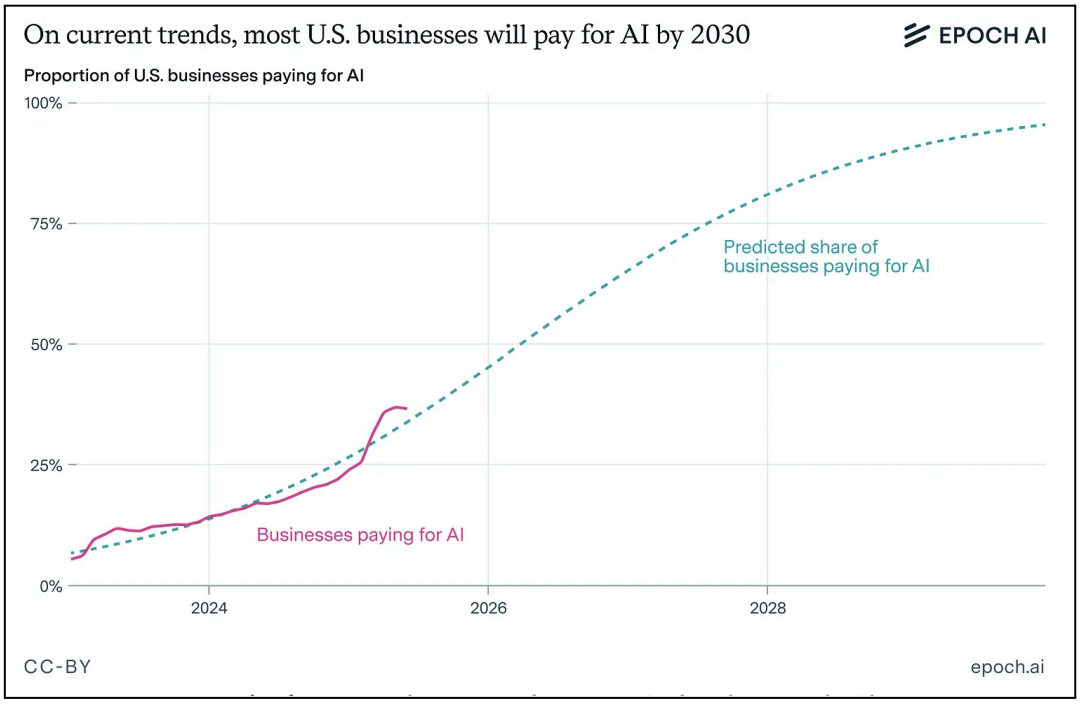

Abbildung 2: Eine aktuelle Schätzung aus den USA geht davon aus, dass bis 2030 praktisch alle Unternehmen KI kommerziell nutzen werden. Quelle: Berg & Ho (2025)

KI-Rechenkapazität ist im KI-Zeitalter eine strategische Kernressource. Eine Grundvoraussetzung, um das immense Potential von KI auszuschöpfen, sind große Mengen hochkomplexer KI-Chips wie GPUs (Graphics Processing Units). Sie sind das Herzstück spezieller KI-Rechenzentren und werden sowohl für die Entwicklung ("Training") als auch den Betrieb ("Inferenz") von KI-Modellen benötigt (siehe Abschnitt 2). Ein Unternehmen, das KI-Modelle nutzen oder trainieren will, kann dazu entweder eigene Rechenzentren betreiben oder die nötige Rechenleistung bei externen Anbietern mieten (Cloud Computing). Handelt es sich beim Cloud-Anbieter um ein ausländisches Unternehmen, ist die auf diese Weise bezogene Rechenleistung nicht souverän (siehe Box 1).

Box 1: Souveräne Rechenkapazität: Vor- und Nachteile

- Dieser Bericht versteht Rechenkapazität nur dann als souverän, wenn das jeweilige Rechenzentrum einen physischen Standort in Deutschland (bzw. der EU) hat und von einem deutschen (bzw. europäischen) Unternehmen betrieben wird (siehe Abbildung 3).
- Deutsche Unternehmen können souveräne Rechenkapazität daher auf zwei Arten beziehen: 1) Betrieb eines eigenen Rechenzentrums in Deutschland (bzw. der EU), 2) Nutzung eines deutschen (bzw. europäischen) Cloud-Dienstes, bei dem das relevante Rechenzentrum physisch in Deutschland (bzw. der EU) liegt. Rechenleistung, die von Cloud-Anbietern mit Sitz im (nicht-europäischen) Ausland bezogen wird, ist nicht souverän im Sinne dieses Berichts – auch dann nicht, wenn sich das jeweilige Rechenzentrum physisch in Deutschland befindet.

- Der Ursprung der Hardware-Komponenten ist für diese Definition von Souveränität unerheblich. Eine rein europäische KI-Chip-Produktion ist in den nächsten Jahren nicht realistisch; ein handlungsrelevanter Souveränitätsbegriff sollte diese nicht zur Voraussetzung machen.
- Souveräne KI-Rechenzentren bieten einige Vorteile:
    - **Geopolitische Unabhängigkeit.** Die Nachfrage nach KI-Rechenkapazität ist in den vergangenen Jahren enorm gestiegen. Wer diese nur über ausländische Cloud-Dienste mieten kann, schwächt die eigene Verhandlungsposition und macht sich abhängiger von ausländischen Regierungen, die den Zugang zu Cloud-Rechenkapazität kappen oder an strenge Bedingungen knüpfen könnten. Wer KI-Chips hingegen einmal importiert und in eigenen Rechenzentren installiert hat, kann über diese wesentlich unabhängiger verfügen.
    - **Datensouveränität.** Eigene KI-Rechenzentren erlauben eine vollständige Kontrolle über sensible Daten. Dazu zählen sensible Geschäftsdaten und geistiges Eigentum in Unternehmen sowie sensible personenbezogene Daten, etwa beim Einsatz von KI in der öffentlichen Verwaltung oder Medizin.
    - **Nationale Sicherheit.** In besonders sicherheitskritischen Bereichen – z. B. KI-Anwendungen in der kritischen Infrastruktur – sollten KI-Rechenzentren umfassend vor Sabotage und Spionage geschützt sein. Nur in souveränen KI-Rechenzentren können höchste Sicherheitsstandards unabhängig und gemäß nationaler Standards implementiert und jederzeit geprüft werden.
    - **Latenz.** Wird KI in komplexe Industrieprozesse integriert, ist eine niedrige Latenz wichtig – d. h. eine geringe Verzögerung zwischen einer Anfrage an ein KI-System und der berechneten Antwort. Auch in vielen anderen Bereichen kommt es auf niedrige Latenzen an (z. B. autonomes Fahren, Robotik, Hochfrequenzhandel). Dafür sorgen KI-Rechenzentren, die sich in physischer Nähe zu den Unternehmen befinden, die diese nutzen.
    - **Wertschöpfung.** Souveräne KI-Rechenzentren stimulieren die Wirtschaft und verhindern, dass ein beträchtlicher Teil der KI-Wertschöpfung an ausländische Hyperscaler abfließt.
    - **Ökosystem-Effekte.** Je mehr Rechenleistung zur Verfügung steht, desto attraktiver wird Deutschland als Forschungs- und Start-up-Standort für in- und ausländisches Spitzentalent.
- Gleichzeitig geht souveräne Recheninfrastruktur aktuell mit zwei Nachteilen einher:
    - Zum einen bieten US-Hyperscaler wie AWS, Microsoft Azure oder Google Cloud ein umfassendes Hardware-/Software-Ökosystem, das Unternehmen eine besonders leichte Entwicklung und Bereitstellung eigener KI-Anwendungen ermöglicht. Dieser “Platform-as-a-Service”-Ansatz nimmt insbesondere Start-ups viel Aufwand ab und lässt sich durch kleinere Cloud-Anbieter nur schwer replizieren.
    - Zum anderen erfordert der Aufbau souveräner Recheninfrastruktur einen sehr hohen Kapitalaufwand. Neue KI-Rechenzentren und die dafür nötige Energieinfrastruktur sind eine Milliardeninvestition, die häufig die Finanzkraft einzelner Unternehmen übersteigt.

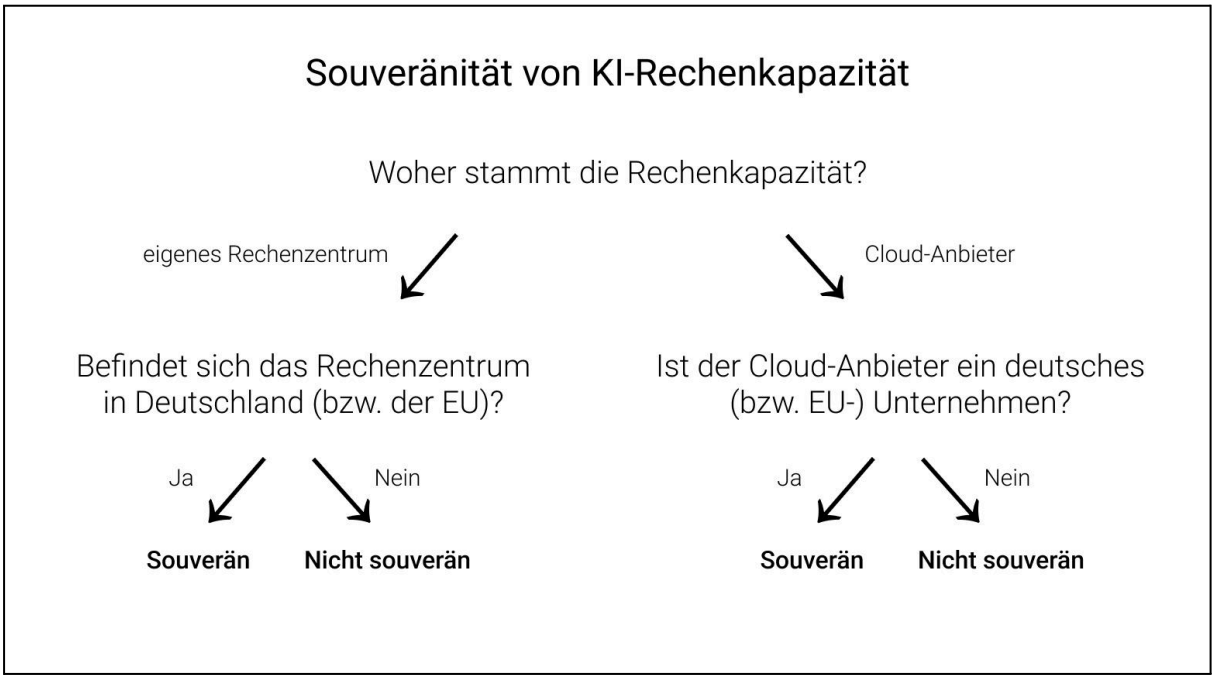

Abbildung 3: Ob Rechenkapazität im Sinne dieses Bericht als souverän gilt, hängt vom physischen Standort des Rechenzentrums und dem Sitz des Betreibers ab.

Wer über mehr Rechenkapazität verfügt, kann leistungsfähigere KI-Modelle trainieren und diese Modelle in mehr Prozesse integrieren. Sogenannte „Scaling Laws" beschreiben in der KI-Forschung einen empirischen Zusammenhang zwischen (u. a.) der Rechenleistung, die für das Training eines KI-Modells eingesetzt wird, und der Performance eines Modells – vereinfacht gesagt führt mehr Rechenleistung zu „intelligenterer" KI. Neuere Erkenntnisse zeigen zudem, dass sich die Modell-Performance nicht nur durch mehr Rechenleistung während des Trainings steigern lässt. Die Performance von „Reasoning Models" wie GPT-5-Thinking (OpenAI) oder Gemini 2.5 Pro Deep Think (Google) erhöht sich auch, wenn dem Modell während des Betriebs zusätzliche Rechenleistung zur Verfügung gestellt wird. Diese Modelle können Rechenleistung in längere Denkprozesse umwandeln, um Probleme logisch stringent zu analysieren oder mehrere Lösungsansätze auszuprobieren – und damit besonders komplexe Aufgaben lösen.

Deutschlands KI-Rechenkapazität fällt im internationalen Vergleich zurück. Daten des Forschungsinstituts Epoch AI zeigen: Aktuell verfügt Deutschland über weniger als 2% der globalen KI-Rechenkapazität (Pilz et al. 2025). Zwar gehört Deutschland damit noch zu der kleinen Gruppe an Ländern, die relativ gesehen nach den KI-Schwergewichten USA und China über die größte KI-Rechenkapazität verfügen. Dieses Bild verschlechtert sich aber deutlich, sobald man berücksichtigt, wie verschiedene Ländern ihre Kapazität in den nächsten Jahren ausbauen wollen: Deutschland bleibt dann in der vorletzten von insgesamt vier Leistungsgruppen – zu der in diesem Fall auch Länder wie Brasilien, Russland, Kanada, Vietnam und Malaysia gehören –, während Frankreich und das Vereinigte Königreich in die zweithöchste Gruppe aufsteigen (Abbildung 4). Mit einer europäischen KI-Gigafabrik (siehe Box 2) könnte es Deutschland knapp in die zweithöchste Klasse schaffen – vorausgesetzt, Deutschland erhält einen Zuschlag der Europäischen Union, die Gigafabrik entsteht wie geplant und andere Ländern bauen ihre KI-Infrastruktur nicht stärker aus als aktuell erwartet.

