An assessment of data center infrastructure's role in AI governance

By Konstantin Pilz
July 2023

This comment relates insights on data centers from my initial report to their role in AI development and governance. It focuses on aspects of data center infrastructure, an abstraction that includes both how hardware is set up within data centers and the supporting systems it requires.

The first part presents figures and models for understanding how data center infrastructure impacts AI development, building upon the compute governance frameworks, as presented by Heim, 2023. The second part explores how data centers can be an opportunity for AI governance.

Readers should be familiar with a) the motivation for the governance of advanced AI systems, e.g., as outlined by Dafoe, 2020 and b) the basics of the data center industry, e.g., having read the summary of my report Compute at Scale - A broad investigation into data centers.

Executive summary

Scope

Data center infrastructure concerns **a) the way hardware is set up in a data center** and **b) the supporting infrastructure required to operate that hardware**. This abstraction allows focusing on the physical aspects of data centers rather than their role of providing virtual compute resources. For an intuitive understanding of this abstraction, see Figures 1, 2, 3a, and 4. [⇩]

Data center infrastructure's role in AI development

AI training and deployment of major AI systems such as ChatGPT, Bard, or Claude are exclusively run on AI compute clusters that consist of thousands of specialized AI accelerators with high-bandwidth connections. These **clusters require dedicated supporting infrastructure** provided by data centers, including power distribution, cooling, security, and connectivity equipment, each with backup components for uninterrupted operation. Yet, most data centers serve ordinary purposes, such as providing various internet services, and less than 10% of the 500 large (>10 MW[1]) data centers globally likely host AI compute clusters. [⇩]

Discussion of the compute supply chain often focuses on semiconductor production, but data centers equally present an integral part, constituting the link to the compute end-user. Building and operating data center infrastructure requires a variety of inputs, such as land, electrical equipment, cooling systems, water, and dozens of MW of power. The supply chain thus involves multiple companies in data center construction, infrastructure operation, and hardware operation. Big Tech firms such as Google, Microsoft, and AWS already control considerable parts of their supply chains and thus increasingly dominate the data center industry. [⇩]

Supporting infrastructure required to host a compute cluster is significant in size and operational requirements, with large data centers being more economical due to economies of scale. Constructing a large data center (20 MW) currently costs $100M to $200M just for the supporting infrastructure[2]. With an increasing demand for larger compute clusters, data center size, and construction costs could increase by an order of magnitude in the next five years. Yet, although it requires extensive planning, there are dozens of companies in the world able to construct large data centers. Therefore, if funds are available, new actors can simply contract a construction company, buy a fully fitted-out "turnkey" data center, or rent capacity in a colocation facility. [⇩]

AI compute clusters present considerable barriers to entry in terms of investment, knowledge/talent, and partnerships. The hardware components needed for a state-of-the-art AI compute cluster currently cost hundreds of millions, which could reach up to billions of dollars if clusters continue to grow in size. Furthermore, designing

[1] For reference, this is the equivalent of the energy consumption of about 10,000 US households.
[2] I.e., excluding the cost to buy and install the hardware.

high-bandwidth interconnected cluster require specialized skills. Additionally, waiting times for cutting-edge AI hardware can be considerable, with large cloud providers having special partnerships with chip suppliers for priority access and discounts. [⇩]

Implications for AI governance

Data center infrastructure is a source of information on the (AI) compute capacity of various actors. It is also an input for forecasting future developments, such as fundamental limits to scaling AI compute clusters or which inputs would likely limit growth in transformative scenarios. [⇩]

Data centers are an opportunity for regulation:

- **Verifying compute ownership**: To regulate AI compute usage, it is likely necessary to reliably detect hardware ownership. While tracking AI accelerator sales appears like the most promising method of doing so, an additional approach could require data center infrastructure providers to report the hardware they host, which could be verified via routine inspections. [⇩]
- **Verifying compute usage:** In the absence of chip-based mechanisms, various measures, such as network traffic and utilization, could be tracked at the data center level to help verify claims compute owners make about their usage. Further research is required to find out to what extent this is possible and which measures are most informative. [⇩]
- **Data center security**: Data centers train and host AI models that can be misused or might be inherently dangerous. They are thus responsible for preventing unauthorized access to computational infrastructure and the model parameters and training-relevant data contained within. This involves physical as well as cyber-security. [⇩]
- **Containment of dangerous models**: Emergency shutdown mechanisms could be implemented on several levels of the infrastructure stack to potentially quickly turn off dangerous models. Air-gapping the training of new models until they have passed safety auditing could further limit the impacts of misaligned models and serve as an extra protection against theft of model weights. [⇩]

Considerable challenges remain for policy work:

