Bachelor Thesis

Sustainable AI - A Technical, Legal, and Economic Perspective

Vladica Spasojevic

vladica@spasojevic.io
Matr.Nr. 12118213
Datum: November 14, 2025

Abstract

Artificial Intelligence (AI) is rapidly becoming a cornerstone of modern society, yet its exponential growth has introduced environmental concerns and sustainability challenges. Training and deploying large-scale models requires immense computational power, leading to substantial energy consumption and associated carbon emissions. This thesis explores the concept of **Sustainable AI** through a multidisciplinary lens, integrating technical, economic, and legal perspectives to examine how AI can be made more environmentally responsible without hindering innovation. The **technical perspective** explores methods to improve efficiency, including model compression, specialized hardware, software optimizations, and energy-efficient data centre design. The **legal perspective**, focused on the EU, Austria, and Germany, reviews regulatory frameworks such as the Energy Efficiency Act and AI Act shaping the future of AI and how they address its environmental impact. The **economic perspective** highlights how sustainability aligns with cost savings and new business opportunities, making sustainable AI not only ecologically necessary but also financially beneficial. The findings show that technical advances enable Sustainable AI, economic incentives support it, and legal frameworks establish it as standard practice.

Contents

1 Introduction

Artificial Intelligence (AI) has long been a goal of humanity, though the scientific field dedicated to its development has only emerged not too long ago. Warren McCulloch and Walter Pitts published the first recognised works of artificial intelligence in 1943[1]. Later on, the famous Turing Test, proposed by Alan Turing in 1950, provided a more operational definition of intelligence[2] and shortly after the 2nd World War, the name AI was coined in 1956[3]. Although it is a fairly new scientific field, its subfields encircle a broad area, ranging from general knowledge like Large Language Models (LLMs) or more specific areas including autonomous driving, playing chess, solving mathematics[4], or the very recent breakthrough in designing Protein Structures, AlphaProteo[5]. AI is a prominent field that has experienced remarkable growth not only in its effectiveness but also in popularity, especially since the launch of LLMs and more specifically OpenAI's ChatGPT, which gained 1 million users within the first five days of its free launch[6]. Today there exists a large variety of proprietary or free AI products which can assist in writing tasks, perform complex calculations, generate videos, and more. While AI has driven remarkable progress across multiple fields, its rapid development has also increased concerns about sustainability. AI systems require a lot of computational costs, particularly during training and inference, which is a key contributor to rising energy consumption and carbon emissions. A massive growth in data consumption can be observed, as data ingestion for recommendation systems has increased by 3.2-fold from 2019 to 2021[7], which translates to significant demand on resources and energy consumption. Humanity is facing a climate crisis as global temperatures continue to rise, resulting in insecure food and water supplies, more frequent natural disasters, and many more problems which need to be addressed[8]. While it is impossible to completely reverse the damage already done, efforts can be directed towards a cleaner world. One of the increasingly critical practices is sustainability, which is why this thesis will cover **Sustainable AI** in three core aspects: Technical, Legal, and Economic.

1.1 Context and Problem Statement

The rapid growth of AI has not only yielded positive outcomes but also raised concerns for the environmental impact of training and operating AI models. It has caused an unprecedented surge in resource demand, with some studies showing a 300,000x increase in computing power used for training AI between 2012 and 2018[9]. Figure 1 illustrates the 'modern era', which began in 2012, highlighting the increase in computing power used for training AI models during this time. During this period, a notable trend can be observed where the need for computational resources has roughly doubled every 3.4 months, indicating an exponential increase. In contrast, the 'first era' instead shows a rough tracking of Moore's Law, doubling only every two years[10]. Another rising concern is that AI typically requires significant power, which is why it is often powered in data centres. Currently, nearly 3% of energy consumption in the EU comes from data centres, which is projected to increase by 28% by 2030[11]. These statistics underscore that the rise of AI must be contributing to an increase in carbon footprint, but more on its impact will be explored in the 2nd Chapter.

[1] Russell et al. 1995, p. 16.
[2] Ibid., p. 2.
[3] Ibid., p. 1.
[4] Ibid., p. 1.
[5] Zambaldi et al. 2024.
[6] Burmagina 2025.
[7] Wu et al. 2022, p. 1.
[8] w.A. 2025i.
[9] Amodei and Hernandez 2018.
[10] Ibid.
[11] Butler 2023.