Selbst dann hätte Deutschland aber voraussichtlich weniger Kapazität als China, Frankreich, Indien, Saudi-Arabien und das Vereinigte Königreich – geschweige denn die Vereinigten Arabischen Emirate und die USA, deren Ausbaupläne mit Abstand am ambitioniertesten sind (Patel et al. 2025a).[1]

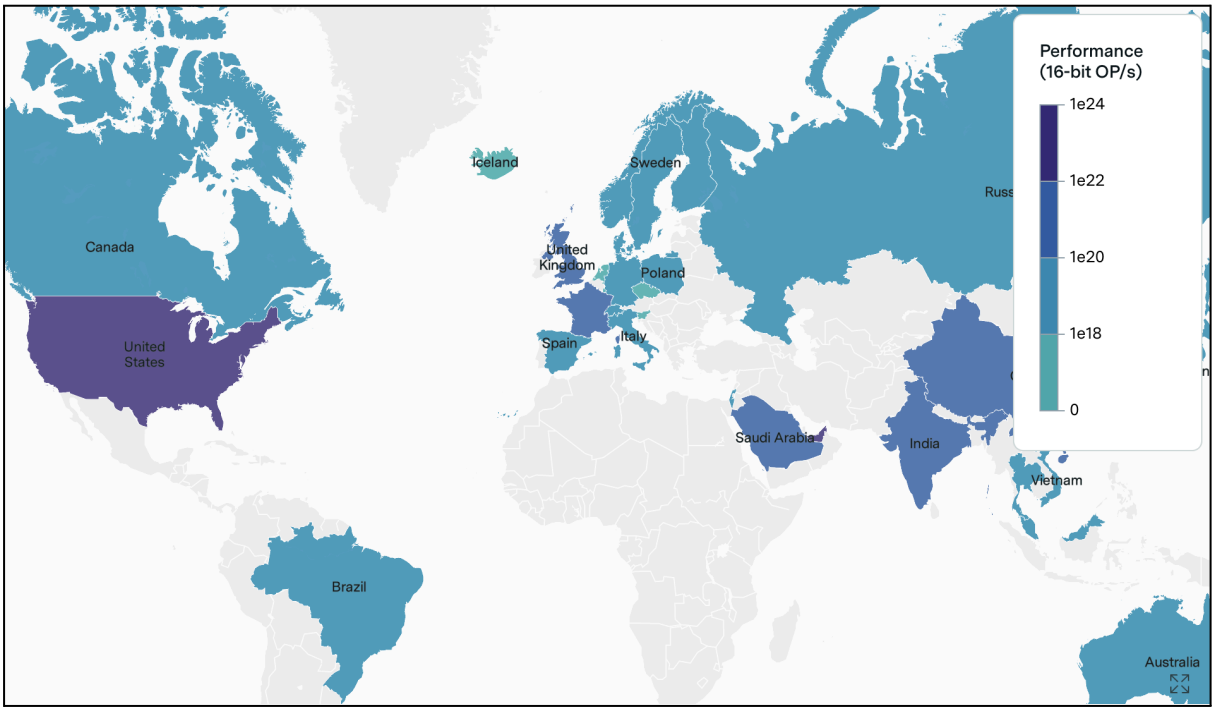

Abbildung 4: Deutschlands Pläne für den Ausbau von Rechenkapazität sind im internationalen Vergleich wenig ambitioniert. Eine mögliche europäische KI-Gigafabrik ist hier nicht eingerechnet. Aber auch mit einer Gigafabrik läge Deutschland im Ranking voraussichtlich hinter Ländern wie Frankreich, Indien, Saudi-Arabien oder den USA. Grundlage für den Vergleich sind Daten über die Rechenleistung von Supercomputern, gemessen in 16-Bit-Operationen pro Sekunde (FLOP/s) (Pilz et al. 2025). Der Datensatz deckt etwa 10-20% der globalen Supercomputer-Performance ab, ist die derzeit umfassendste Analyse dieser Art und liefert eine repräsentative Übersicht der globalen Verteilung von KI-Rechenkapazität.

Box 2: Europäische KI-Gigafabriken

- Die Europäische Union plant im Zuge ihres AI Continent Action Plan bis zu 5 europäische KI-Gigafabriken: große Rechenzentren für KI-Training und -Inferenz, ausgestattet mit jeweils etwa 100.000 leistungsfähigen KI-Chips (European Commission 2025a, EuroHPC 2025).
- Dafür stellt die EU 20 Milliarden € im Rahmen der InvestAI-Initiative zur Verfügung und rechnet je Gigafabrik mit Kosten von 3-5 Milliarden € (EuroHPC 2025). Andere halten hingegen 6-8 Milliarden € für realistischer (Hess 2025). Nicht eingerechnet sind Ausgaben für den laufenden Betrieb, z. B. Energiekosten. Die EU bzw. ihre Mitgliedsstaaten unterstützen den Bau einer Gigafabrik mit bis zu 35% des Kapitalaufwands (CAPEX). Weitere Kosten, einschließlich der gesamten Betriebskosten (OPEX), werden vom Privatsektor getragen (EuroHPC 2025).

[1] Für das Vereinigte Königreich ergibt sich dieser Befund aus der jüngsten Ankündigung, 120.000 GPUs der neuen Blackwell-Generation anzuschaffen (NVIDIA 2025b) – was mindestens 300.000 H100-Äquivalenten entspricht (siehe Fußnote 2). Für die anderen Länder folgt dies aus den Daten von Epoch AI (Pilz et al. 2025).

- Die Bundesregierung will mindestens eine Gigafabrik in Deutschland ansiedeln. Nachdem die Verhandlungen führender deutscher Unternehmen über eine gemeinsame Bewerbung scheiterten, haben verschiedene Unternehmen eigene Interessenbekundungen eingereicht – darunter die Schwarz Gruppe, die Deutsche Telekom, Ionos (gemeinsam mit Hochtief), der Freistaat Bayern, das Start-up Lyceum sowie ein Konsortium unter der Führung von Silicon Saxony (Handelsblatt 2025, FAZ 2025).
- Es ist derzeit unklar, ob ein deutsches Unternehmen bzw. Konsortium den Zuschlag erhält und, falls ja, ob in der Folge eine KI-Gigafabrik in Deutschland entsteht – Lyceum plant diese beispielsweise aufgrund günstigerer Energie und schnellerer Genehmigungsverfahren in Dänemark (FAZ 2025).
- Das offizielle Bewerbungsverfahren soll 2026 beginnen und bis zum Ende des 2. Quartals abgeschlossen sein. Selbst bei einer erfolgreichen Bewerbung würde eine Gigafabrik deshalb wohl frühestens 2027 in Betrieb gehen. Verzögerungen im Ausschreibungsverfahren, bei der Finanzierung, im Planungs- und Genehmigungsprozess oder bei der Netzanbindung könnten auch zu einem späteren Betriebsbeginn führen.

Deutschland ist mit seinen KI-Supercomputern international nicht konkurrenzfähig und wird in den nächsten Jahren weiter zurückfallen. Deutschland droht, aufgrund mangelnder Ambition beim Ausbau von Rechenkapazität als KI-Standort weiter abgehängt zu werden. Ein Blick auf deutsche KI-Supercomputer im internationalen Vergleich stützt diesen Befund (Tabelle 1). Selbst das Leuchtturmprojekt “Jupiter” am Forschungszentrum Jülich liegt mit seinen ca. 24.000 H100-Äquivalenten (FZ Jülich 2025) deutlich hinter den größten US-Systemen, die mit bis zu 275.000 H100-Äquivalenten ausgestattet sind (siehe Tabelle 1). Ein Blick auf geplante Systeme zeigt, dass der Abstand in Zukunft eher noch weiter wächst: Megaprojekte wie ein Cluster mit 20 Millionen H100-Äquivalenten in den VAE oder 5 Millionen H100-Äquivalenten in Südkorea stellen die im Koalitionsvertrag für Deutschland vorgesehene KI-Gigafabrik mit ihren 100.000 H100-Chips in den Schatten (Pilz et al. 2025). Auch ein aktueller Bericht des BMWK aus der vergangenen Legislaturperiode kommt zu dem Ergebnis, dass Deutschland in Sachen Rechenkapazität trotz der aktuellen Pläne weiter hinter Länder wie die USA zurückfällt (BMWK 2025).

KI-Supercomputer in Deutschland und international

Deutschland						International					
System	**Standort**	**Eröffnung**	**# GPUs**	**Typ**	**H100-Äquivalente[2]**	**System**	**Standort**	**Eröffnung**	**# GPUs**	**Typ**	**H100-Äquivalente**
Vorhanden											
Jupiter AI Factory	Jülich	2025	23.762	NVIDIA GH200	**23.762**	xAI Colossus	Memphis, USA	2025	230.000	NVIDIA H100/200/ GB200	**ca. 275.000[3]**
JUWELS-Booster	Jülich	2020	3.744	NVIDIA A100	**1181**	Meta 100k Cluster	USA	2024	100.000	NVIDIA H100	**100.000**
Hunter	Stuttgart	2025	752	AMD Instinct MI300A	**745**	OpenAI/Micro-soft Cluster	Goodyear, USA	2024	100.000	NVIDIA H100	**100.000**
Capella	Dresden	2024	624	NVIDIA H100	**624**	Oracle OCI Supercluster	USA	2025	65.536	NVIDIA H200	**65.536**
hessian.AI fortythree	Darmstadt	2024	912	NVIDIA A100/ H100	**479**	Tesla Cortex Phase 1	USA	2024	50.000	NVIDIA H100	**50.000**
Helma	Nürnberg	2024	384	NVIDIA H100	**384**	El Capitan	Livermore, USA	2024	44.544	AMD Instinct MI300A	**44.143**

[2] H100-Äquivalente sind ein gängiges Maß, um die Rechenleistung unterschiedlicher GPU-Typen miteinander zu vergleichen (siehe technischer Appendix).
[3] xAI-Geschäftsführer Elon Musk gab im Juni 2025 an, dass Colossus über 150.000 H100-, 50.000 H200- und 30.000 GB200-GPUs verfüge (Y Combinator 2025).

Geplant											
KI-Gigafabrik	Deutsch-land	?	100.000	?	**100.000**	Stargate 5GW Campus	Abu Dhabi	2030	?	?	**20.262.759**[4]
Telekom/Nvi-dia Industrial Cloud[5]	Deutsch-land	2026	10.000	?[6]	**max. 25.265**	3GW Cluster	Südkorea	2028	?	?	**5.103.588**
ELBJUWEL	Dresden	?	?	?	**25.265**	xAI Colossus 2	Memphis, USA	2026	800.000	NVIDIA H100/ GB200	**2.221.223**
HammerHAI AI Factory	Stuttgart	?	?	?	**?**[7]	Fluidstack 1GW Campus	Frank-reich	2028	500.000	NVIDIA GB200	**1.263.264**
Herder	Stuttgart	2027	?	?	**?**[8]	Reliance Industries SC	Indien	2027	450.000	NVIDIA GB200	**1,136,938**
Blue Lion	Garching	2027	~4.400	?	**?**[9]	Stargate OCI Supercluster 2	Abilene, USA	?	?	?	**1,010,615**

Tabelle 1: Eine Auswahl bedeutender KI-Supercomputer in Deutschland und international zeigt, dass die deutsche Rechenkapazität sowohl bei vorhandenen als auch geplanten Clustern hinter anderen Ländern zurückfällt. Die Daten für geplante (und teilweise auch für vorhandene) Cluster basieren teilweise auf inoffiziellen Berichten und sind entsprechend unsicher, liefern aber einen hinreichend repräsentativen Überblick über die globale Verteilung von KI-Rechenkapazität. Quelle: Pilz et al. 2025, eigene Recherche

[4] Bei einigen angekündigten Projekten sind die Anzahl und der Typ der verwendeten GPUs unbekannt. Epoch AI schätzt in diesen Fällen die Anzahl an H100-Äquivalenten anhand verfügbarer Informationen wie der angekündigten Stromkapazität des Rechenzentrums.