- **Information availability**: The data center industry is relatively opaque, making it difficult to collect information and scope regulation well. I caution against advocating any data center-level regulation before more thorough research is conducted. [⇩]
- **Governments' willingness to regulate**: The current Overton window does not include regulation of data centers in the context of risks from AI. Yet, as this could change quickly, it is advisable to proactively develop and prepare policy proposals leveraging the discussed unique levers for data centers. [⇩]
- **Regulatory arbitrage**: Since data centers exist all over the world, future regulation could be avoided by setting up AI compute clusters in other jurisdictions. International coordination may be required for successful monitoring and access restrictions. [⇩]

- **Technical feasibility**: It is unclear to what extent the proposed mechanisms, such as logging of hardware data during training or emergency shutdown systems, can be implemented on a technical level, and more research is required. [⇩]

Conclusion

Data centers present several opportunities for limiting risks from advanced AI systems and contributing to their beneficial governance. Further work is needed to assess technical feasibility and develop roadmaps for policy engagement. [⇩]

Scope (What is data center infrastructure?)

Data center infrastructure, a commonly used industry term, refers to a) the arrangement of hardware within a data center and b) the supporting infrastructure required to host it, including power supply, cooling, connectivity infrastructure, various backup components, and security measures (see Figures 1, 2, 3, and 4). *Data center infrastructure* thus focuses on the physical aspects of data centers and allows disentangling them from the virtual aspects, contained in the abstraction of *compute provision*.

As defined by Digital Realty, 2021; "If it's physically inside or forms the physical structure of a data centre facility, then it's data centre infrastructure."

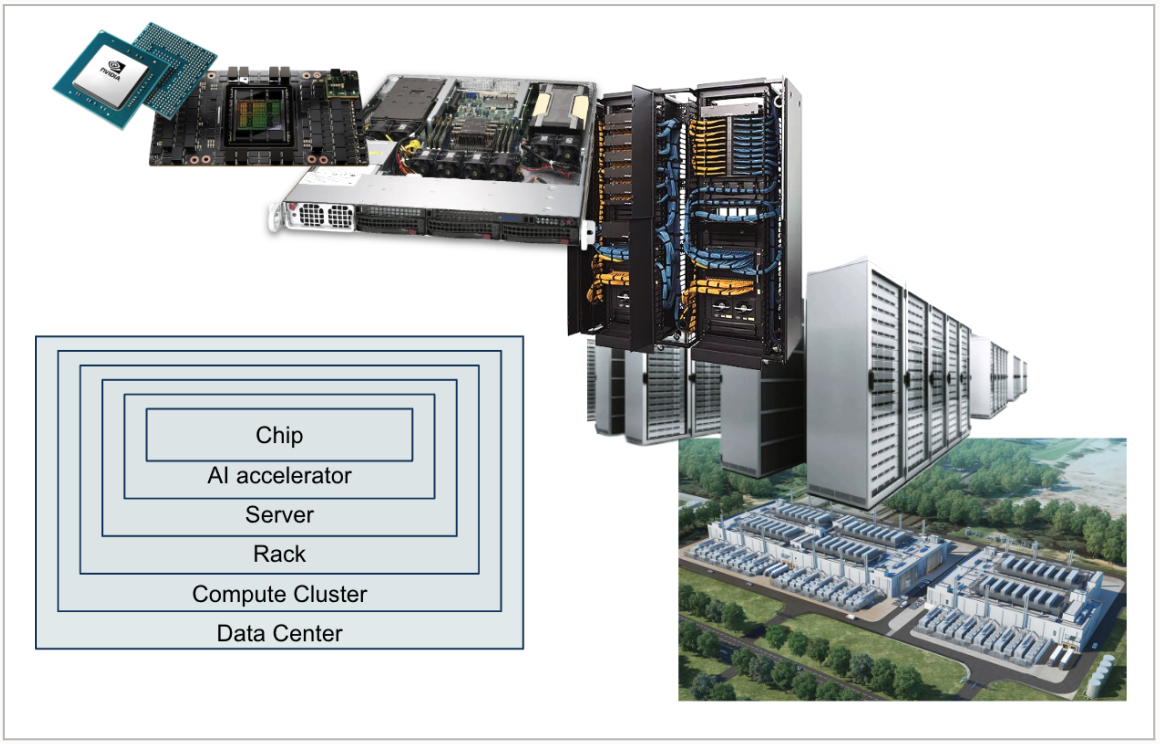

Figure 1 - Data center infrastructure encompasses a) the ***different layers of hardware*** *in the data center, how they are set up, and interact with the supporting infrastructure. The data center (bottom) contains one to several compute clusters, which consist of racks containing servers. Among other parts, the server contains an AI accelerator that has a semiconductor chip running the computations.*