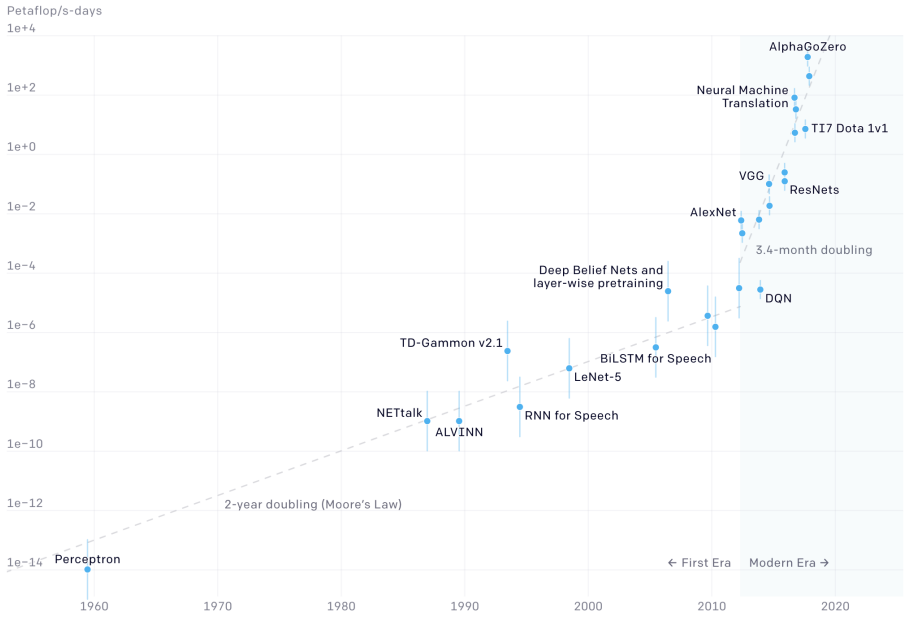

Fig. 1: Illustration of the contrast between the 'first era' and 'modern era' of computing power used in AI. Source: Amodei and Hernandez 2018

.

1.2 Scope, Objective Research Questions

This thesis delves deeper into the concept of sustainable AI, which is how to operate or use AI more environmentally friendly. Achieving sustainability will, however, require a multidisciplinary approach, which is why this thesis will, at its core, examine three aspects that need to work together to achieve environmentally friendly AI:

- **The Technical Aspect** will span from a broad perspective – comparing models and architectures that achieve high performance with significantly lower energy usage – and then dive deeper into how such efficiency is achieved through compression techniques, specialised hardware, software, and algorithmic innovations.
- **The Legal Aspect** will overview legal frameworks and policies which foster sustainable AI, primarily focusing on the European Union, while also zooming into Austria's and Germany's regulatory practices. Furthermore, it will examine how legal measures and incentives for local AI development and data centres can promote the sustainability agenda and finally highlight policy gaps.
- **The Economic Aspect** will discuss incentives on how implementing sustainable AI practices can reduce energy consumption and costs, as well as examine the innovation it sparks and the opportunities it offers.

Key questions of this thesis will include how can AI models and infrastructure be designed to be more energy-efficient without significant loss of performance? How are current laws facilitating sustainable AI practices, and what is the future outlook? How is adopting sustainable AI beneficial for companies? Following this introduction, Chapter 2 provides background on AI's environmental impact and the need for sustainability. This sets the stage for the core chapters, where each aspect will be discussed in depth. A discussion will highlight the interplay between technical, legal, and economics and appeal to the need for collaboration. To end this thesis, the conclusion will summarise the findings and talk about the future outlook.

2 Environmental Impact and the Need for Sustainability

This chapter explores the rising environmental footprint of Artificial Intelligence, emphasising the need for sustainability in AI. It begins by highlighting the rising energy consumption of AI systems and then compares real-world consumption metrics which contextualise the impact. The following section will discuss why sustainability is important and the implications for our climate. Furthermore, a brief overview of why exactly data consumption in AI is rising will follow. This chapter concludes by defining sustainable AI and briefly discussing current initiatives, which are reviewed more in depth in Chapter 4.

2.1 Rising Energy Footprint of AI

With the rise of deep learning and the emergence of large language models, the environmental footprint of AI has increased significantly. The previous section highlights the rapid growth of data consumption of AI over the past years, but it might be difficult to imagine how it impacts the environment. Figure 2 shows how the training of the GPT-3 175 billion parameter model consumed several thousand petaflops a day[12]. Although OpenAI remained vague on the energy consumption associated with training this model, subsequent research by Luccioni et al. 2022 offers a more precise estimation of the computational demands involved. The training of GPT-3 was estimated to have consumed 1,287 MWh of power, which is equal to 502 tonnes of carbon emissions[13]. To put this into perspective, the average Austrian household consumes about 4,415 kWh in a year. Based on this figure, the energy used to train GPT-3 could have powered about 291 Austrian homes for an entire year. For further perspective, streaming one hour of Netflix requires roughly 0.8 kWh, or 0.0008 MWh. This means to consume the same amount of power, one would have to watch over 1.6 million hours[14] or roughly 182 years of content on Netflix. Newer models have even more parameters to train, e.g. GPT-4 has 1.76 trillion. Obtaining estimates on every model is difficult since most companies training them do not enclose such data; however, more parameters do not necessarily translate to more energy consumption. For instance, a comparable LLM BLOOM with 176 billion parameters consumed an estimated 433 MWh of power during training, resulting in only 70 tonnes of carbon emissions, which would be equal to 20× lower than that of GPT-3[15]. This can mostly be attributed to the carbon intensity of the energy source used for training, as the carbon intensity of the electric grid BLOOM was trained on is 57 gCO_2eq/kWh, compared to 429 gCO_2eq/kWh for GPT-3[16]. The following Chapter 3 will explore the techniques and strategies more in depth which can be used to achieve efficient AI.