[5] Die Deutsche Telekom sieht die gemeinsam mit Nvidia geplante Industrial Cloud als Vorstufe einer europäischen KI-Gigafabrik (Telekom 2025). Sollte der Zuschlag für eine Gigafabrik an die Telekom gehen, könnten die max. 25.265 H100-Äquivalente der Industrial Cloud daher Teil der 100.000 GPUs für die Gigafabrik werden.

[6] Laut offizieller Ankündigung wird die Industrial Cloud mit 10.000 GPUs ausgestattet sein, darunter Nvidia-Chips der neuen Blackwell-Generation (NVIDIA 2025a). Unklar ist, ob neben GPUs des Typs B200 auch ältere GPUs eingesetzt werden (und, falls ja, wie viele).

[7] Im Rahmen der EuroHPC-Initiative plant das HLRS Stuttgart den Bau einer europäischen AI Factory. Typ und Anzahl der GPUs sind noch unbekannt. Das Projektbudget beträgt 85 Millionen € (HLRS 2024).

[8] Der Herder-Supercomputer am HLRS Stuttgart folgt auf das Übergangssystem Hunter (siehe Tabelle) aus dem Jahr 2025 und ist als Exascale-System angekündigt. Typ und Anzahl der Chips werden laut HLRS bis Ende 2025 festgelegt (HLRS 2023).

[9] Blue Lion soll am LRZ in Garching bei München entstehen und mit Nvidia-GPUs der neuesten Rubin-Architektur ausgestattet sein (LRZ 2025). Schätzungen zufolge könnten dabei 4.400 der noch nicht verfügbaren VR200-GPUs zum Einsatz kommen, was ca. 26.000 H100-Äquivalenten entsprechen könnte (The Next Platform 2024). Das Projektvolumen beträgt 250 Millionen €.

Ein Mangel an Rechenkapazität gefährdet Deutschlands Wettbewerbsfähigkeit und geopolitische Souveränität. Mit der geringen Ambition beim Ausbau von KI-Infrastruktur gehen für die Zukunft gravierende Risiken einher:

- **Abgehängte Wirtschaft.** Das deutsche BIP-Wachstum könnte langfristig hinter Staaten zurückfallen, die über mehr Rechenkapazität für die Entwicklung und Nutzung von KI verfügen und damit einen größeren Teil der Wertschöpfung vereinnahmen können.
- **Abgehängte Verwaltung.** Die staatliche Verwaltung könnte im internationalen Vergleich zunehmend rückschrittlich und dysfunktional werden, wenn durch KI ermöglichte Effizienzgewinne ungenutzt bleiben.
- **Datenabfluss.** Deutschland könnte gezwungen sein, sensible Industrie- oder personenbezogene Daten an ausländische Cloud-Anbieter preiszugeben.
- **Talentabfluss.** Deutschland könnte als Standort für heimisches oder ausländisches KI-Talent noch unattraktiver werden, weil andere Länder eine bessere KI-Infrastruktur bieten.

Angesichts der herausragenden strategischen Bedeutung von KI sollte die Bundesregierung dringend den Ausbau von KI-Rechenkapazität vorantreiben. Die folgenden Abschnitte stellen dazu konkrete Strategien und Empfehlungen vor.

Box 3: DeepSeek – (k)ein Gamechanger?

- Im Januar 2025 veröffentlichte das chinesische KI-Unternehmen DeepSeek das Reasoning-Modell R1, dessen Performance nur knapp hinter führenden US-Modellen lag. Laut eigenen Angaben erreichte DeepSeek diese Performance mit deutlich weniger Rechenleistung – und damit wesentlich günstiger – als die amerikanische Konkurrenz. Für das zugrunde liegende Modell V3 bezifferte DeepSeek die Kosten für den finalen Trainingslauf – die nur einen Teil der gesamten Entwicklungskosten abbilden – auf bloß 5,6 Millionen Dollar (DeepSeek-AI 2024). Dieser Wert liegt einerseits deutlich unter den 100 Millionen Dollar, die beispielsweise OpenAI im Jahr 2023 für die Entwicklung ihres Modells GPT-4 aufgewendet hat. Andererseits bleiben darin viele Kostenfaktoren unberücksichtigt, was einen Vergleich mit der Zahl von OpenAI erschwert.
- Auf die Veröffentlichung folgte eine heftige Reaktion an den Finanzmärkten. Die Börsenwerte von Chipunternehmen wie Nvidia brachen ein – manche Marktteilnehmer gingen möglicherweise davon aus, dass Spitzenmodelle in Zukunft keine so hohen Mengen an Rechenleistung mehr benötigen würden. Parallel dazu wuchs in den USA die Angst, im globalen KI-Wettlauf gegen das womöglich effizientere China zu verlieren. In Europa ließ der Erfolg DeepSeeks hingegen die Hoffnung aufkommen, bald zu vergleichbar niedrigen Kosten eigene Spitzenmodelle trainieren zu können. Diesen Reaktionen liegen aber zum Teil irreführende Annahmen zugrunde:
    - Die öffentlich häufig genannte Summe von 5,6 Millionen Dollar bezieht sich allein auf den finalen Trainingslauf von V3. Weitere Kosten, wie z. B. für GPUs, Strom oder Gehälter, sind nicht eingerechnet. Experten schätzen, dass die tatsächlichen Entwicklungskosten deutlich höher liegen (Patel et al. 2025b).
    - DeepSeek hat gezeigt, dass sich die Effizienz von KI-Modellen durch kluge Ideen in der Tat beachtlich steigern lässt. Fachleuten war allerdings schon

vorher bekannt, dass sich ein gegebenes Performance-Niveau nach einigen Monaten mit deutlich weniger Ressourcen erreichen lässt (Pilz, Heim & Brown 2023). Neu waren deshalb nicht die Effizienzsteigerungen selbst, sondern dass diese von einem chinesischen Unternehmen realisiert wurden. Damit bleibt aber der grundlegende Trend unberührt, der aktuell die massiven Investitionen in KI-Recheninfrastruktur antreibt: Wer mehr Rechenkapazität besitzt, kann bessere Modelle trainieren – was sich auch daran zeigt, dass über ein halbes Jahr nach der Veröffentlichung von DeepSeeks R1 die führenden Modelle weiterhin aus den USA stammen.
    - Rechenkapazität braucht es nicht nur für KI-Training, sondern auch für KI-Inferenz (siehe Box 4). Selbst wenn China ähnlich gute Modelle zu geringeren Kosten trainieren könnte, könnte es diese weniger intensiv nutzen als andere Länder. Zum Beispiel kann DeepSeek Berichten zufolge die eigenen Modelle nicht wie gewünscht in der Breite anbieten, weil dafür aufgrund US-amerikanischer Exportkontrollen die Chips fehlen (The Information 2025).
- Für die deutsche KI-Rechenzentrenstrategie ergeben sich zwei wichtige Implikationen:
    - Rechenkapazität ist weiterhin ein entscheidender Faktor für leistungsfähige KI. Das Training von Spitzenmodellen erfordert in der Zukunft Milliarden-Investitionen. Auch etwas weniger leistungsfähige Modelle, deren Performance einige Monate hinter den besten US-Modellen liegt, lassen sich in den nächsten Jahren nicht mit wenigen Millionen Euro entwickeln.
    - Selbst wenn sich Spitzenmodelle in der Zukunft günstiger trainieren ließen – oder Deutschland ganz auf das Training solcher Modelle verzichten würde –, bräuchte Deutschland zusätzliche Rechenkapazität für KI-Inferenz, um KI flächendeckend in Wirtschaft und Verwaltung einzusetzen.

2. Strategien

Je nach Ambitionslevel ergeben sich für Deutschland drei mögliche Compute-Strategien, die jeweils unterschiedliche Schwerpunkte auf Training und Inferenz legen. Die Bundesregierung sollte Klarheit darüber schaffen, wie ambitioniert die deutsche KI-Infrastruktur bis zum Ende der Legislatur ausgebaut werden soll. Die Entscheidung darüber hängt von zwei zentralen Faktoren ab:

1. Wie hoch schätzt die Bundesregierung die zukünftige strategische Bedeutung von KI ein (für Wettbewerbsfähigkeit, Souveränität und nationale Sicherheit)?
2. Wie hoch ist die Bereitschaft der Bundesregierung, die notwendigen politischen und wirtschaftlichen Rahmenbedingungen für einen erfolgreichen KI-Standort zu schaffen?

Je nach den Antworten auf diese Fragen könnte sich Deutschland primär als Technologienehmer sehen, der ausländische KI-Modelle in die eigene Wirtschaft und Verwaltung integriert, oder als Technologieführer, der auch die Entwicklung eigener KI vorantreibt. Davon hängt ab, wie viel Kapazität jeweils für KI-Inferenz bzw. KI-Training benötigt wird (siehe Box 4). Als Technologienehmer bräuchte Deutschland vor allem

Inferenz-Kapazität und müsste die eigene KI-Infrastruktur dafür mit niedriger bis mittlerer Ambition ausbauen. Als Technologieführer bräuchte Deutschland darüber hinaus Kapazität für das Training (und dann auch die Inferenz) eigener Spitzenmodelle, was enorme zusätzliche Investitionen in entsprechende Rechenkapazität bedeuten würde – und ein entsprechend hohes Ambitionslevel erfordert.

Box 4: Training vs. Inferenz

- Rechenleistung kommt in zwei Phasen des KI-Lebenszyklus zum Einsatz:
    - Während des **Trainings** werden die Parameter – die internen Stellschrauben eines Modells – anhand großer Datenmengen optimiert, damit das Modell zunehmend bessere Outputs liefert. Dabei identifiziert das Modell Muster in den Daten und lernt, diese auf neue Fälle anzuwenden. Unterschieden wird zwischen dem *Pre-Training*, bei dem ein Modell vor allem auf der Grundlage von Internetdaten allgemeines Wissen erlangt, und dem *Post-Training*, bei dem kuratierte Daten und menschliches bzw. KI-generiertes Feedback genutzt werden, um das Modell für spezifische Aufgaben oder Verhaltensweisen zu optimieren (z. B. mathematische Probleme zu lösen oder Code zu schreiben). Während das Pre-Training traditionell der rechenintensivste Schritt ist, wird bei neueren Modellen immer mehr Rechenleistung für das Post-Training aufgewendet (You 2025).
    - Während der **Inferenz** wendet ein Modell das im Training erlangte Wissen an und generiert für einen gegebenen Input den entsprechenden Output. Kommerzielle Anbieter stellen ihre Modelle dafür teils Hunderten Millionen Menschen zur Verfügung (Chatterji et al. 2025). Seit Ende 2024 zeigt sich ein Trend zu immer rechenintensiverer Inferenz: “Reasoning Models” liefern bessere Ergebnisse, wenn sie während der Inferenzphase mehr Rechenleistung zur Verfügung haben und ausführlicher über ein Problem nachdenken können.
- Grundsätzlich kann jedes KI-Rechenzentrum gleichermaßen für Training und Inferenz genutzt werden. Allerdings haben die Eigenschaften eines KI-Rechenzentrums einen großen Einfluss darauf, welche Nutzungsart im Einzelfall sinnvoller ist. In der Praxis stellen Training und Inferenz deshalb jeweils unterschiedliche Anforderungen an KI-Rechenzentren (Fist & Datta 2024):
    - **Für Inferenz eignen sich auch ältere Chips.** Chips der jeweils neuesten Generation haben meist sowohl für Training als auch für Inferenz die beste Performance. Dennoch lassen sich Chips älterer Generationen, die für das Training kompetitiver Modelle nicht mehr geeignet sind, in der Regel noch wirtschaftlich für Inferenz nutzen. Bei unklarem strategischen Fokus kann es trotzdem sinnvoll sein, Chips der neuesten Generation zu erwerben, da diese sich vielseitig für Training und Inferenz einsetzen lassen.
    - **Für Inferenz genügen kleinere Rechenzentren.** Für das Training von Spitzenmodellen braucht es eine hohe Anzahl an Chips, die zentral an einem Ort verbunden werden. Weil während der Inferenzphase hingegen kleinere Datenmengen verarbeitet werden, können für Inferenz kleinere Rechenzentren genutzt werden, die weniger Energie benötigen und physisch über ein Land verteilt sein können.

- **Für Inferenz spielt mitunter die physische Nähe zum Nutzer eine Rolle.** Anders als beim Training spielt bei KI-Inferenz bei manchen Anwendungen – z. B. am Finanzmarkt oder in der Industrieproduktion – die Latenz eine wichtige Rolle (d.h. die Verzögerung zwischen der Eingabe eines Nutzers und der Antwort des Modells). Diese lässt sich minimieren, indem KI-Rechenzentren in der Nähe ihrer Nutzer gebaut werden. Bei Rechenzentren für KI-Training spielen dagegen andere Faktoren eine wichtigere Rolle, wie z. B. die Verfügbarkeit großer Energiemengen.

Drei paradigmatische Compute-Strategien stehen Deutschland zur Verfügung.