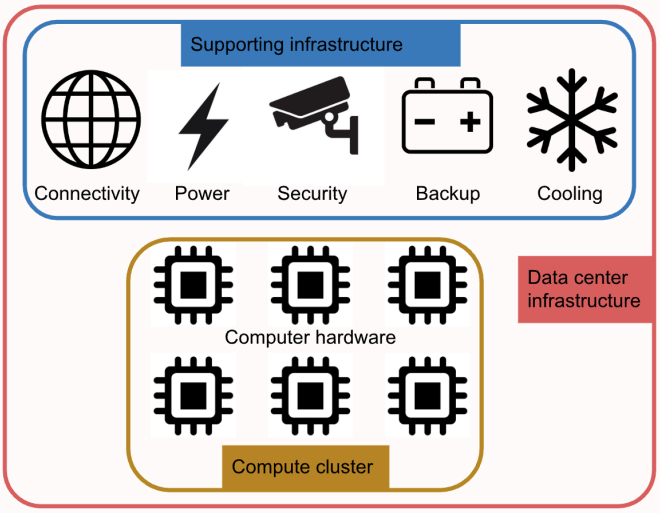

*Figure 2 - Data center infrastructure encompasses b) the **supporting infrastructure** enabling the operation of the hardware, including connectivity, power distribution, security, backup components, and cooling systems. These support the compute cluster. Together they make up the data center infrastructure.*

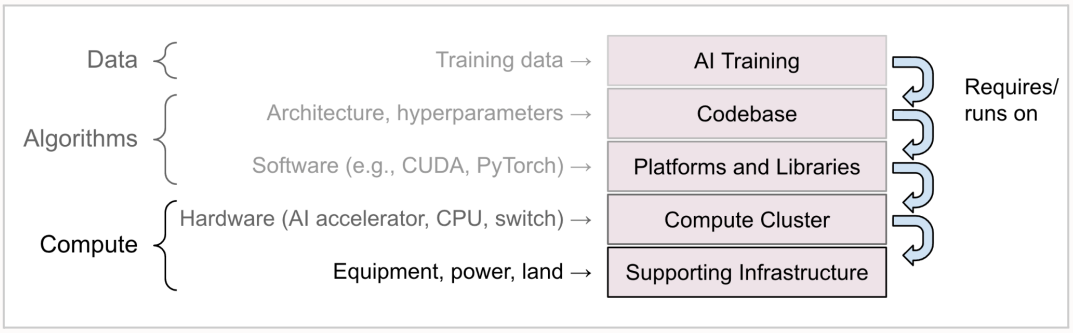

Figure 3 - Simplified AI tech stack. Data center infrastructure encompasses the two bottom layers of the AI infrastructure stack. i) Supporting infrastructure enables the operation of a compute cluster. ii) The compute cluster runs the different layers of software, consisting of iii) platforms and libraries that allow the execution of AI-specific applications and iv a codebase based on these platforms and libraries that specify the model and training parameters. The codebase then enables v) AI training, requiring data as additional input.