2.2 Why Sustainability is important

Society is facing human-induced climate change, which has been a growing concern for well over a century. The so-called pioneer of climate change research, Svante Arrhenius, was the first researcher to publish articles in 1896 on how CO_2 would change the global temperature[17]. Today, gases which trap heat and therefore increase global temperature are called greenhouse gases. CO_2 is a major greenhouse gas and its concentration has significantly increased due to human activities. As global temperature continues to rise, society is witnessing not only hotter summers in some regions but also a growing number of natural disasters. Massive floods, record temperatures above 40° Celsius, melting glaciers, droughts, and many more have become regular news headlines. To fight this, there have already been numerous climate change pledges, with

[12] Brown et al. 2020, p. 39.
[13] Luccioni et al. 2022, p. 7.
[14] Vincent 2024.
[15] Luccioni et al. 2022, p. 7.
[16] Ibid., p. 7.
[17] Wulff 2020, p. 1.

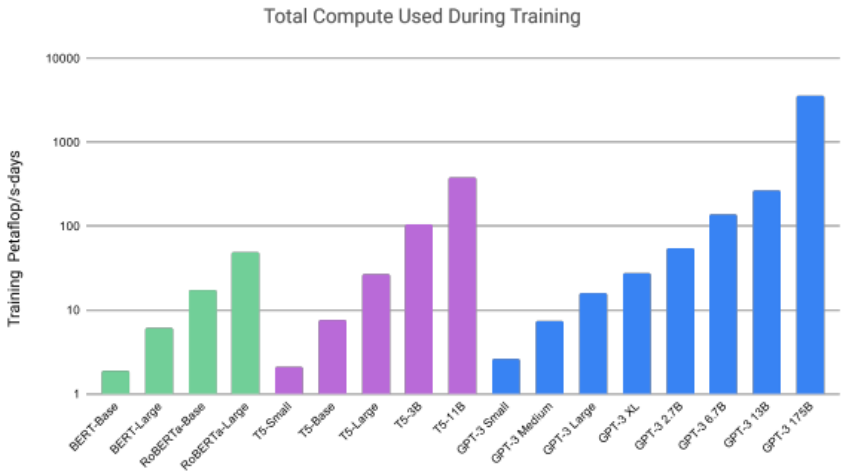

Fig. 2: A comparison of computing power used for training AI Models. Source: Brown et al. 2020, p. 9

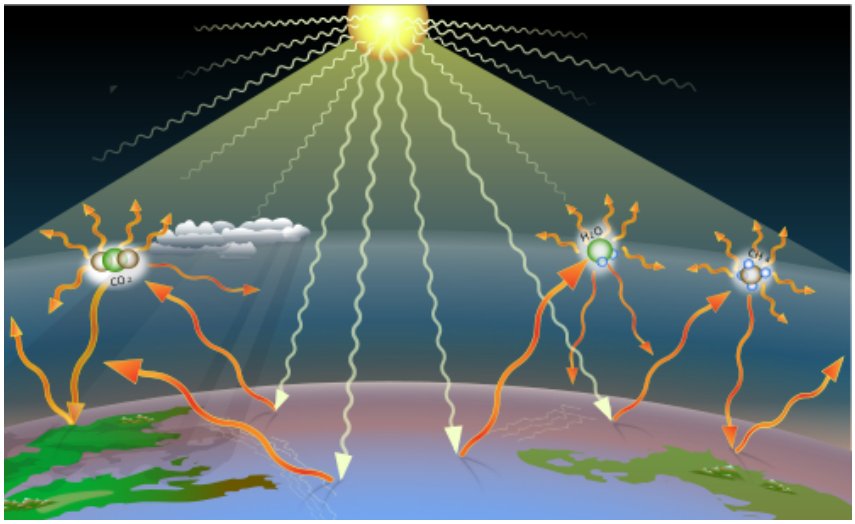

Fig. 3: This illustration shows the three greenhouse gases carbon dioxide, water vapour, and methane, and how some of the heat, radiated by the sun, is trapped by them. Source: w.A. 2025f

the Paris Agreement being the largest. It has been ratified by almost every country and pledges to keep global warming below 1.5°C[18]. To reach this goal, the EU committed to reducing its greenhouse gas emissions by 55% by 2030, compared to its 1990 levels, and aims to become climate-neutral by 2050[19]. Given AI's rapid growth in energy consumption, managing its carbon footprint and developing a sustainable AI lifecycle is crucial for meeting these climate targets.

2.3 Why is data consumption rising?

Following the examination of AI's rapid growth in data consumption and its impact on the environment, it is important to understand the reason behind this surge.