Strategie 1 (niedrige Ambition)

- Diese Strategie sieht vor, sich auf die Nutzung (Inferenz) ausländischer Spitzenmodelle zu konzentrieren und dafür in begrenztem Maße eigene Kapazität zu schaffen, z. B. um KI-Agenten in die Wirtschaft und Verwaltung zu integrieren. Darüber hinaus würde Deutschland nur branchenspezifische KI-Modelle und -Anwendungen entwickeln, deren Training vergleichsweise wenig Rechenleistung benötigt. Wir gehen davon aus, dass der Gesamtbedarf für Inferenz und branchenspezifisches Training in Deutschland Ende 2028 bei ca. 4,3 Mio. H100-Äquivalenten liegt (siehe technischer Appendix). Weiterhin nehmen wir an, dass Deutschland bei einer niedrigen Ambition nur 20% dieses Bedarfs durch eigene KI-Rechenzentren abdecken würde.[10] Dafür bräuchte Deutschland ca. 850.000 H100-Äquivalente bzw. 0,4% der globalen Rechenkapazität bis Ende 2028. Das entspricht einer IT-Anschlussleistung von ca. 0,8 GW.

Strategie 2 (mittlere Ambition)

- Diese Strategie sieht vor, den Gesamtbedarf für Inferenz und branchenspezifisches Training zu einem hohen Teil – 80% – selbst abzudecken. Dafür bräuchte Deutschland ca. 3,4 Mio. H100-Äquivalente bzw. 1,5% der globalen Rechenkapazität bis Ende 2028. Das entspricht einer IT-Anschlussleistung von ca. 3,4 GW. Damit wäre Deutschland deutlich besser positioniert, KI-Agenten in allen Bereichen der Wirtschaft und Verwaltung einzusetzen und würde seine Abhängigkeit von ausländischen Cloud-Diensten reduzieren.

Strategie 3 (hohe Ambition)

- Diese Strategie sieht vor, darüber hinaus Trainings- und Inferenz-Kapazität für eigene, kompetitive Spitzenmodelle mit allgemeiner Anwendung zu schaffen. Deutschland würde diese Modelle gemäß eigener Sicherheits- und Wertmaßstäbe entwickeln und ausländische Regierungen könnten den Zugang zu diesen Modellen nicht unmittelbar einschränken. Dafür bräuchte Deutschland eine deutlich höhere Rechenkapazität von ca. 6 Mio. H100-Äquivalenten bzw. 2,7% der globalen Rechenkapazität bis Ende 2028 – was unter dem deutschen Anteil am globalen BIP (nominal) von derzeit ca. 4,2% liegt (IMF 2025). Das entspricht einer IT-Anschlussleistung von 5,9 GW.

Tabelle 2 fasst diese Strategien und die dafür notwendigen politischen Erfolgsbedingungen zusammen. Ein Appendix am Ende dieses Berichts schlüsselt auf, wie die o. g. GPU-Kapazitäten bzw. IT-Anschlussleistungen berechnet wurden.

[10] Ob Deutschland die restliche Nachfrage durch ausländische Cloud-Dienste bedienen könnte, hängt von globalen Marktdynamiken und den geopolitischen Entscheidungen anderer Regierungen ab, die derzeit unsicher sind und auf die Deutschland nur begrenzt Einfluss hat.

Mögliche Compute-Strategien für Deutschland

Ambitionslevel	Strategie bis 2028	Beispielhaft: politische Erfolgsbedingungen
Niedrig	Fokus: ● Deutschland schafft eigene Inferenz-Kapazität für sicherheits- und latenzkritische Anwendungen und die Nutzung von KI-Agenten in Schlüsselsektoren. ● Zudem schafft Deutschland Trainings-Kapazität für branchenspezifische KI-Modelle und -Anwendungen (z. B. Industrial Foundation Models). Benötige GPUs: 850.000 H100-Äquivalente (= 0,4% der globalen Kapazität) ● davon z. B. 250.000 H100-Äquivalente in einer EU-Gigafabrik primär für KI-Training ● der Rest in kleineren, dezentralen Cluster für Inferenz	● Die Bundesregierung sieht KI als ein wichtiges Digitalthema, vergleichbar mit der elektronischen Patientenakte. ● “Business-as-usual”-Mentalität: Die Bundesregierung behandelt KI mit derselben Priorität wie die meisten anderen Themen im Tagesgeschäft. ● Die Bundesregierung hebt in ihrer aktualisierten KI-Strategie die Bedeutung schneller Genehmigungsprozesse für KI-Rechenzentren hervor und beruft zwecks Umsetzung eine Bund-Länder-Arbeitsgruppe ein. ● Verfügbare IT-Anschlussleistung für KI-Rechenzentren von ca. 0,8 GW. ● Die Bundesregierung fördert KI-Rechenzentren und das breitere KI-Ökosystem mit einem niedrigen zweistelligen Milliardenbetrag. ● Die Bundesregierung schafft im zuständigen Ministerium ein Referat, das den Infrastruktur-Ausbau koordiniert und sich mit relevanten Ressorts abstimmt.
Mittel	Fokus: ● Wie bei Ambitionslevel “niedrig”, aber höhere Inferenz-Kapazität, um KI-Agenten in allen Bereichen der Wirtschaft und Verwaltung zu nutzen und damit die Abhängigkeit von ausländischen Cloud-Diensten zu reduzieren Benötigte GPUs: 3.400.000 H100-Äquivalente (= 1,5% der globalen Kapazität) ● davon z. B. ca. 500.000 H100-Äquivalente in einem zentralen Cluster primär für KI-Training ● der Rest in kleineren, dezentralen Clustern für Inferenz	● Die Bundesregierung sieht KI als zentrale Zukunftstechnologie, die die Welt so stark verändern wird wie das Internet. ● “Fast-track”-Mentalität: Die Bundesregierung schreitet bei KI schneller voran als bei früheren Technologien und schafft staatliche Kapazität, die das Thema vorrangig behandelt. ● Die Bundesregierung novelliert binnen sechs Monaten die relevanten Gesetze (z. B. BauGB, BImSchG) und verkürzt Genehmigungsprozesse für KI-Rechenzentren damit auf weniger als ein Jahr (statt wie derzeit zwei Jahre oder sogar länger). ● Verfügbare IT-Anschlussleistung für KI-Rechenzentren von ca. 3,4 GW. ● Die Bundesregierung fördert KI-Rechenzentren und das breitere KI-Ökosystem mit einem mittleren zweistelligen Milliardenbetrag.

		• Die Bundesregierung setzt eine Taskforce mit zuständigen Staatssekretären/-sekretärinnen unter Leitung des Kanzleramtsministers ein, die den Infrastruktur-Ausbau über die Ressorts hinweg koordiniert.
Hoch	Fokus • Wie bei Ambitionslevel “mittel”, aber zusätzlich Kapazität, für das Training und die Inferenz eigener Spitzenmodelle mit allgemeiner Verwendung Benötigte GPUs: 6.000.000 H100-Äquivalente (= 2,7% der globalen Kapazität) • davon z. B. ca. 3.000.000 H100-Äquivalente in einem zentralen Supercluster für das Training von Spitzenmodellen • der Rest in kleineren und mittleren Clustern für Inferenz und das Training branchenspezifischer Modelle	• Die Bundesregierung sieht KI als entscheidende Technologie unserer Zeit, deren Einfluss mit der Industriellen Revolution vergleichbar ist und die globale Machtstrukturen tiefgreifend und rasant verändern wird. Sie hält eine Allgemeine Künstliche Intelligenz (AGI) noch vor Ende des Jahrzehnts für möglich. • “Whatever it takes”-Mentalität: Die Bundesregierung erklärt KI-Souveränität zur obersten Priorität, schwört alle relevanten Akteure auf dieses Ziel ein und setzt alle Hebel in Bewegung, um es mit “Warp Speed” zu erreichen. • Die Bundesregierung bringt in wenigen Wochen ein “KI-Beschleunigungsgesetz” durch den Bundestag, das ausgewählte KI-Rechenzentren als im “überragenden öffentlichen Interesse” einstuft, für die gezielte Ausweisung geeigneter Grundstücke sorgt und den Planungs- und Genehmigungsprozess damit auf max. sechs Monate verkürzt. • Verfügbare IT-Anschlussleistung für KI-Rechenzentren von ca. 5,9 GW. • Die Bundesregierung fördert KI-Rechenzentren und das breitere KI-Ökosystem mit einem Sondervermögen im niedrigen dreistelligen Milliardenbereich. • Die Bundesregierung setzt eine Taskforce mit allen relevanten Bundesministern/-ministerinnen unter Leitung des Kanzleramtsministers ein, die den Infrastruktur-Ausbau über die Ressorts hinweg koordiniert (angelehnt an die “Bunkerrunde” zum Bau von LNG-Terminals während der Energiekrise 2022).

Tabelle 2: Je mehr KI-Souveränität Deutschland anstrebt, desto ambitionierter sollte die nationale Compute-Strategie sein. Damit diese nicht ins Leere laufen, müssen auf politischer Ebene bestimmte Mindestbedingungen gegeben sein.

Selbst bei einem niedrigen Ambitionslevel gemäß Tabelle 2 müsste Deutschland seine KI-Rechenkapazität wesentlich stärker ausbauen als aktuell vorgesehen. Werden alle derzeit bekannten Projekte realisiert, würde Deutschland seine KI-Rechenkapazität in den nächsten zwei bis drei Jahren auf höchstens 225.000 H100-Äquivalente erhöhen. Das ist nur etwa ein Viertel der 850.000 H100-Äquivalente, die Strategie 1 als Ziel für die deutsche KI-Rechenkapazität vorsieht. Unabhängig vom geplanten Ambitionslevel besteht daher dringender Handlungsbedarf.

Deutschland braucht unabhängig von seinem Ambitionslevel mindestens 200.000 H100-Äquivalente in souveränen KI-Rechenzentren. Souveräne Rechenzentren (wie in Box 1 definiert) sind essentiell, um KI in sensiblen oder sicherheitskritischen Sektoren zu nutzen. Dazu gehören der Einsatz von KI in der kritischen Infrastruktur, der ein Höchstmaß an Souveränität und Ausfallsicherheit verlangt, oder die Verarbeitung sensibler Daten, wie z. B. Betriebsgeheimnisse in der Industrie. In diesen Bereichen braucht Deutschland eigene Trainings- und Inferenz-Kapazität, die höchsten Sicherheitsstandards entspricht und über die sich jederzeit unabhängig von äußeren Einflüssen verfügen lässt. Der genaue Bedarf für die nächsten Jahre lässt sich aufgrund von Unsicherheiten über technische Faktoren und das Diffusionstempo von KI nur schwer präzise ermitteln. Dennoch wäre es ein strategisches Versäumnis, aufgrund dieser Unsicherheit nicht zu handeln. Angesichts des rasanten KI-Fortschritts und der wachsenden Bedeutung von KI für staatliche und wirtschaftliche Schlüsselfunktionen besteht das größte Risiko darin, durch eine zaudernde Haltung in Kernbereichen gefährliche Abhängigkeiten einzugehen. Deutschland sollte daher bis spätestens 2028 eine souveräne Mindestkapazität von 200.000 H100-Äquivalenten aufbauen, die aus Sicherheitsgründen dezentral über das Land verteilt sein sollte.

Je ambitionierter die Compute-Strategie, desto mehr politische Voraussetzungen müssen erfüllt sein. Der massive Ausbau von Recheninfrastruktur gemäß Strategie 3 wäre ein Vorhaben von nationaler Tragweite, das mit einem enormen finanziellen und logistischen Aufwand verbunden wäre. Mit einer “Business-as-usual”-Mentalität wäre so ein Vorhaben zum Scheitern verurteilt. Stattdessen wäre es nötig, KI als das zentrale Thema unserer Zeit zu begreifen und alle relevanten Akteure aus Politik und Wirtschaft darauf einzuschwören, KI-Infrastruktur über die nächsten Jahre mit unbedingter Priorität auszubauen. Dieser Ausbau trägt außerdem nur dann Früchte, wenn die Bundesregierung das gesamte KI-Ökosystem umfassend und mit großzügigen Ressourcen stärkt. Denn nur im Zusammenspiel mit weiteren Faktoren wie Kapital, Ideen und Talent wird Rechenkapazität zum Katalysator des deutschen KI-Standorts. Eine Strategie, die das Training eigener Grundlagenmodelle auf Spitzenniveau beinhaltet, erfordert deshalb eine “Whatever-it-takes”-Mentalität. Sind diese politischen Voraussetzungen nicht erfüllt, droht der Ausbau von Rechenkapazität zu versanden oder im schlimmsten Fall teure KI-Rechenzentren zu hinterlassen, deren Potential mangels eines starken KI-Ökosystems nicht ausgeschöpft wird.