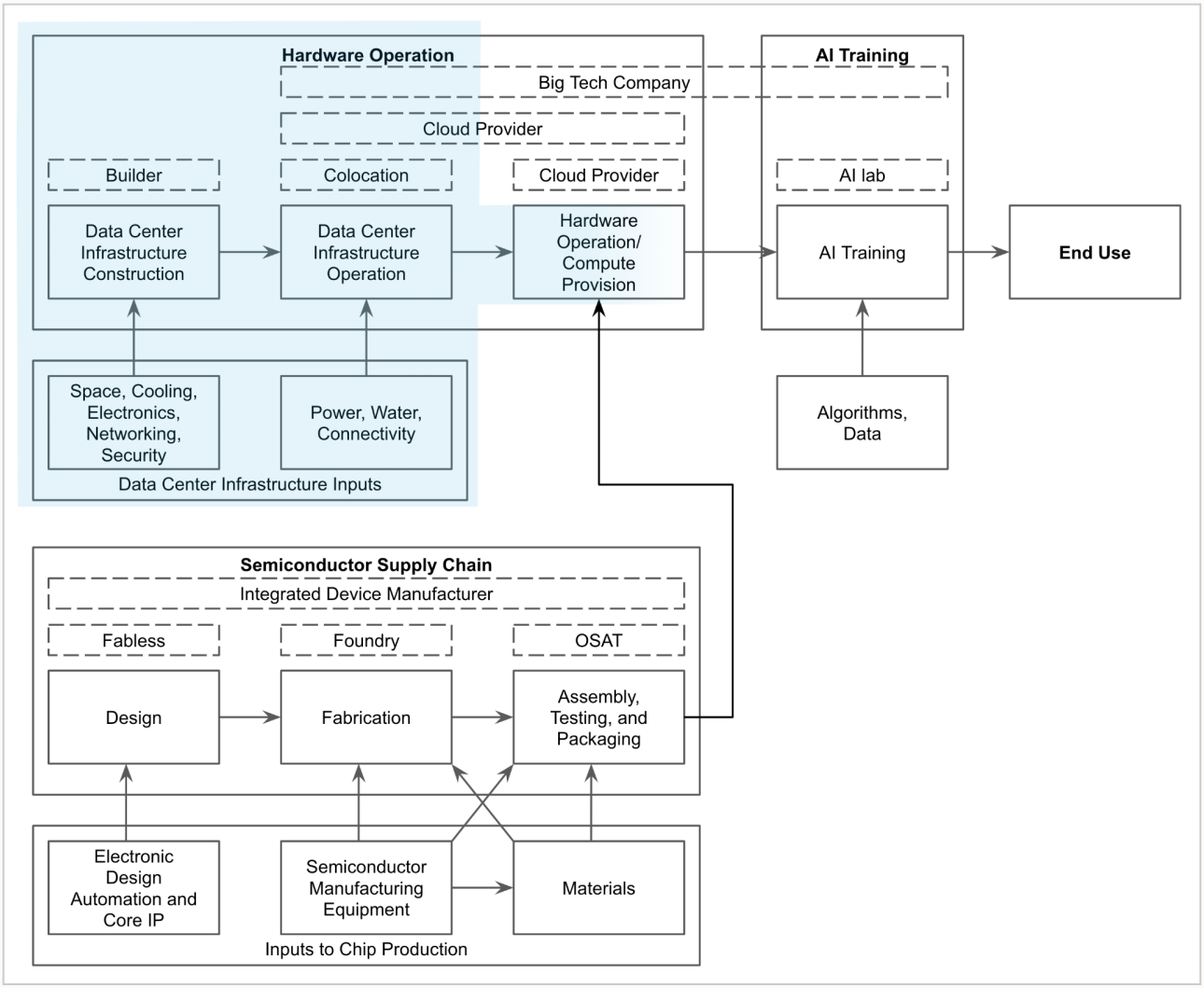

Figure 4a - In the compute supply chain, data center infrastructure is concerned with ***infrastructure construction*** *(including all the equipment needed) and* ***infrastructure operation*** *(including the operational inputs such as power, water, and connectivity). It also includes the operation of the hardware in terms of how it is designed and installed and how it interacts with the supporting infrastructure. Important actors are construction companies, equipment suppliers, infrastructure owners/ providers, and hardware installers and operators. See Figure 4b for a more thorough explanation.*

Data center infrastructure's role in AI development

AI training and deployment at scale are exclusively executed from AI compute clusters

While most compute today relies on CPUs, efficient AI training demands specialized AI accelerators (Khan, 2020) set up in clusters with high-bandwidth connections.[3] Training the most notable models, such as GPT-4, PaLM, or Claude, required thousands of AI accelerators in custom-designed clusters, sometimes called AI supercomputers (Microsoft, 2020). Recent trends have shown that big AI labs continuously increase the amount of compute they use to train AI systems (Sevilla et al., 2022). This suggests that training state-of-the-art models will increasingly demand larger AI compute clusters.[4]

Model inference, which is required for the deployment of AI models, uses a fraction of the compute needed for training, making the deployment process significantly less compute-intensive. However, deploying an AI model economically on a large scale, such as rolling out ChatGPT to more than 100 million customers (Reuters, 2023), similarly requires an AI compute cluster.[5]

AI compute clusters require supporting infrastructure

Deploying hardware efficiently at a large scale necessitates a dedicated infrastructure that includes power, cooling, connectivity, security, and backup equipment for each. Consequently, a substantial AI compute cluster can only be established within a data center.

One way of characterizing the role data center plays in AI development is the AI infrastructure stack (Figure 3) (Business Bliss Consultants FZE, 2018). The stack shows how different layers of hardware and software are built on top of one another. Data centers constitute the foundation of the stack and, therefore, a necessity for all layers above.

Today's largest AI compute clusters consume up to several dozen MW of power and cooling—a demand that can comfortably be met by today's largest data centers.[6] However, if AI hardware continues to increase in power consumption and AI compute clusters continue to grow in the number of AI accelerators, it may get increasingly difficult to provide the appropriate power and cooling infrastructure to host such AI compute clusters.

---

[3] Although it is, in principle, possible to train such models on CPU clusters (e.g., HPC clusters primarily used for scientific simulations.) or geographically decentralized AI accelerators, this involves considerable costs in terms of time, effort, and money, and there are no significant precedents for either. However, to what extent AI training will rely on concentrated compute clusters remains an open question and will be a crucial consideration for evaluating how much leverage regulating data centers can have on AI development.

[4] However, this trend can only last for a few more years, see Heim, 2023.

[5] Inference, just like training, can be highly parallelized. The efficiency increases with the number of queries per time.

[6] Based on internal estimates.

Most data centers do not contain AI compute clusters

The primary purpose of data centers is to host the digital infrastructure powering the wide variety of internet services of our modern economy. (See "What are data centers?" in the initial report.) Most data centers, thus, don't contain significant AI compute clusters.[7]

Data centers are the final part of the compute supply chain

Seen as a product, compute resources have a long supply chain. Its first part is equivalent to the semiconductor supply chain (Khan, 2021) and involves design, fabrication, and packaging. Its final product is a functional piece of computer hardware, such as a CPU or an AI accelerator.