Throughout the history of computer science, there has always been a focus on algorithms and how to increase their performance. Traditional improvements relied on Moore's Law and better algorithms, but in AI, abundant data has been a more critical factor. An enormous

[18] w.A. 2025h.
[19] Ibid.

variety of sentence structures in forums, billions of different pictures, and work on word-sense disambiguation - which is how a word might correlate to another word given context - along with rapidly increasing availability of diverse data, have all been key drivers behind the increasing effectiveness of AI systems[20]. Banko and Brill showed in 2001 that focusing on the algorithm is not going to be of advantage by comparing average and excellent algorithms on different datasets. The observations concluded that the average algorithm trained on 100 million words of training data bests an excellent algorithm trained on 1 million words of training data[21]. Another work examines how filling photos using a collection of pictures improved exponentially by increasing the collection of photos by a factor of 200[22]. In summary, AI will perform better by examining the learning methods and providing enough training data. This underscores why AI has been able to have such exponential growth and how it has benefited from the increase in data availability. However, this hunger for data translates to ever-growing datasets and computational loads, which translates to higher energy consumption.

2.4 Defining Sustainable AI

Sustainable AI refers to the design, development, and deployment of artificial intelligence systems in ways that minimise environmental impact. It emphasises energy efficiency across the full AI lifecycle, ranging from model training to inference and its utilisation. It aligns with broader sustainability goals such as reducing carbon emissions, integrating renewable energy, and supporting long-term societal benefits.

In research, there is currently a differentiation between two approaches to AI. One is called *Red AI*, which is, at the time of this writing, the more prevalent approach. It strives for state-of-the-art results by sacrificing efficiency and achieving more accuracy through using massive computational power in its training and deployment[23]. While this has brought valuable contributions to AI research, it may not be the optimal approach for several reasons. The first, and most obvious, is the already discussed impact it brings on the environment. The second becomes clear when examining the relationship between performance and complexity. Model complexity can be measured by the number of parameters or the inference time, whereas inference time describes the time it takes for an AI model to make a prediction. Figure 4 shows that beyond a point, doubling model size yields only a small accuracy improvement, showing a roughly logarithmic relationship[24]. This suggests that Red AI might not be optimal and future-proof, as AI development might soon hit a ceiling of data that can be consumed.

Green AI, the alternative and more environmentally friendly approach, is a call for efficiency to be the primary success metric, instead of accuracy[25]. Along with accuracy, it has been a widely accepted metric for research. Green AI encourages researchers to report and optimise computational cost (e.g. FLOPs, energy or CO_2) of new models in addition to accuracy, thus making it more inclusive by enabling broader participation due to a lowered resource barrier[26]. This approach will be discussed further in Section 3.2.1 of the following Chapter. Chapter 5 will discuss how companies utilised AI to efficiently organise resources, therefore reducing energy consumption and their carbon emissions.

[20] Russell et al. 1995, p. 28.
[21] Ibid., p. 28.
[22] Ibid., p. 28.
[23] Schwartz et al. 2019, p. 2.
[24] Ibid., p. 2.
[25] Ibid., p. 5.
[26] Ibid., p. 5.

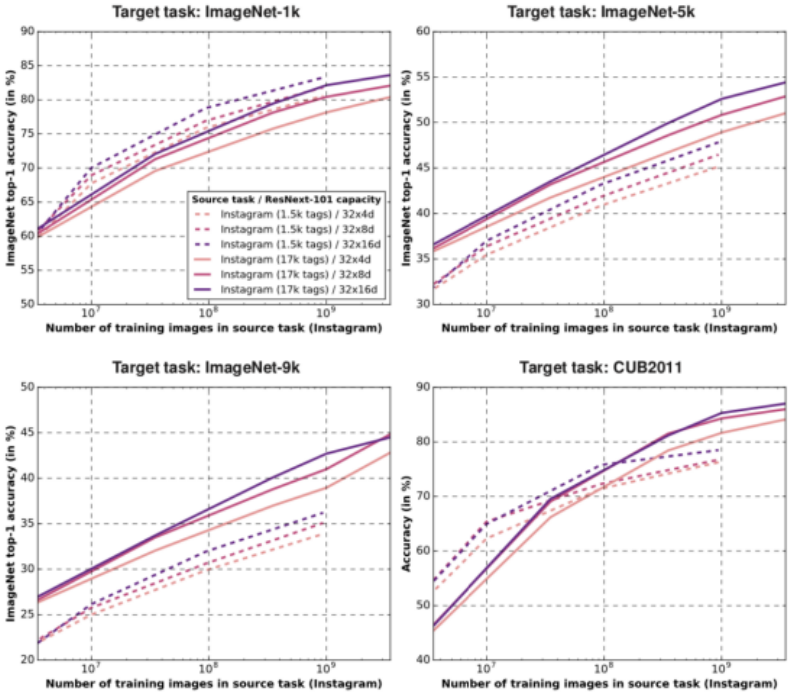

Fig. 4: In this illustration, three different ImageNet and one CUB2011 dataset are shown, which are famous datasets used for image recognition tasks. In all of them, it can be observed that only with an exponential increase in model size a linear growth in accuracy can be achieved, showing that the relationship is at best logarithmic. Source: Schwartz et al. 2019, p. 5

2.5 Current Initiatives and Awareness

Growing climate awareness has not only led to more sustainable industry practices, as well as academic initiatives, but also to initiatives leading research on sustainable AI. Investment in renewable energy is increasing, which could power data centres operating AI. Hardware is getting continually more energy efficient, Green AI is gaining more prominence, and metrics regarding energy usage per inference are increasingly reported. Additionally, growing awareness of the information and communications sector can be observed, which accounts for five to nine percent of global energy usage[27]. The core chapters of this thesis will discuss sustainable initiatives, practices and awareness.