Eigene Spitzenmodelle sind auch als europäisches Verbundprojekt mit geteilter Infrastruktur im EU-Ausland denkbar. Der Kapitalaufwand für das Training eigener Spitzenmodelle (Strategie 3) ist enorm. Als drittgrößte Volkswirtschaft wäre Deutschland unter den richtigen politischen Voraussetzungen in der Lage, die dafür nötigen Ressourcen bereitzustellen. Die damit verbundene Kraftanstrengung wäre dennoch groß und würde mit anderen gesellschaftlichen Prioritäten konkurrieren. Diese Anstrengung ließe sich

reduzieren, indem Deutschland souveräne Spitzenmodelle gemeinsam mit anderen Staaten entwickelt. Dafür kommen primär andere europäische Staaten in Frage – entweder als bilaterale Partnerschaft mit Frankreich oder Großbritannien oder als EU-weites Verbundprojekt –, aber auch andere Nationen wie Kanada, Australien, Japan oder Südkorea. Eine solche Zusammenarbeit würde Ressourcen bündeln und es Deutschland ermöglichen, von ausländischen Standortfaktoren zu profitieren – zum Beispiel der verlässlichen Versorgung mit Nuklearenergie in Frankreich oder der günstigen Energie durch Wasserkraft in Ländern wie Norwegen.

Deutschland muss sich angesichts einer unsicheren Zukunft entscheiden, für welche KI-Szenarien es sich wappnen will. Es ist nicht auszuschließen, dass sich der KI-Fortschritt in den nächsten Jahren verlangsamt. Selbst wenn Modellfähigkeiten morgen ein Plateau erreichen würden, könnten der Einfluss auf die Weltwirtschaft durch Hilfssysteme (*scaffolding*) und eine effizientere Einbettung in bestehende Prozesse vergleichbar mit dem Internet sein – dessen Beitrag zum globalen BIP-Wachstum laut einer Studie von 2011 bis zu 21% betrug (Pélissie du Rausas et al. 2011). Aktuell entwickeln sich KI-Fähigkeiten allerdings rasant weiter – in einigen Bereichen womöglich sogar exponentiell (METR 2025). Es gibt derzeit keinen wissenschaftlichen Konsens über das Ausmaß und Timing weiterer Fähigkeitssprünge. Viele Fachleute – darunter der meistzitierte KI-Forscher der Welt, Yoshua Bengio, und der Nobelpreisträger und KI-"Urvater" Geoffrey Hinton – halten es für möglich, dass eine Allgemeine Künstliche Intelligenz (AGI)[11] noch vor Ende des Jahrzehnts erreicht wird. Damit würde Intelligenz in einem nie dagewesenen Maß kopierbar bzw. skalierbar – ein manchmal als "Millionen Genies in einem Datenzentrum" beschriebenes Szenario. Der Einfluss von KI könnte dann mit der Industriellen Revolution vergleichbar sein: eine systemverändernde Disruption wirtschaftlicher und gesellschaftlicher Gleichgewichte, die sich allerdings im Zeitraffer vollziehen würde. Einige Experten gehen sogar davon aus, dass auf AGI in kurzer Zeit eine Superintelligenz (ASI) folgen könnte, die selbst die intelligentesten Menschen kognitiv in den Schatten stellen würde (Kokotajlo et al. 2025). Die damit verbundenen Veränderungen könnten präzedenzlos sein. Aus dieser Bandbreite an möglichen Szenarien ergibt sich eine strategische Unsicherheit: Für welches Maß an Veränderung sollte sich Deutschland wappnen?

Compute-Strategien sind gegenüber verschiedenen Zukunftsszenarien unterschiedlich robust. Abhängig davon, welchen Einfluss KI auf die Welt haben wird, kann eine Compute-Strategie mehr oder weniger sinnvoll sein. Das gilt besonders für Strategie 1 und 2 in Tabelle 3. Verändert KI die Welt in etwa dem Maße wie das Internet, handelt es sich hierbei um vielversprechende Strategien, um zu vertretbaren Kosten an der globalen KI-Wertschöpfung zu partizipieren. Sie gehen aber auch mit einem hohen Risiko einher: Ist der Einfluss von KI transformativer als frühere Technologien oder sogar zivilisationsverändernd, könnte Deutschland durch den Verzicht auf Inferenz-Kapazität und eigene Grundlagenmodelle wirtschaftlich und geopolitisch irreversibel abgehängt werden vergleichbar mit einer Gesellschaft, die sich der Dampfmaschine oder dem Ackerbau verweigert hätte. Strategie 3 ist dagegen in allen Szenarien positiv: Im transformativen Szenario würde sich Deutschland als führende KI-Nation Wohlstand und globalen Einfluss sichern, wenn auch zu potentiell hohen Opportunitätskosten. In den anderen beiden

[11] Der Begriff AGI hat keine kanonische Definition, meint im Allgemeinen aber eine KI, die dem Menschen in kognitiver Hinsicht ebenbürtig ist.

Szenarien erscheint eigene Spitzen-KI ohnehin als der einzige Weg, um in Zukunft souverän und wettbewerbsfähig zu sein.

	Einfluss von KI auf die Welt		
	1. disruptiv (≈Internet)	2. transformativ (≈Industrielle Revolution, im Zeitraffer)	3. zivilisationsverändernd (≈Neolithische Revolution, im Zeitraffer)
Strategie 1 (niedrige Ambition)	Wertschöpfung primär im Ausland, aber Partizipation an wirtschaftlicher Entwicklung	Abhängigkeit von ausländischen Technologieführern	Abhängigkeit von ausländischen Technologieführern
Strategie 2 (mittlere Ambition)	Führend in Schlüsselsektoren	Teilweise Abhängigkeit	Weitgehende Abhängigkeit
Strategie 3 (hohe Ambition)	Wachstum und geopolitische Souveränität, ggf. hohe Opportunitätskosten	Wachstum und geopolitische Souveränität	Wachstum und geopolitische Souveränität

Tabelle 3: Die verschiedenen Compute-Strategien beeinflussen die deutsche Wettbewerbsfähigkeit und geopolitische Souveränität unterschiedlich – abhängig davon, welchen Einfluss KI auf die Gesellschaft, Wirtschaft, Verwaltung und internationale Ordnung haben wird.

3. Empfehlungen

Wer im KI-Zeitalter souverän und wettbewerbsfähig sein will, braucht ausreichend Rechenleistung für das Training und den Betrieb von KI-Modellen. Dem Staat kommt dabei eine besondere Verantwortung zu: Angesichts der hohen Kosten für neue Infrastruktur und regulatorischer bzw. technologischer Unsicherheiten sind die Investitionsanreize für einzelne Unternehmen derzeit limitiert. Die Bundesregierung sollte den Privatsektor aufgrund der langfristigen strategischen Bedeutung von KI deshalb gezielt unterstützen. Es geht darum, die richtigen Rahmenbedingungen zu schaffen und den Infrastruktur-Ausbau auch gezielt über Public-Private-Partnerships zu fördern. Sechs Prioritäten müssten dabei im Zentrum stehen.

1. Kapazität für KI-Training und -Inferenz gemäß strategischer Prioritäten ausbauen

Deutschlands aktuelle Pläne zum Ausbau von KI-Infrastruktur sind bisher nicht ambitioniert genug. Selbst wenn alle derzeit geplanten Projekte realisiert werden sollten, würde die deutsche KI-Rechenkapazität in den nächsten 3 Jahren nur auf etwa 175.000-250.000 H100-Äquivalente steigen. Damit fiele Deutschland im internationalen Vergleich zurück. Statt eigener Kapazität für die Nutzung fortgeschrittener KI-Modelle, geschweige denn deren Training, würde sich Deutschland zunehmend abhängig von anderen KI-Nationen machen.

Eine europäische KI-Gigafabrik wäre ein erster Schritt und sollte bei einer positiven deutschen Bewerbung so schnell wie möglich gebaut werden, bliebe aber deutlich hinter dem prognostizierten Bedarf zurück.

Die deutsche Industrie ist vor allem aufgrund der hohen Anfangsinvestitionen noch verhalten, sich an großen KI-Infrastrukturprojekten zu beteiligen.

Empfehlungen:

- Die Bundesregierung sollte angesichts der oft mehrjährigen Planungs- und Bauphase jetzt damit beginnen, den Ausbau von KI-Rechenkapazität gezielt zu fördern.
- Parallel dazu gilt es, so schnell wie möglich strategische Klarheit darüber zu schaffen, mit welcher Ambition die deutsche KI-Rechenkapazität bis zum Ende der Legislatur ausgebaut werden soll und welcher Fokus dabei jeweils auf KI-Training bzw. -Inferenz liegt (siehe Abschnitt 2).
- Public-Private-Partnerships sollten genutzt werden, um finanzielle Risiken für Privatunternehmen zu begrenzen und dafür zu sorgen, dass der Infrastruktur-Ausbau dem langfristigen wirtschaftlichen und geopolitischen Interesse Deutschlands entspricht.

2. Politische Voraussetzungen schaffen und Kompetenzen bündeln

Der Ausbau von KI-Infrastruktur ist ein über Jahre andauerndes Milliarden-Projekt, dessen Erfolg von politischer Ausdauer und Ambition, einer klaren Strategie und gebündelter Koordinierung abhängt. Aktuell mangelt es an diesen Faktoren. Zwar betont die Bundesregierung die strategische Bedeutung von KI, einschließlich eines Bedarfs an eigenen KI-Grundlagenmodellen (Welt 2025). Die infrastrukturellen und politischen Voraussetzungen für ein derart ambitioniertes Ziel sind aber aktuell bei Weitem nicht gegeben. Auch die Schaffung von Inferenz-Kapazität, die es für weniger ambitionierte Strategien bräuchte, wird noch nicht systematisch unterstützt. Eine Herausforderung besteht darin, KI-Infrastruktur stärker als ein ressortübergreifendes, zentral zu steuerndes Thema zu behandeln. Das Beispiel der LNG-Terminals während der Energiekrise 2022 zeigt, dass sich strategisch wichtige Infrastrukturprojekte auch in kurzer Zeit realisieren lassen, wenn diese zentral aus dem Bundeskanzleramt heraus koordiniert werden. Neben den innenpolitischen Voraussetzungen braucht es außerdem eine koordinierte Außenwirtschaftspolitik, um den Zugang zu essentieller US-Hardware abzusichern, die sich absehbar nicht durch europäische Produktion ersetzen lassen wird.

Empfehlungen:

- Je nach Compute-Strategie sollte die Bundesregierung die in Tabelle 2 genannten politischen Voraussetzungen schaffen. Dazu gehört u. a. ein in der gesamten Regierung herrschendes Bewusstsein über die unmittelbare strategische Relevanz von KI, ein Ausbau der für KI-Rechenzentren notwendigen Energiekapazität, bürokratische Entlastungen sowie eine gezielte Förderung des KI-Ökosystems.
- Eine zentrale Instanz – z. B. ein KI-Infrastruktur-Board im Bundeskanzleramt (KI Bundesverband 2024, Bitkom 2025) – sollte den Ausbau an Rechen- und Energiekapaziät steuern. Diese Instanz sollte Zuständigkeiten über die relevanten Ressorts hinweg koordinieren (Digitales, Energie, Wirtschaft, Inneres, Forschung),

bürokratische Hürden abbauen und eng mit europäischen Initiativen wie EuroHPC zusammenarbeiten.
- Das Bundeskanzleramt sollte im engen Austausch mit anderen Ressorts eine verlässliche Versorgung mit den modernsten US-amerikanischen KI-Chips sicherstellen. Insbesondere sollte sie für Deutschland die Möglichkeit bewahren, US-Chips weiterhin für souverän betriebene KI-Rechenzentren erwerben zu können – statt ausnahmslos in KI-Rechenzentren, die von US-Firmen betrieben werden, wie zuletzt Berichte nahelegten (Export Compliance Daily 2025).

3. Günstige, verlässliche Energie für KI-Rechenzentren bereitstellen

Je nach Strategie benötigt Deutschland für KI-Rechenzentren eine IT-Anschlussleistung von 0,8 GW, 3,4 GW bzw. 5,9 GW (siehe Abschnitt 2). Um diesen Bedarf verlässlich zu decken, sollte Deutschland so schnell wie möglich die dafür nötige Energieinfrastruktur ausbauen – ein Unterfangen, das selbst unter den besten Voraussetzungen Jahre dauern wird. Darüber hinaus sollte Strom nicht nur verfügbar, sondern auch wettbewerbsfähig bepreist sein.