Besides semiconductor production, the compute supply chain involves an extra stage: deploying the hardware and enabling its efficient usage (see Figure 1). As outlined above, doing so at an industrial level necessitates dedicated infrastructure in terms of power supply, cooling, internet connectivity, security, and backup components. Data centers provide this infrastructure, thereby serving as an additional part of the compute supply chain.

Data centers add complexity to the compute supply chain (Figure 4b) as their construction and operation require a wide range of inputs. Furthermore, a variety of companies may be involved in data center construction, infrastructure operation, and hardware operation, adding further actors. However, cloud providers such as Google Cloud or AWS already control several links of their supply chain, e.g., by owning their data center infrastructure themselves, designing their own hardware, and having in-house AI labs.

[7] The initial report estimates there are about 500 data centers with a power capacity of more than 10 MW globally. My informed guess is that less than 10% of them contain clusters of more than 5,000 AI accelerators (such as A100, H100, TPUv4 or hardware of similar capabilities). Further, of all such data centers, only a fraction features sufficiently advanced networking infrastructure to support highly parallel training and inference.

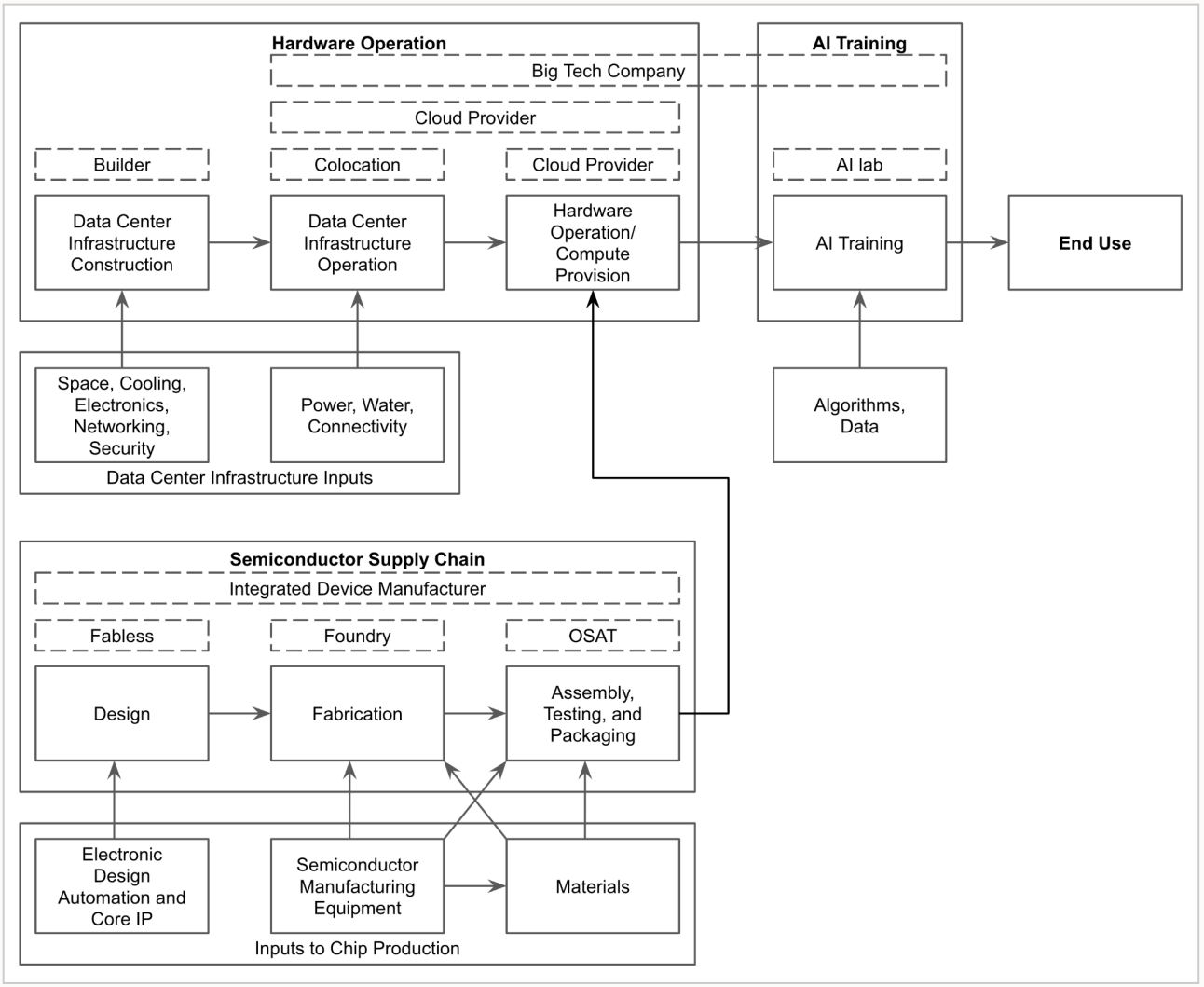

*Figure 4b - The compute supply chain. Hardware is produced in the semiconductor supply chain, as described by Khan, 2021. The hardware is then hosted in a data center. This requires supporting infrastructure that first needs to be **constructed**, requiring a variety of inputs such as land, electrical equipment, and cooling systems. Data center infrastructure also needs to be **operated**, requiring considerable amounts of power and water. In some cases, the infrastructure operator is a different entity from the compute owner. However, large compute providers typically own and operate their data center infrastructure in-house. The compute is **provided** to AI labs and used for **AI training**, additionally requiring algorithms and data. Big Tech Companies may both operate and use AI compute in-house. The AI model is finally integrated into products and supplied to the end user.*

Supporting infrastructure demands considerable resources

The supporting infrastructure for AI compute clusters is significant in terms of land use, power use, equipment required, and construction difficulty, which creates a financial barrier to entry for new actors in the market (See Figure 2 in the initial report). To illustrate, the infrastructure for a large data center that can support 20 MW of hardware costs between $100M and $200M to construct (see construction cost in initial report). This figure could increase by an order of magnitude as demand for ever larger AI compute clusters continues over the next five years.

While there are considerable planning requirements to construct a large data center, dozens of companies around the world can do so. Thus, given the required investment, new actors can simply contract data center construction companies or buy "turnkey" data centers that are fully fitted out. Furthermore, colocation providers[8] have specialized in providing and operating supporting infrastructure, granting easy access for new actors.[9]

AI compute clusters present additional barriers to entry

Meanwhile, AI compute clusters present further barriers to entry in terms of investment, tacit knowledge/talent, and partnerships.