Following the deep dive into the impact of AI on the environment, the importance of sustainability, and the concept of sustainable AI itself, will be the subsequent technical, economic, and legal analyses, which highlight the multi-disciplinary effort of achieving sustainable AI, forming the core of this thesis. As explored earlier, the large language model BLOOM succeeded in reducing its carbon footprint by a factor of 20 compared to GPT-3's, while maintaining similar model complexity. The strategies to achieve such outcomes will be discussed in Chapter 3 **Technical Strategies for Sustainable AI**.

[27] Butler 2023.

3 Technical Strategies for Sustainable AI

As AI systems' environmental impact continues to grow, developing technical strategies to reduce energy consumption and carbon footprint has become critical. These strategies include energy-efficient model designs, techniques for achieving such designs, specialised hardware, and software optimisations to maximise efficiency, and efficient training methods and intelligent resource scheduling.

This chapter starts off by outlining important Neural Networks, as they are mentioned many times throughout this thesis.

It will continue by exploring key approaches to energy-efficient AI models, diving deeper into the **Green AI** philosophy, as well as different model designs and architectures through real-world examples.

The discussion then moves towards model compression techniques, such as pruning, quantization, and distillation, which are presented as practical methods for building lightweight and low-energy AI models.

Furthermore, the section on hardware and software optimisations highlights how specialised hardware and optimised software can impact energy consumption. Finally, it will also examine how energy-efficient data centre design can help reduce carbon emissions.

Additionally, strategies for more efficient training algorithms will be addressed, while also including an optimised way of handling resources. Throughout this chapter, trade-offs between performance and energy consumption are examined, offering insights and real-world examples of how AI Models can come to a balance between accuracy and efficiency, and be more environmentally responsible.

By outlining these strategies, this chapter illustrates the critical role that technical innovation plays in the broader effort to make AI development sustainable.

3.1 Overview of Neural Network Architectures in AI Systems

Artificial Intelligence (AI) systems have seen rapid evolution over the past decade, largely driven by the development of specialised neural network architectures. These architectures form the computational backbone of modern AI and are tailored for different types of data and tasks. This section provides an overview of commonly used systems, focusing on their structural differences and typical applications.

3.1.1 Deep Neural Networks

Deep Neural Networks (DNNs) are a general class of artificial neural networks with multiple hidden layers between input and output. DNNs represent networks that consist of multiple (typically 4 or more) hidden layers[28]. Hidden layers consist of neurons, weights, biases (also called their parameters) and activation functions, such as Sigmoid or ReLU. Calculating inference, which is the process of using a trained model to make predictions, is described as:

1. Multiplying our input data with the weights and adding our biases[29]
2. Aggregating the results into a single value[30]
3. Applying an activation function onto the state to modulate activity[31]

[28] Dhilleswararao et al. 2022, p. 5.
[29] Kaz Sato 2017.
[30] Ibid.
[31] Ibid.

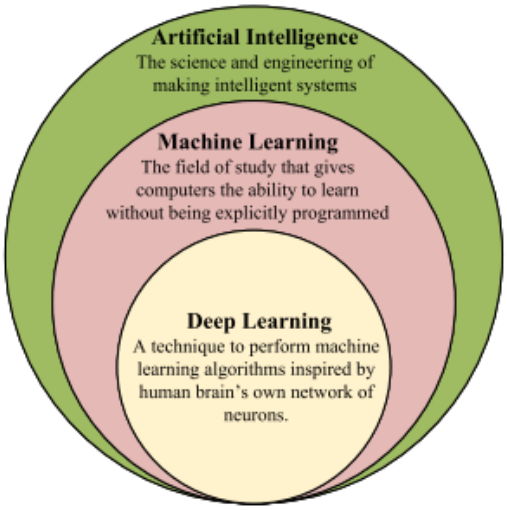

Fig. 5: Illustration showing the science of AI and how it is broken down to Deep Learning. Source: Dhilleswararao et al. 2022

This sequence can also be written as matrix multiplications. While conceptually simple, DNNs can model intricate relationships and are widely used in tasks like tabular data classification, speech recognition, and fraud detection. Standard DNNs often lack inductive biases that make them efficient for certain data types (e.g., images or sequences), and they can become computationally expensive as depth increases. They serve as the foundation from which more specialised architectures, like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have emerged. Recent architectures like the Transformer (used in GPT-3 and other LLMs) are also part of this landscape.