Empfehlungen:
- Der Ausbau der KI-relevanten Energieinfrastruktur sollte als nationale Priorität zentral vom Bundeskanzleramt aus gesteuert werden.
- KI-Rechenzentren sollten (wie im Koalitionsvertrag angekündigt) in die Strompreiskompensation einbezogen werden. Zudem sollten sie über einen Industriestrompreis gefördert werden (eco 2025). Die Bundesregierung sollte dafür die beihilferechtlichen Voraussetzungen schaffen, z. B. die Aufnahme der KI-Rechenzentrenbranche in die KUEBLL-Liste der EU (eco 2023).
- Darüber hinaus sind Steuervergünstigungen für KI-Rechenzentren denkbar, wie sie z. B. in den USA und einigen europäischen Staaten üblich sind (Lorentz et al. 2025).
- Im Rahmen möglicher Verbundprojekte sollte die Bundesregierung überprüfen, einen Teil der in Deutschland benötigten Rechenkapazität in anderen EU-Ländern anzusiedeln – vorzugsweise dort, wo mehr oder günstigere Energie vorhanden ist.

4. Planungs- und Genehmigungsprozesse beschleunigen

Langwierige Planungs- und Genehmigungsprozesse sind neben der Energieversorgung eine der größten Hürden beim Bau eines KI-Rechenzentrums in Deutschland. Wer Milliarden in KI-Infrastruktur investieren will, muss wissen, dass das geplante Vorhaben nicht jahrelang durch bürokratische Prozesse verzögert wird. Der Genehmigungsprozess dauert in anderen EU-Ländern wie Dänemark teils nur wenige Monate, statt wie in Deutschland zwei Jahre oder sogar länger (FAZ 2025, German Datacenter Association 2024). Die Anbindung eines neuen Rechenzentrums an das Stromnetz dauert in Deutschland ebenfalls lange: laut der Internationalen Energieagentur bis zu 7 Jahre, verglichen mit 3-5 Jahren in Spanien und 1-3 Jahren in den USA (IEA 2025). Das macht Investitionen entsprechend unattraktiv.

Empfehlungen:
- Die Bundesregierung sollte “Special Compute Zones” (SCZs) einrichten, in denen eine vereinfachte Regulatorik für neue KI-Rechenzentren und die dafür nötige Energieinfrastruktur gilt (Wiseman, McClements & Horsley 2024, Juijn et al. 2025, Datta & First 2025). Die Idee hinter SCZs besteht darin, dass die Bundesregierung

gemeinsam mit Ländern und Kommunen besonders geeignete Flächen (z. B. stillgelegte Industriestandorte mit vorhandener Netzkapazität) identifiziert. Für Genehmigungen in diesen Flächen wäre dann eine einzelne Behörde zuständig und es würde eine “presumption of conformity” herrschen: Legt die zuständige Behörde nicht innerhalb einer festen Frist (z. B. drei Monate) einen begründeten Widerspruch ein, gilt das Projekt als genehmigt.

- Auch darüber hinaus sollten Planungs- und Genehmigungsprozesse vereinfacht und beschleunigt werden, z. B. indem KI-Rechenzentren als im “überragenden öffentlichen Interesse” eingestuft werden, verschiedene Verfahren (wie Umweltverträglichkeitsprüfung und Baugenehmigung) parallelisiert werden und KI-Rechenzentren in die Liste bauplanungsrechtlich privilegierter Außenbereichsvorhaben aufgenommen werden (BMWK 2024).
- Die Eignung bundeseigener Grundstücke für den Bau von KI-Rechenzentren sollte überprüft werden.
- Bei der Netzanbindung sollten Vergabeverfahren vereinfacht und strategisch wichtige KI-Rechenzentren prioritär behandelt werden. Die Netzanbindung könnte außerdem für solche KI-Rechenzentren beschleunigt werden, die sich bereit erklären, ihre Leistung an Tagen mit Spitzenauslastung im Stromnetz zu drosseln. Daten aus den USA zeigen, dass dies nur an wenigen Tagen im Jahr nötig wäre (Norris et al. 2025). Zur Entlastung des bestehenden Stromnetzes sollte die Bundesregierung auch “Behind-the-meter-”Projekte unterstützen, bei denen KI-Rechenzentren durch eigene, vom Betreiber selbst auf dem Gelände errichtete Kraftwerke versorgt werden.

5. Sicherheit von KI-Rechenzentren erhöhen

Je leistungsfähiger KI-Modelle werden und je tiefer sie in Wirtschaft und Verwaltung integriert werden, desto attraktiver werden sie als Ziel von Cyberattacken – teilweise auch durch bestens ausgestattete, staatlich finanzierte Gruppen. Ein Bericht des Beratungsunternehmens Gladstone kommt zu dem Ergebnis, dass aktuell kein KI-Datenzentrum ausreichend gegen Angriffe durch letztere geschützt ist (Harris & Harris 2025). Neben Sabotage droht der Diebstahl sensibler Daten bis hin zu den Modellparametern selbst. Wer diese Parameter physisch oder per Cyberangriff aus einem KI-Rechenzentrum entwendet, könnte anschließend auf die Fähigkeiten des Modells zugreifen – und diese zum Beispiel für militärische Zwecke missbrauchen. Weil es einem Bericht der RAND Corporation zufolge mehrere Jahre dauern wird, KI-Rechenzentren effektiv vor staatlichen Cyberangriffen zu schützen, muss die Sicherheit hochsensibler Daten von Anfang an mitgedacht werden (Nevo et al. 2024). Tut Deutschland das nicht, könnten die USA irgendwann den Betrieb US-amerikanischer Spitzenmodelle auf deutschen KI-Rechenzentren untersagen, um nicht den Verlust wertvollen geistigen Eigentums zu riskieren.

Empfehlungen:

- KI-Rechenzentren, die Spitzenmodelle beherbergen oder für Anwendungen in der kritischen Infrastruktur relevant sind (z. B. Energieversorgung, Gesundheitswesen, Finanzmärkten), sollten mit einer ausreichend hohen Sicherheitsstufe ausgestattet sein (z. B. RAND SL4).
- Es sollte darauf hingearbeitet werden, KI-Rechenzentren, die hochsensible, unmittelbar für die nationale Sicherheit relevante Daten verarbeiten, mit der höchstmöglichen Sicherheitsstufe zu schützen (idealerweise RAND SL5).

- Staatliche Fachleute für IT-Sicherheit sollten eng mit privaten Betreibern von KI-Rechenzentren zusammenarbeiten, um diese auch auf Basis nicht-öffentlicher Informationen vor möglichen Angriffen zu schützen.
- Deutsche Fachbehörden (BSI, Cyberagentur) und Forschungseinrichtungen sollten sich auf internationaler Ebene aktiv an der Erarbeitung, Umsetzung und Prüfung geteilter Sicherheitsstandards beteiligen.

6. KI-Ökosystem umfassend stärken

Dem Ausbau der KI-Recheninfrastruktur muss ein wachsendes Ökosystem an Organisationen gegenüberstehen, die das durch neue Rechenkapazität entstehende Potential nutzen. Dieses Ökosystem ist in Deutschland gerade nicht ausgeprägt genug. Zwar gibt es einige Positivbeispiele wie die Start-ups DeepL oder Parloa, und auch in der Industrie spielen Industrial Foundation Models eine zunehmende Rolle. Dennoch ist die viel beschworene “KI made in Germany” derzeit eher ein Buzzword als eine ökonomische Realität. Daten aus dem Jahr 2024 zeigen außerdem, dass die KI-Adoptionsrate in deutschen Unternehmen zwar über dem EU-Durchschnitt liegt (19,75% vs. 13,48%), aber hinter Ländern wie Dänemark (27,58%), Schweden (25,09%) oder Belgien (24,71%) zurückfällt (Eurostat 2024). Aufgrund des langen Zeithorizonts für den KI-Infrastrukturausbau sollte das KI-Ökosystem parallel dazu in den Bereichen Wirtschaft, öffentliche Verwaltung und Forschung gestärkt werden.

Empfehlungen:

- Wirtschaft:
  - Pensionskassen, Versicherungen und Förderbanken sollten leichter in Tech-Startups investieren können, um dem akuten Mangel an Risikokapital in Europa entgegenzuwirken.
  - Die Bundesregierung sollte die Startup- und Scaleup-Strategie der EU unterstützen, die den europäischen Kapitalmarkt stärker vereinheitlichen, Investitionsflüsse erleichtern und regulatorische Unsicherheit beseitigen will (European Commission 2025b).
  - Die KI-Adoption in Unternehmen sollte gezielt gefördert werden, z. B. durch Sonderabschreibungen für KI-Investitionen oder eine bürokratiearme Umsetzung der europäischen KI-Verordnung.
  - Vereinfachte Visa-Prozesse sollten genutzt werden, um den deutschen KI-Standort für internationales Top-Talent attraktiv zu machen.
- öffentliche Verwaltung:
  - Die öffentliche Hand sollte bei technischer Gleichwertigkeit als Ankerkunde für heimische KI-Lösungen auftreten und damit verlässliche Umsätze für vielversprechende Unternehmen in der Wachstumsphase generieren.
  - Die Bundesregierung sollte den Bau ausgewählter KI-Rechenzentren attraktiver machen, indem sie Abnahmegarantien für einen Teil der Rechenleistung übernimmt und diese nutzt, um die öffentliche Verwaltung effizienter und zukunftsfähiger zu machen.
  - Damit der Staat seine Rolle als wichtiger Ankerkunde erfüllen kann, sollten öffentliche Beschaffungsregeln hinreichend einfach ausgestaltet sein.
- Forschung:

- Der Transfer zwischen universitärer Forschung und wirtschaftlicher Anwendung sollte vereinfacht werden, z. B. durch leichtere Ausgründungen, einen schnellen und unbürokratischen Zugang zu öffentlicher Rechenleistung für Start-ups, einen attraktiven Rechtsrahmen für Mitarbeiterbeteiligungen sowie eine vereinfachte Regulatorik gerade für kleine Unternehmen.
- Die öffentliche Forschungsförderung sollte darauf abzielen, gemeinsam mit europäischen Partnerländern ambitionierte Moonshot-Projekte zu wagen, damit der nächste technische Durchbruch für sichere, verlässliche und vertrauenswürdige KI aus Europa stammt. Auch Kooperationen mit Ländern wie Großbritannien, Kanada, Japan oder Australien sind denkbar, um einen Gegenpol zu US-amerikanischer und chinesischer KI zu schaffen.

4. Fazit

Deutschland steht vor einer kritischen Weichenstellung. KI-Rechenkapazität ist eine strategische Kernressource, die die deutsche Wettbewerbsfähigkeit und geopolitische Souveränität über Jahrzehnte hinweg bestimmen wird. Dennoch droht Deutschland im internationalen Vergleich zurückzufallen, selbst wenn man den Bau einer europäischen Gigafabrik berücksichtigt.

Noch kann Deutschland umsteuern. Die Bundesregierung sollte den Ausbau von KI-Infrastruktur zur Priorität machen und mit einer koordinierten Strategie vorantreiben. Entscheidend sind dabei sechs Hebel: der Ausbau von Trainings- und Inferenz-Kapazität gemäß strategischer Prioritäten, eine Bündelung politischer Kompetenzen, vereinfachte Regulatorik, verlässliche und günstige Energie, der Schutz von KI-Rechenzentren vor Sabotage und Diebstahl geistigen Eigentums sowie die umfassende Stärkung des gesamten KI-Ökosystems.

Technischer Appendix

Die drei Compute-Strategien in Abschnitt 2 sind durch verschiedene Rechen- bzw. Stromkapazitäten gekennzeichnet, die Deutschland je nach Ambitionslevel bis Ende 2028 aufbauen könnte. Dieser Appendix erklärt, wie diese berechnet wurden.

A.1 Berechnung von H100-Äquivalenten

Dieser Bericht gibt KI-Rechenkapazität primär in H100-Äquivalenten an. Dabei handelt es sich um ein gängiges Maß, um die Rechenleistung unterschiedlicher GPU-Typen miteinander zu vergleichen. Wir nutzen eine gängige Umrechnungsformel (Pilz, Sanders et al. 2025, S. 18). Diese teilt die Leistung eines beliebigen GPUs in der niedrigsten unterstützten Präzision (aus 32-, 16- oder 8-Bit) durch die Leistung eines Nvidia H100-GPUs in 8-Bit-Präzision. Die Umrechnung liefert einen groben Referenzwert, um die Kapazität von KI-Rechenzentren mit verschiedenen GPU-Typen zu vergleichen. Vor allem für jüngere Rechenzentren, die 2021 und später entstanden sind, ist dieser akkurat (Pilz et al. 2025).