Nvidia's A100, the currently most widely used AI accelerator, costs about $10,000[10] (CNBC, 2022). The cluster that trained GPT-4 likely consisted[11] of tens of thousands of these chips and additionally required expensive hardware such as CPUs and networking equipment. This means compute clusters can easily cost more than $100M.[12] As AI companies conduct larger training runs and build highly parallelized clusters for inference, hardware for the largest compute clusters could cost billions in a matter of five years. Besides financial resources, designing and implementing an AI supercomputer requires latent knowledge, adding another layer of complexity and a barrier to entry for new players in the market.

Furthermore, large cloud providers have special partnerships with semiconductor suppliers for priority access to hardware.[13] Lacking these partnerships presents a considerable handicap for smaller companies, leading to an increasing concentration of AI compute clusters in the hands of a small number of large cloud providers (*Statista*, 2023).

[8] Colocation providers host the hardware of their customers and operate the supporting infrastructure entirely.
[9] For more, see the section "Data center ownership" in the initial report.
[10] Other sources claim considerably higher prices (*Tom's Hardware*, 2023).
[11] According to an unconfirmed leak of a Morgan Stanley document.
[12] As Sam Altman recently confirmed, according to WIRED, 2023.
[13] E.g., NVIDIA has developed a special partnership program (NVIDIA, 2023). The largest partners, like AWS, get early access to their hardware and collaborate to set up efficient clusters (NVIDIA, 2023).

Implications for AI governance

Source of information

Identifying relevant actors

Data center capacity can be seen as a coarse proxy for (potential) AI capabilities.[14] Since public sources on AI compute clusters are rare, this proxy can be used to answer a range of questions on compute capacities, such as:

- What is the data center capacity of certain countries (E.g., US, China, UK, etc.)?
- Which companies own and use the highest number of data centers?
- What data center demands do militaries have, and how are they supplied?
- What are the barriers to entry for AI development? I.e., how difficult is it for an emerging actor to set up a cutting-edge AI supercomputer? (My current impression is that there are significant barriers, see Supporting infrastructure is significant in size.)

Forecasting compute developments

The growth potential of the data center industry is a determining factor in the rate at which AI can be integrated into the economy and thus have transformative effects. Assessing fundamental and economic limits for scaling data center or AI compute cluster level is vital, as it helps determine this potential.

Further, forecasting trends in concentration and ownership indicates opportunities for regulation.

Questions relevant to forecasting include:

- How quickly can data center capacity increase in high-demand scenarios? Which inputs could become bottlenecks if demand was high?
- What proportion of data centers could be easily repurposed for AI compute clusters conditional on exceptional demand?
- If AI compute clusters continue to grow, are there physical or infrastructural limitations on how large data centers can be to accommodate them?
- Will data center energy consumption increase, and could it be a limit to future growth?

[14] Data on the relationship between companys' data center capacity and their AI compute capacity is limited. Yet, e.g., the three largest cloud providers, AWS, Google Cloud, and Microsoft Azure, also each have large AI compute capacity.