3.1.2 Transformers

Transformers have become a staple deep learning model in various artificial intelligence fields such as natural language processing, computer vision and audio processing[32]. They have become a go-to model for natural language processing, as they show state-of-the-art accuracy when transformer-based pre-trained models are used[33]. There are numerous Transformers which differ in strengths and weaknesses; however, the vanilla Transformer is a sequence-to-sequence model consisting of an encoder and a decoder[34]. The key modules of the decoder and encoder are:

- **Attention Modules** in Transformers are adopted in a Query-Key-Value model[35]. It helps it keep attention on more important information.
- **The Position-wise FFN** can operate separately and identically on each position[36].
- **Residual connection and normalisation:** by employing a residual connection around each module, followed by a normalisation layer, a deep model is built[37].
- **Position encodings** help the Transformer grasp its current position[38]

[32]Lin et al. 2022, p. 1.
[33]Ibid., p. 1.
[34]Ibid., p. 2.
[35]Ibid., p. 2.
[36]Ibid., p. 2.
[37]Ibid., p. 2.
[38]Ibid., p. 2.

3.1.3 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are networks trained for solving visual tasks and their deployment has already been in research since the late 1980s[39]. It has been one of the most important neural networks, as it allows for tasks once deemed impossible, such as facial recognition and autonomous driving[40]. CNNs are a type of feed-forward network which automatically extract features from structured data, especially images[41]. They are inspired by biological visual perceptions as they mimic the way neurons respond respectively to specific stimuli[42]. They are efficient due to their reduced number of parameters from implementing local connections and weight sharing[43]. A CNN consists of five core components:

1. **Convolution** applies kernels to input data to extract features, thus producing feature maps[44].
2. **Padding** enlarges the input, usually with zeros, to preserve information during convolution[45].
3. **Stride** controls the density of convolving[46].
4. **Pooling** combats overfitting by removing redundancies (e.g. max-sampling downsamples the feature maps)[47].
5. **Dilated Convolution** helps convolution kernels perceive larger areas of the image without additional weights[48].

Together, they enable CNNs to efficiently learn hierarchical and spatially aware features, making them foundational in deep learning for vision tasks.

3.2 Energy-Efficient AI Models and Algorithms

While the field of Artificial Intelligence has made continuous progress, many of these breakthroughs have largely been driven by the Red AI mindset. Although this approach enabled rapid progress, it did so by sacrificing environmental care through the usage of massive amounts of power. This established multiple norms, which include expensive training and inference costs, large-scale data processing and running extensive experiments (hyper-parameter searches, model ensembling, etc.)[49]. It is effectively buying stronger results with brute-force computation, while also excluding smaller research groups that cannot compete due to a lack of resources. In response, research has started advocating and demonstrating Green AI, a more environmentally and inclusive approach. Section 3.2.1 will dive deeper into Green AI.

Moreover, to emphasise how focusing on efficiency can still have extraordinary performance, the subsequent section will compare real-life models like DistillBert, which are compressed or distilled models that are trained on a larger 'teacher' model. Using this method, DistillBert managed to retain most of the larger model's performance, but with a reduced size and improved speeds. The discussion then moves to architectural innovations, illustrated by EfficientNet, which

[39] Dash 2025, p. 1.
[40] Li et al. 2022, p. 1.
[41] Ibid., p. 2.
[42] Ibid., p. 2.
[43] Ibid., p. 2.
[44] Ibid., p. 2.
[45] Ibid., p. 2.
[46] Ibid., p. 2.
[47] Ibid., p. 2.
[48] Ibid., p. 2.
[49] Schwartz et al. 2019, p. 4.

used neural architecture search and compound scaling to achieve state-of-the-art performance with far fewer resources compared to traditional scaling methods.

Together, these insights demonstrate that both smart model compression and thoughtful architectural design are essential strategies for advancing Sustainable AI.

3.2.1 'Green AI' Approach

The more prominent way of AI research has been Red, with its focus being on model accuracy. Green AI, however, seeks to have a more positive impact on the environment by prioritising efficiency alongside accuracy, resulting not only in reduced energy consumption but also greater inclusiveness, by opening AI research to a broader audience[50]. A common way to measure efficiency is to examine the amount of computational power that is required to train a model.

However, there may be various factors that can influence efficiency results, such as local electricity infrastructure, the hardware used, and other conditions, which is why it is important to find a stable measure across different locations, times, and hardware configurations:

- **Carbon Emission** has already been mentioned a fair amount across this thesis. It is however, not always practical to use as a stable efficiency measure, since it can be influenced by different factors, such as the local energy infrastructure and location[51].
- **Electricity usage** is similar to carbon emission, only that location and time are not relevant. Mostly, it is measured by the amount of electricity consumed by a GPU when generating AI results, thus making it also influenced by hardware[52].
- **Elapsed real time** would usually be a very obvious answer to measure efficiency. The faster a process can finish, the less computational power is used, resulting in more efficiency. Although plausible, it is still not a stable enough measure for efficiency, due to it also being influenced by the hardware performance (e.g. how efficient is the hardware, are there any other jobs running, how many cores are available for training)[53].
- **Number of parameters** is a common measure for efficiency, which describes the number of parameters used by an AI Model. Similar to runtime, it is correlated with the amount of work performed; however, unlike other measures, it is independent of the underlying hardware. Yet, it still does not represent the most suitable efficiency measure, due to algorithms utilising their parameters differently, e.g. by having a wider instead of deeper model. This often leads to different efficiencies among different models with similar parameters, e.g. a 100 million–parameter CNN might outperform a 100 million–parameter RNN on images due to architecture[54].