A.2 Berechnung der nötigen IT-Anschlussleistung

Wir verwenden folgende Formel, um die für den Betrieb einer bestimmten Anzahl an GPUs benötigte Menge an elektrischer Leistung zu berechnen (Pilz, Sanders et al. 2025, S. 25):

Thermal Design Power × Anzahl der GPUs × Overhead-Faktor × Power Usage Effectiveness

Die Thermal Design Power (TDP) variiert nach GPU-Typ und liegt bei einem Nvidia H100 beispielsweise bei 700 Watt. Ein Overhead-Faktor berücksichtigt weitere Hardware neben GPUs und wird hier pauschal als 2,03 angenommen. Die Power Usage Effectiveness (PUE) ist eine Kennzahl für die Energieeffizienz eines Rechenzentrums und wird hier pauschal als 1,1 angenommen.

A.3 Globale GPU-Kapazität

Die globale GPU-Kapazität wächst jedes Jahr aufgrund von zwei Faktoren: 1) Chips werden stetig effizienter und 2) Chip-Hersteller wie TSMC (CPUs, GPUs) oder SK Hynix (Speicherchips) weiten ihre Produktion stetig aus. Fachleute nehmen an, dass die Chip-Effizienz ca. um den Faktor 1,35x/Jahr wächst und die Chip-Produktion ca. um den Faktor 1,65x/Jahr, was einem Gesamtwachstum von 2,25x/Jahr entspricht (Epoch AI 2025b, Kokotajlo et al. 2025). Aktuell beträgt die globale GPU-Kapazität etwa 10 Mio. H100-Äquivalente (Kokotajlo et al. 2025). Setzen sich die aktuellen Trends fort, ergeben sich 225 Mio. H100-Äquivalente bis Ende 2028. Diese Schätzung ist unsicher, weil sie lediglich auf öffentlich verfügbaren Informationen basiert und sich die zugrunde liegenden Hardware-Trends verlangsamen oder beschleunigen könnten. Sie spielt keine Rolle bei der folgenden Berechnung der GPU-/Stromkapazitäten für die einzelnen Strategien. Ihr Zweck ist allein zu illustrieren, welchen ungefähren Anteil Deutschland je nach Strategie an der globalen Rechenkapazität hätte (siehe Tabelle 2).

A.4 Benötigte GPU-Kapazität unter Strategie 1 & 2

Das britische Wissenschaftsministerium (DSIT) hat McKinsey beauftragt, die Nachfrage nach KI-Rechenleistung im Vereinigten Königreich bis 2035 auf Grundlage eines volkswirtschaftlichen Modells zu prognostizieren (DSIT 2025). Weil das McKinsey-Modell proprietär ist, sind nicht alle Details öffentlich einsehbar. Es fußt aber – wie auch Strategie 1 und 2 im hier vorliegenden Bericht – auf der Annahme, dass Rechenleistung vor allem für Inferenz sowie für das Training branchenspezifischer Modelle benötigt wird, nicht aber für das Training allgemeiner Spitzenmodelle (DSIT 2025, S. 6, S. 14).

Wir wenden das McKinsey-Modell auf den deutschen Kontext an und betrachten dazu das "Medium"-Szenario, in dem die Nachfrage nach KI-Rechenleistung jährlich um 27% wächst. Zu einem ähnlichen Ergebnis kommt eine Studie von Deloitte: Dort ergibt ein S-Kurven-Modell zur Diffusion neuer Technologien, dass sich die zur Deckung der Nachfrage die deutsche KI-Rechenkapazität von 1,6 GW im Jahr 2025 auf 4,8 GW im Jahr 2030 verdreifachen müsste (Lorentz et al. 2025). Das entspricht einem jährlichen Wachstum von 24,5%.

Wir verwenden hier die Wachstumsrate aus dem McKinsey-Modell. Als Ausgangswert nehmen wir basierend auf der Deloitte-Studie eine IT-Anschlussleistung von 1,6 GW im Jahr 2025 an. In diesem Fall würde der deutsche Bedarf bis Ende 2028 auf 4,2 GW steigen. Wir nehmen weiterhin an, dass dieser Gesamtbedarf durch ca. 1,9 Mio. B200-GPUs gedeckt wird. Nimmt man für diesen GPU eine Thermal Design Power von 1000 W an (Patel & Nishball 2024), ergibt sich unter Verwendung der in A.2 genannten Formel die o. g. Gesamtleistung von 4,2 GW.

Für die Berechnung der H100-Äquivalente ergibt sich ein Umrechnungsfaktor von 2,27: Die 8-Bit-Rechenleistung eines H100 entspricht 1.979 TFLOP/s; die eines B200 4.500 TFLOP/s (Patel & Nishball 2024). Der für Deutschland relevante Gesamtbedarf liegt in diesem Fall bei ca. 4,3 Mio. H100-Äquivalenten.

Strategie 1 deckt ca. 20% der deutschen Nachfrage (0,84 GW) durch eigene KI-Rechenzentren ab (statt über ausländische Cloud-Dienste). Dafür würde Deutschland bis Ende 2028 etwa 850.000 H100-Äquivalente benötigen. Dies entspricht zu diesem Zeitpunkt ca. 0,4% der globalen KI-Rechenkapazität (bei 225 Mio. H100-Äquivalenten insgesamt – siehe A.3).

Strategie 2 deckt ca. 80% der deutschen Nachfrage (3,36 GW) durch eigene KI-Rechenzentren ab. Dafür würde Deutschland bis Ende 2028 ca. 1,5 Mio. B200-GPUs – d. h. 3,4 Mio. H100-Äquivalente – benötigen. Dies entspricht zu diesem Zeitpunkt ca. 1,5% der globalen KI-Rechenkapazität.

A.5 Benötigte GPU-Kapazität unter Strategie 3

Grundlage für diese Strategie sind eigene Berechnungen basierend auf einem umfangreichen Datensatz des KI-Forschungsinstituts Epoch AI (Epoch AI 2025a). In einem ersten Schritt schätzen wir die benötigte Rechenleistung, um Ende 2028 ein Spitzenmodell mit allgemeiner Anwendung (*frontier model*) zu trainieren. Diese Rechenleistung (*training compute*) ist seit 2010 etwa um den Faktor 4x/Jahr gewachsen. Einige Fachleute halten es

nicht nur für technisch möglich, sondern auch für wahrscheinlich, dass sich dieser Trend bis zum Ende des Jahrzehnts fortsetzt (Sevilla et al. 2024, Epoch AI 2025c). Allerdings verlangt dieses enorm schnelle Wachstum von KI-Unternehmen immer größere Investitionen in Chips, KI-Rechenzentren und Energie. Allein OpenAI plant, in den nächsten vier Jahren 500 Mrd. Dollar in KI-Infrastruktur zu investieren (OpenAI 2025). Es ist ungewiss, ob diese Ambition wirtschaftlich tragfähig ist oder technische Hürden die Skalierung von Rechenleistung stärker ausbremsen als erwartet. Der KI-Experte Ryan Greenblatt geht z. B. davon aus, dass sich das Wachstum auf etwa 3,5x im Jahr verlangsamen wird (Greenblatt 2025). Sie könnte aber auch deutlich niedriger liegen. Bei einem hohen Ambitionslevel würde die Bundesregierung – wie auch die derzeit führenden KI-Unternehmen – davon ausgehen, dass sich der historische Wachstumstrend in den nächsten Jahren in etwa fortsetzt. Wir nehmen für Strategie 3 deshalb an, dass die jährliche Wachstumsrate der Trainings-Rechenleistung zwischen 3,5x und 4x liegt. Bei einer 16-Bit-Präzision im Training (Fist & Datta 2024) und einer GPU-Auslastung von 40%[12] impliziert das, dass Ende 2028 der Trainingslauf eines Spitzenmodells eine Rechenleistung von 1,5-2,3 × 10^{28} FLOP erfordert. Bei einer Trainingsdauer von 258 Tagen[13] und einer Leistung von 989 TFLOP/s eines einzelnen H100-GPUs entspricht das insgesamt 1,7-2,6 Mio. H100-Äquivalenten. Ein KI-Rechenzentrum dieser Größenordnung, in dem auf den neuen B200-GPUs trainiert wird, benötigt eine IT-Anschlussleistung von 1,7-2,5 GW.

In einem zweiten Schritt berechnen wir die benötigte Anzahl an GPUs für Inferenz. Weil sich der gesamtwirtschaftliche Bedarf an Inferenz-Rechenleistung schwer prognostizieren lässt – unsichere Variablen sind u. a. das Diffusionstempo von KI-Technologie und die zukünftige Bedeutung von “inference-time scaling” (Ma et al. 2025) –, nehmen wir an, dass der Inferenz-Anteil an der gesamten Rechenleistung 15%, 25% oder 50% betragen könnte.[14] GPUs, die für Inferenz verwendet werden, haben außerdem eine geringere Auslastung von etwa 15%.[15] Damit ergibt sich:

[12] Mit “Auslastung” ist der Anteil der theoretisch möglichen Rechenleistung eines GPUs gemeint, der tatsächlich für das Training genutzt wird. Faktoren wie eine langsame Datenübertragung zwischen Speicher- und Recheneinheiten führen zu einer Auslastung unter 100%. Wir nehmen für das Training von Spitzenmodellen eine GPU-Auslastung von 40% an (Fist & Datta 2024).

[13] Die Trainingsdauer von Spitzenmodellen ist in den vergangenen Jahrzehnten stetig gestiegen. Dieser Trend wird sich in den nächsten Jahren wahrscheinlich fortsetzen, weil längere Trainingsläufe Energie sparen (Fist & Datta 2024). Aufgrund schneller Iterationszyklen ist der Trend aber wahrscheinlich nach oben hin gedeckelt. Fachleute gehen davon aus, dass die Trainingsdauer bis 2027 ihr Maximum von 8,6 Monaten (258 Tage) erreicht (Emberson & Edelman 2025).
Fachleute gehen davon aus, dass dieses Wachstum nach oben hin gedeckelt

[14] Höhere Inferenz-Anteile erscheinen für diesen Zeitrahmen unplausibel, weil ein in Deutschland trainiertes Spitzenmodell im Vergleich zu den USA anfangs geringere Nutzerzahlen hätte. McKinsey geht in einer Auftragsarbeit für das britische Wissenschaftsministerium (DSIT) davon aus, dass der Inferenz-Anteil im Vereinigten Königreich zwischen 2023 und 2035 von 16% auf 57-71% steigen wird.

[15] Diese Schätzung basiert auf Gesprächen mit Hardware-Experten und deckt sich in etwa mit dem Wert von 17%, den Erdil (2025) für das Modell Llama 3 70B berechnet.

H100-Äquivalente für Frontier-Training	Inferenz-Anteil	H100-Äquivalente für Inferenz	H100-Äquivalente insgesamt
1,7 bis 2,6m	15%	0,8 bis 1,2m	2,5 bis 3,8m
	25%	1,5 bis 2,3m	3,3 bis 4,9m
	50%	4,6 bis 6,9m	6,3 bis 9,5m

Tabelle A: Unter Strategie 3 hat Deutschland einen GPU-Bedarf von 2,5 bis 9,5 Mio. H100-Äquivalenten, abhängig von verschiedenen technischen Annahmen über das Wachstum von Trainings-Rechenleistung und dem relativen Anteil von Training und Inferenz. Bei der Summenbildung ergeben sich rundungsbedingte Abweichungen.

Der GPU-Wert für Strategie 3 – 6 Mio. H100-Äquivalente – entspricht dem Mittelwert aus den in Tabelle A genannten Mindest- bzw. Höchstwerten von 2,5 bzw. 9,5 Mio. H100-Äquivalenten. 6 Mio. H100-Äquivalente machen Ende 2028 ca. 2,7% der globalen KI-Rechenkapazität aus. Dies liegt unter dem deutschen Anteil am globalen BIP (nominal) von derzeit ca. 4,2% (IMF 2025).

Unter der Annahme, dass für Training und Inferenz ca. 2,6 Mio. B200-GPUs genutzt werden, ergibt sich daraus eine IT-Anschlussleistung von insgesamt 5,9 GW (2,5 bis 9,3 GW).