Data centers as an opportunity for regulation

Verifying hardware ownership

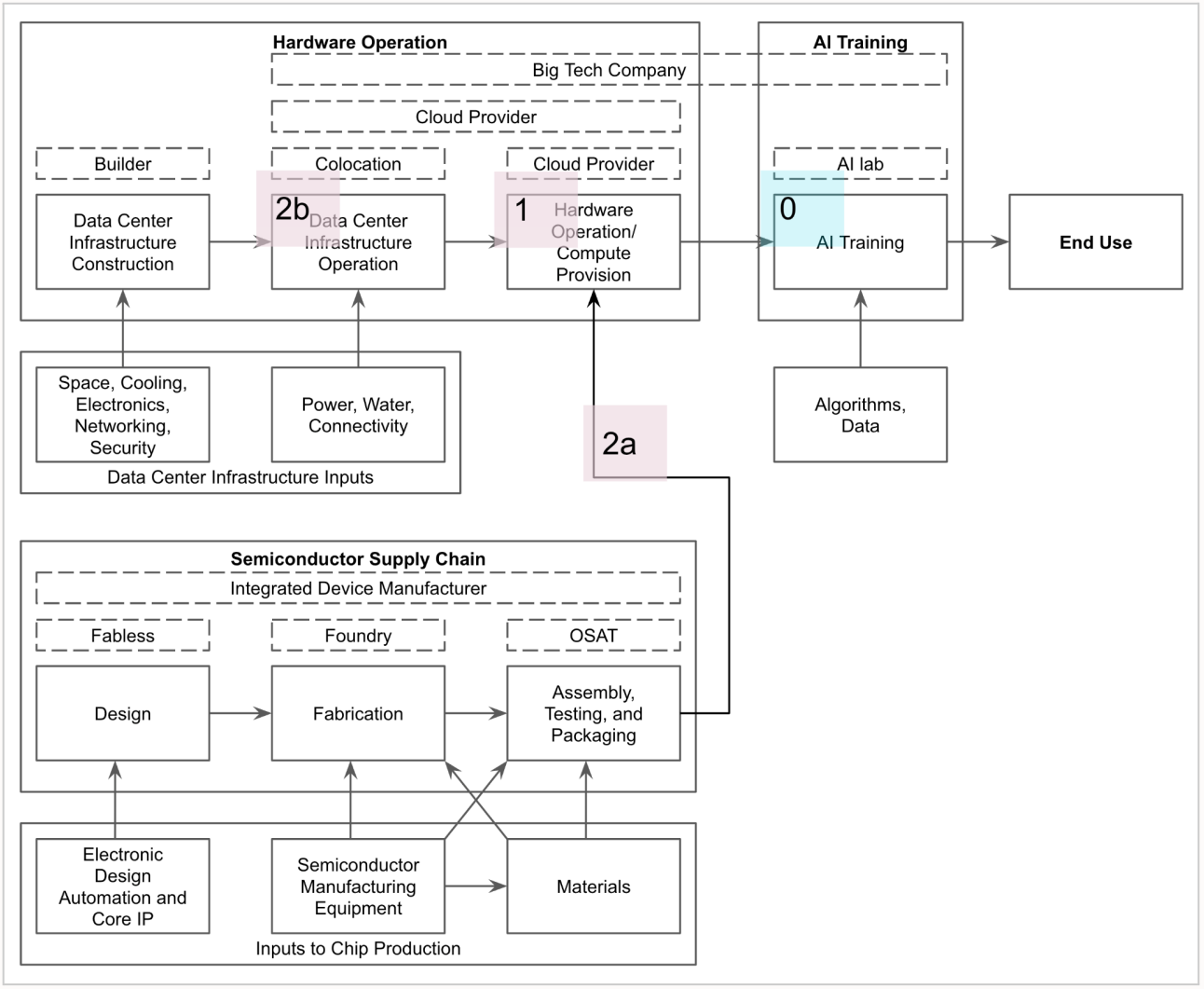

Figure 4c - Different steps of monitoring AI development.
To regulate compute usage in AI development (0), it is likely necessary to monitor hardware ownership (1), which can be verified via tracking hardware sales (2a) and routine inspections on the data center infrastructure level (2b).

Monitoring and regulating compute usage for AI development is a powerful lever for controlling both which actors have access to frontier models and which types of models these actors create. Increasing government monitoring of and control over AI compute clusters is also one of the most commonly shared intermediate goals within the AI governance community (Aird et al., 2023). While regulations on compute usage would be most effective if enforced directly on the AI hardware, even hardware-based mechanisms likely require knowledge of hardware ownership (Shavit, 2023).

To acquire information on hardware ownership, it is most straightforward to mandate compute owners to report the hardware they own to a central entity. However, absent verification mechanisms, this approach is vulnerable to misreporting. Particularly in scenarios

where owning certain amounts of hardware requires compliance with stricter regulations, companies could have incentives to underreport the hardware they own.[15]

Tracking hardware sales could offer a more reliable way to acquire ownership data (2a in Figure 4c). This could consist of major hardware providers and their distributors being mandated to report to whom they sell data center-grade AI accelerators. To verify that end-users do not secretly resell the hardware, a central entity could regularly demand AI accelerators with unique IDs to be presented to authorities.