After thorough examination of these and why they are not optimal and stable enough energy efficiency measures, Green AI researchers advocate for using floating point operations or **FLOP(s)** as a measurement metric. They describe the number of basic mathematical operations a running AI model performs, specifically additions (ADD) and multiplications (MUL). FLOP(s) can be calculated by breaking down complex problems, like multiplying matrices or running big models, as well as breaking them down into multiple ADD and MUL operations and adding up the cost. This shows how much work a machine has done, which can therefore be closely correlated to the energy consumption of a machine. It is also strongly associated with elapsed real time, but unlike the elapsed time, the work done can be quantified at each time step, independent of

[50] Ibid., p. 5.
[51] Ibid., p. 6.
[52] Ibid., p. 6.
[53] Ibid., p. 6.
[54] Ibid., p. 6.

Model	Score	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST-2	STS-B	WNLI
ELMo	68.7	44.1	68.6	76.6	71.1	86.2	53.4	91.5	70.4	56.3
BERT-base	79.5	56.3	86.7	88.6	91.8	89.6	69.3	92.7	89.0	53.5
DistilBERT	77.0	51.3	82.2	87.5	89.2	88.5	59.9	91.3	86.9	56.3

Fig. 6: This table illustrated the comparison of three different models: BERT, DistilBERT and ELMo on the GLUE benchmark, where DistilBERT show a retaining 97% performance of BERT. Source Sanh et al. 2020

hardware slowdowns. With these connections, FLOP(s) are also not influenced by the underlying hardware, making it a fair and stable comparison metric between models. Some tools have already integrated FLOP(s) calculations in popular machine learning algorithms, but some modern AI models are still missing needed parts, which is why Green AI is encouraging the implementation of such functionality[55].

In practice, this thesis will still discuss and not overlook metrics such as carbon emissions or energy usage, since the community and society hasn't yet standardised a measurement metric. Some models can still be compared by, e.g. carbon emissions when the same location and hardware are in use.

3.2.2 Smaller Models, Big Impact

OpenAI's GPT-3 with 175 billion parameters demonstrated remarkable capabilities, but also brought a great amount of energy consumption with it. The trend towards larger models has not only brought environmental concerns, but operating models on the edge will become increasingly difficult, due to growing computational demand and memory usage. With algorithmic innovations like knowledge distillation, researchers have found models maintaining a similar performance with only a fraction of their counterparts' size while boasting a much higher computational speed. These models are called distilled or efficient models and achieve similar results using far fewer resources. Knowledge distillation is discussed briefly here and in more detail in Section 3.3.3.

DistilBERT, a distilled version of BERT, managed to reduce its size by 40% and be 60% faster while retaining about 97% of its language understanding abilities, tested with the GLUE benchmark (see Figure 6)[56]. For training, a compression technique called knowledge distillation was used, in which a distilled model, also called the student model, is trained to replicate the behaviour of a larger model, or the teacher model. The student model is trained on a loss, where it tries to match the teacher's output distribution, and not just predict the correct result. To achieve this, a cross-entropy loss between the soft target probabilities from the teacher and the student's predicted probabilities was used. Normally, a softmax makes one class very confident and others not, but here, a softmax temperature was used. A higher temperature produces a flatter (less confident) distribution, meaning the teacher's probabilities are spread out, so the student can learn which mistakes are 'almost correct'. Finally, the student had the same general architecture as the teacher, with a focus on reducing the number of layers, due to prior studies showing that model depth had the most impact on efficiency[57].

3.2.3 Efficient Model Architectures

DistilBERT exemplifies how compression can yield huge efficiency gains, but designing the right architecture can likewise have a positive impact on computational efficiency. To provide more

[55] Schwartz et al. 2019, p. 6.
[56] Sanh et al. 2020, p. 1.
[57] Ibid., p. 2.

context for this section, a model refers to a specific trained instance, whereas architecture refers to the general network design (the blueprint that could be instantiated and trained).