Literaturverzeichnis

Berg & Ho (2025). “After the ChatGPT Moment: Measuring AI’s Adoption”, Epoch AI Gradient Updates, https://epochai.substack.com/p/after-the-chatgpt-moment-measuring

Bick et al. (2024). “The Rapid Adoption of Generative AI”, National Bureau of Economic Research, Cambridge, MA, https://www.nber.org/system/files/working_papers/w32966/w32966.pdf

Bitkom (2025). “Deutschland zum KI-Hotspot machen: 10 Empfehlungen für eine erfolgreiche KI-Zukunft”, Bitkom e. V. https://www.bitkom.org/sites/main/files/2025-05/bitkom-publikation-deutschland-zum-ki-hotspot-machen.pdf

BMWK (2025). “Standund und Entwicklung des Rechenzentrumsstandorts Deutschland”, Gutachten im Auftrag des Bundesministeriums für Wirtschaft und Klimaschutz, https://www.bundeswirtschaftsministerium.de/Redaktion/DE/Publikationen/Technologie/stand-und-entwicklung-des-rechenzentrumsstandorts-deutschland.pdf?__blob=publicationFile&v=10

Chatterji et al. (2025). “How People Use ChatGPT“, OpenAI, https://cdn.openai.com/pdf/a253471f-8260-40c6-a2cc-aa93fe9f142e/economic-research-chatgpt-usage-paper.pdf

Datta & First (2025). “Compute in America: A Policy Playbook”, Institute for Progress, https://ifp.org/special-compute-zones/

DeepSeek-AI (2024). “DeepSeek-V3 Technical Report”, arXiv, https://arxiv.org/abs/2412.19437

DSIT (2025). “Compute Evidence Annex”, Research Report 2025/025, Department for Science, Innovation & Technology, https://assets.publishing.service.gov.uk/media/687f74f4fdc190fb6b846868/compute-evidence-annex-final.pdf

eco (2023). “Einordnung des BMWK-Konzeptes für einen Industriestrompreis in das EU-Beihilferecht”, eco - Verband der Internetwirtschaft e.V., https://www.eco.de/wp-content/uploads/2023/09/20230810_eco-hintergrundpapier_industriestrompreis.pdf

eco (2025). “eco Allianz fordert Industriestrompreis für Rechenzentren”, eco - Verband der Internetwirtschaft e.V., https://www.eco.de/presse/eco-allianz-fordert-industriestrompreis-fuer-rechenzentren/

Emberson & Edelman (2025). “Frontier training runs will likely stop getting longer by around 2027”, Epoch AI, https://epoch.ai/data-insights/longest-training-run

Epoch AI (2025a). “Data on AI Models”, Online-Zugriff am 03.09.2025, https://epoch.ai/data/ai-models

Epoch AI (2025b). “Machine Learning Trends”, Online-Zugriff am 03.09.2025, https://epoch.ai/trends#hardware

Epoch AI (2025c). “AI in 2030: Extrapolating current trends”, Online-Zugriff am 18.09.2025, https://epoch.ai/files/AI_2030.pdf

Erdil (2025). “Inference economics of language models”, arXiv, https://arxiv.org/abs/2506.04645

EuroHPC (2025). “Call for expression of interest in AI Gigafactories (AIGFs)”, European High Performance Computing Joint Undertaking, https://www.eurohpc-ju.europa.eu/document/download/47492db7-592e-4ad8-b672-9c822f94afa0_en?filename=AI%20GIGAFACTORIES%20CONSULTATION.pdf

European Commission (2025a). “AI Continent Action Plan”, Online-Zugriff am 03.09.2025, https://digital-strategy.ec.europa.eu/en/factpages/ai-continent-action-plan

European Commission (2025b). “EU Startup and Scaleup Strategy”, Online-Zugriff am 03.09.2025, https://research-and-innovation.ec.europa.eu/strategy/strategy-research-and-innovation/jobs-and-economy/eu-startup-and-scaleup-strategy_en

Eurostat (2024). “Use of artificial intelligence in enterprises”, European Commission, Online-Zugriff am 03.09.2025, https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Use_of_artificial_intelligence_in_enterprises

Export Compliance Daily (2025). “New AI Diffusion Rule Will Let Allies Buy US Chips With Conditions, Commerce Head Says”, https://exportcompliancedaily.com/article/2025/06/05/new-ai-diffusion-rule-will-let-allies-buy-us-chips-with-conditions-commerce-head-says-2506040042?BC=bc_685e6c5ba1f0a

Fist & Datta (2024). “How to Build the Future of AI in the United States”, Institute for Progress, https://ifp.org/future-of-ai-compute

FAZ (2025). “Das elfköpfige Start-up, das eine AI Gigafactory bauen will”, Frankfurter Allgemeine Zeitung, https://www.faz.net/pro/digitalwirtschaft/kuenstliche-intelligenz/lyceum-start-up-aus-berlin-will-ai-gigafactory-bauen-110586207.html

FZ Jülich (2025). “Supercomputer JUPITER erreicht Rekord-Rechenleistung in Europa”, Forschungszentrum Jülich, https://www.fz-juelich.de/de/aktuelles/news/pressemitteilungen/2025/supercomputer-jupiter-erreicht-rekord-rechenleistung-in-europa

German Datacenter Association (2024). “GDA-Positionspapier zum Konsultationsverfahren BNetzA über ein Verfahren zur Zuteilung von Entnahmeleistungen aus Netzebenen oberhalb der Niederspannung”, German Datacenter Association e. V., https://www.germandatacenters.com/fileadmin/documents/publications/2024-12-30_GDA_Position_Konsultation_BK_6-24-245.pdf

Greenblatt (2025). “What's going on with AI progress and trends? (As of 5/2025)”, Redwood Research, https://blog.redwoodresearch.org/p/whats-going-on-with-ai-progress-and

Harris & Harris (2025). “America’s Superintelligence Project”, Gladstone AI, https://superintelligence.gladstone.ai/

Handelsblatt (2025). “Das sind die deutschen Bewerber für eine AI Gigafactory”, https://www.handelsblatt.com/technik/ki/ki-rechenzentren-das-sind-die-deutschen-bewerber-fuer-eine-ai-gigafactory-01/100137764.html

Hess (2025). “Eine Gigafactory reicht vorerst: Besonnen statt groß denken!”, Süddeutsche Zeitung Dossier, https://www.sz-dossier.de/gastbeitraege/2025-05-02-julia-hess-eine-gigafactory-reicht-vorerst-besonnen-statt-gross-denken-8e094bdb

HLRS (2023). “Exascale-Supercomputing kommt nach Stuttgart”, Höchstleistungsrechenzentrum Stuttgart, https://www.hlrs.de/de/news/detail/exascale-supercomputing-kommt-nach-stuttgart

HLRS (2024). “HammerHAI wird eine AI Factory für Wissenschaft und Industrie aufbauen”, Höchstleistungsrechenzentrum Stuttgart, https://www.hlrs.de/de/presse/detail/hammerhai-wird-eine-ai-factory-fuer-wissenschaft-und-industrie-aufbauen

IEA (2025). “Energy and AI”, World Energy Outlook Special Report, International Energy Agency, https://iea.blob.core.windows.net/assets/40a4db21-2225-42f0-8a07-addcc2ea86b3/EnergyandAI.pdf

IMF (2025). “World Economic Outlook Database”, International Monetary Fund, Online-Zugriff am 03.09.2025, https://www.imf.org/en/Publications/WEO/weo-database/2025/April/weo-report

Juijn et al. (2025). “Tripling the EU’s data centre stock with special AI compute zones”, Center for Future Generations, https://cfg.eu/special-compute-zones-submission/

KI Bundesverband (2024). “Für ein starkes KI-Deutschland: Impulspapier zur Bundestagswahl 2025”, Bundesverband der Unternehmen der Künstlichen Intelligenz in Deutschland e.V., https://ki-verband.de/wp-content/uploads/2025/05/Impulspapier_Bundestagswahl2025_KI-Bundesverband_2024.12-1.pdf

Kokotajlo et al. (2025). “AI 2027”, AI Futures Project, https://ai-2027.com/

Lorentz et al. (2025). “KI-Infrastruktur: Wie Deutschland im globalen KI-Rennen aufholen kann”, Deloitte, https://www.deloitte.com/content/dam/assets-zone2/de/de/docs/issues/sustainability-climate/2025/Deloitte-KI-Infrastruktur-Studie.pdf

LRZ (2025). “Blue Lion mit Vera Rubin-Architektur”, Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften, https://www.lrz.de/news/detail/blue-lion-mit-vera-rubin-architektur

Ma et al. (2025). “Inference-Time Scaling for Diffusion Models beyond Scaling Denoising Steps”, arXiv, https://arxiv.org/abs/2501.09732

METR (2025). “How Does Time Horizon Vary Across Domains?”, Model Evaluation & Threat Research, https://metr.org/blog/2025-07-14-how-does-time-horizon-vary-across-domains/

Nevo et al. (2024). “Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models”, RAND Corporation, https://www.rand.org/content/dam/rand/pubs/research_reports/RRA2800/RRA2849-1/RAND_RRA2849-1.pdf

Norris et al. (2025). “Rethinking Load Growth: Assessing the Potential for Integration of Large Flexible Loads in US Power Systems”, Nicholas Institute for Energy, Environment & Sustainability, Duke University, https://hdl.handle.net/10161/32077

NVIDIA (2025a). “NVIDIA Builds World’s First Industrial AI Cloud to Advance European Manufacturing”, https://nvidianews.nvidia.com/news/nvidia-builds-worlds-first-industrial-ai-cloud-to-advance-european-manufacturing

NVIDIA (2025b). “NVIDIA and United Kingdom Build Nation’s AI Infrastructure and Ecosystem to Fuel Innovation, Economic Growth and Jobs”, https://nvidianews.nvidia.com/news/nvidia-and-united-kingdom-build-nations-ai-infrastructure-and-ecosystem-to-fuel-innovation-economic-growth-and-jobs

OpenAI (2025). “Announcing The Stargate Project”, https://openai.com/index/announcing-the-stargate-project/

Patel & Nishball (2024). “Nvidia Blackwell Perf TCO Analysis – B100 vs B200 vs GB200 NVL72”, SemiAnalysis, https://semianalysis.com/2024/04/10/nvidia-blackwell-perf-tco-analysis/

Patel et al. (2025a). “DeepSeek Debates: Chinese Leadership On Cost, True Training Cost, Closed Model Margin Impacts”, SemiAnalysis, https://semianalysis.com/2025/01/31/deepseek-debates/

Patel et al. (2025b). “AI Arrives In The Middle East: US Strikes A Deal with UAE and KSA”, SemiAnalysis, https://semianalysis.com/2025/05/16/ai-arrives-in-the-middle-east-us-strikes-a-deal-with-uae-and-ksa/

Pélissie du Rausas et al. (2011). “Internet matters: The Net's sweeping impact on growth, jobs, and prosperity”, McKinsey Global Institute, https://www.mckinsey.com/~/media/mckinsey/industries/technology%20media%20and%20telecommunications/high%20tech/our%20insights/internet%20matters/mgi_internet_matters_exec_summary.pdf

Pilz, Heim & Brown (2023). “Increased Compute Efficiency and the Diffusion of AI Capabilities”, arXiv, https://arxiv.org/abs/2311.15377

Pilz et al. (2025). “Data on GPU Clusters”, Epoch AI, Online-Zugriff am 03.09.2025, https://epoch.ai/data/gpu-clusters

Pilz, Sanders et al. (2025). “Trends in AI Supercomputers”, arXiv, https://arxiv.org/abs/2504.16026

Sevilla et al. (2024). “Can AI Scaling Continue Through 2030?”, Epoch AI, https://epoch.ai/blog/can-ai-scaling-continue-through-2030

Telekom (2025). “KI-Turbo für Gigafactories: Telekom bildet europäische KI-Infrastruktur mit NVIDIA für Industrie”, Deutsche Telekom, https://www.telekom.com/de/medien/medieninformationen/detail/ki-turbo-nvidia-und-deutsche-telekom-1093524

The Information (2025). “DeepSeek’s Progress Stalled by U.S. Export Controls”, https://www.theinformation.com/articles/deepseeks-progress-stalled-u-s-export-controls

The Next Platform (2023). “LRZ Adopts Nvidia Engines For €250 Million “Blue Lion” Supercomputer In 2027”, https://www.nextplatform.com/2024/12/16/lrz-adopts-nvidia-engines-for-e250-million-blue-lion-supercomputer-in-2027/

Welt (2025). “‘Werte und Wissen widerspiegeln’ – Digitalminister fordert europäische KI-Modelle”, https://www.welt.de/wirtschaft/article256390626/Karsten-Wildberger-Digitalminister-fordert-europaeische-KI-Modelle.html

Wiseman, McClements & Horsley (2024). “Getting AI datacentres in the UK”, Inference Magazine, https://inferencemagazine.substack.com/p/getting-ai-datacentres-in-the-uk

Y Combinator (2025). “Elon Musk: Digital Superintelligence, Multiplanetary Life, How to Be Useful”, YouTube-Video, Online-Zugriff am 03.09.2025, https://www.youtube.com/watch?v=cFIlta1GkiE&t=1894s

You (2025). “How far can reasoning models scale?”, Epoch AI Gradient Updates, https://epoch.ai/gradient-updates/how-far-can-reasoning-models-scale

Über diesen Bericht

Autoren: Philip Fox, Daniel Privitera, Monika Schnitzer

Kontakt: philip@kira.eu

Zitierempfehlung: Fox, P., Privitera, D. & Schnitzer, M. (2025). “KI-Rechenzentren in Deutschland: Aktuelle Kapazität, zukünftiger Bedarf”, KIRA Center, Berlin.