In case this approach is intractable or insufficient, an additional way of verifying hardware ownership could be to monitor installed hardware at the data center. Regulations could require the infrastructure provider to report which hardware is hosted. (2b in Figure 4c) This would be a particularly reliable way of verifying hardware ownership, as routine data center inspections could reveal whether or not the compute hosted in a given data center is correctly reported. However, such inspections would necessitate knowing the locations of all data center facilities.

A comprehensive system of data center-level controls could follow the following steps.

1. Already existing data centers and future construction projects would require registration and disclosure of the hosted compute clusters, which would need to be approved by a central entity.
2. Colocation providers (renting out data capacity to other companies) would need to perform Know Your Customer (KYC) checks and report their customers' hardware, similarly subject to approval.
3. Routine inspections would be conducted to verify the presence of the reported hardware and ensure that no unreported hardware is being used.

A verification regime based on tracking chip sales, data center-level controls, or a combination of both could enable a variety of regulations on the compute level, such as requiring licenses for owning amounts of compute that could be used to train dangerous models or mandatory reporting of compute usage to verify responsible use.

Verifying compute usage

Even with a monitoring regime that reliably tracks hardware ownership, an additional challenge is verifying the claims compute users make about their usage.

Shavit, 2023 introduces the concept of AI accelerators logging their usage in the context of a verification mechanism. Until such mechanisms are implemented on AI accelerators, comparable hardware-based mechanisms could potentially be implemented at the compute cluster level. E.g., network switches could log the communication between GPUs, or power

[15] This issue is not hypothetical: similar cases of misreporting have happened in the past, such as in the Volkswagen emissions scandal, where the company actively manipulated emission test results to meet regulatory standards.

distribution units could log the power consumption of different system components to show patterns in utilization.

Data center security

The parameters of advanced AI models or the instructions for how to train them could get stolen from large AI labs. This could result in a) an increased number of actors on the forefront of AI capabilities, which could contribute to race dynamics and complicate coordination, and b) leakage and proliferation of powerful models to bad actors. More speculatively, various AI takeover scenarios involve the AI system exploiting cyber vulnerabilities to get access to more compute.[16] Improving such standards could thus make it more difficult for an out-of-control model to accumulate more resources.

Data center infrastructure primarily involves physically securing the hardware running advanced models. Additionally, several layers of potential vulnerabilities in the software stack are best addressed by the infrastructure provider, such as hardware security and the security of fundamental levels of software, such as virtualization and operating systems.

Containment of dangerous models

To quickly shut down models causing harm, data centers can incorporate mechanisms, such as emergency power-off or network shutdown, which could be implemented at various levels of the data center infrastructure and hardware stack.

For an additional layer of safety, compute clusters training large models could be air-gapped, i.e., disconnected from the internet, until safety measures such as finetuning have been implemented, and the models have undergone capability and alignment evaluations as well as risk assessment.

Challenges for governing data center infrastructure

Major challenges remain that need to be addressed before engaging in policy work.

Information availability

Throughout my investigation, I consistently observed that the data center industry is relatively opaque. E.g., there is no reliable data on data center capacity and ownership and few reports on common business practices. The lack of transparency makes it challenging to scope regulation well and could lead to potential downsides being overlooked. I caution against advocating for any data center-level regulation before more thorough research is conducted and relevant policymakers are consulted.

[16] E.g., as outlined by Shlegeris, 2022.

Are governments willing to regulate?

The data center industry has largely escaped public scrutiny, with most governance currently being industry-led. For example, reliability standards are defined by a private auditing firm, the Uptime Institute (*Uptime Institute, 2021*). However, as AI issues continue to gain prominence, it's likely that regulatory attention will increase. It is advisable to proactively develop and prepare such measures focused on the governance implications outlined above so that when an opportunity arises, advocates are well-positioned to promote beneficial governance regimes.

Regulatory arbitrage

Regulation is weak if it can be easily avoided. Since data centers exist all over the world, international coordination may be required to prevent data center operators from moving to parts of the world with less regulation. However, as outlined above, data center regulation could be reinforced by other measures, such as tracking chip sales, which could potentially help enforce regulations globally.

Technical feasibility

None of the proposed mechanisms for air-gapping, remote shut-down, or logging of hardware activity have been tested for feasibility. More research is needed to scope proposals for such mechanisms accordingly.

Conclusion

Data centers could present several opportunities for implementing regulations contributing to mitigating risks from advanced AI systems. Yet, data centers are only one piece of a broader set of regulations that needs to be developed.

Future work should explore the potential value of the outlined mechanisms, examining their technical feasibility and outlining road maps for implementation. A forthcoming piece will feature some of the most important research questions on data centers in the context of AI governance.