ConvNets (see Section 3.1.3) are usually designed at a fixed budget, and further scaled to a desired point if more resources are available. Scaling is primarily used to increase the accuracy of a model, e.g. GPipe, a very large pipelined neural network, managed a 84.3% Top-1 image recognition accuracy by scaling up its baseline by 4×[58]. However, different baselines can have a variety of impacts, which is why EfficientNet tried to focus on energy efficiency while keeping high accuracy. Their baseline network, EfficientNet-B0, was built using multi-objective neural architecture search, which tries to optimise accuracy with fixed FLOPs, effectively controlling the trade-off between accuracy and FLOPs[59]. Further steps include applying their compound scaling method by setting the compound coefficient ϕ, which uniformly scales the width, depth and resolution of ConvNets. The variable ϕ controls how many resources are available, using the constants α, β and γ, which have been found using a small grid search (the compound scaling method is explained in more depth in Tan and Le 2020, p. 5). Using this compound scaling method, and their baseline, EfficientNet-B7 resulted in having a 84.3% Top-1 ImageNet accuracy (the same as GPipe), while being 8.4× smaller and 6.1× faster than it and other leading ConvNets[60].

After exploring the role of models and architectures, the following Section 3.3 will delve deeper into model compression techniques that allow lightweight and resource-efficient AI systems.

3.3 Model Compression and Lightweight Models

In the previous Section, it was discussed how a compressed model can still deliver good results while using less computational power, directly translating to energy savings. This section will go more in-depth about techniques which can achieve such feats and create compressed and lightweight models. The key techniques which are examined are **Pruning** and **Quantization**, two forms of network compression that typically boost efficiency at the cost of only a small accuracy drop. Pruning uses various criteria to remove unneeded computation, and it can be categorised as static, which refers to performing pruning offline or dynamic, if it is performed during run time[61]. Quantization, instead, is a method of lowering data type precision, like weights, biases and activation, to reduce model size[62]. Both these techniques can be used either by themselves or in tandem to maximise model compression and improve efficiency.

Furthermore, the already mentioned **knowledge distillation**, used by DistillBert (see Section 3.2.2), will be explored. It refers to training a student model by trying to mimic a 'teacher' model's output, resulting in models that are much smaller and faster, and also retain most of the 'teacher' model's accuracy.

Finally, a discussion of real-world case studies of model compression in industry-scale systems will conclude the chapter.

3.3.1 Pruning

The development of pruning techniques started back in the 1990s, where it was used for memory size and bandwidth reduction, allowing AI systems to be deployed in smaller environments such as Internet of Things (IoT) systems. Optimal Brain Damage, proposed by LeCun in 1990, was an early pruning method to prune single non-essential weights[63]. Some case studies suggest that larger networks may be redundant. For example, GoogLeNet achieves 69.8% top-1 accuracy

[58] Tan and Le 2020, p. 1.
[59] Ibid., p. 5.
[60] Ibid., p. 1.
[61] Liang et al. 2021, p. 1.
[62] Ibid., p. 1.
[63] Ibid., p. 7.

on ImageNet with only 7 million parameters, comparable to VGG-16's accuracy but with far fewer parameters. Similarly, MobileNet reaches about 70% top-1 accuracy with just 4.2 million parameters and 1.14 GFLOPs[64]. Pruning works by locating parameters or neurons inside a neural network which do not contribute to inference accuracy, due to their weight coefficient being zero, close to zero, or replicated by another parameter. These are redundant parameters which are identified and removed using various pruning techniques to increase computational efficiency and reduce model size. In some cases, a pruned model can be fine-tuned or retrained to increase accuracy by removing overfitting and escaping local minima[65]. Although pruning may be classified by different aspects, a common way is to categorise it by when the pruning steps are performed: **Static and Dynamic Pruning** (illustrated in Figure 7).

- **Static Pruning** is a compression technique which describes the removal of parameters as an offline process, usually once after training, but before inference. It is commonly split into three parts: 1) selection of parameters to prune, 2) selection of pruning methods and 3) optionally re-training the pruned model, which may increase accuracy but with significant computational overhead[66]. However, during this process, the original structure is destroyed, and model capabilities may be decreased, which might not be the optimal choice in some cases[67].
- **Dynamic Pruning**, on the other hand, instead of trying to prune once before training, it focuses on removing layers, channels and neurons on the fly[68]. With this approach, the limitation of static pruning can be overcome by taking advantage of changing the input data and thus saving on computational overhead[69]. Dynamic methods typically do not require retraining during inference; they rely on the decision policy learned during training[70]. Most pruning techniques in practice are static (for simplicity), but dynamic pruning is an active research area for achieving extra efficiency at runtime.

Pruning methods usually only differ in the way they choose what to prune. Some methods calculate how sensitive the network is if a specific neuron or weight is removed, and then remove the least sensitive parameters first[71]. Other methods use regularisation and add penalty terms to loss functions (such as L0/L1 norm penalties), which encourages weights to become zero or close to zero, effectively pruning the model[72]. A good pruning method baseline, which works well in practice, is magnitude-based pruning, where the smallest-magnitude weights are removed. A study, which used the magnitude-based pruned model ResNet-50, showed higher accuracy than state-of-the-art models, while boasting the same computational complexity[73]. Overall, pruning can drastically shrink model size and compute demands. Researchers have reported, for example, compressing a VGG network by 20× (and 5× fewer computations) with only 0.1% accuracy loss using automated layer-wise sparsity tuning[74]. In summary, pruning removes redundant parameters from the networks to compress them in size, thus reducing their complexity, speeding up their processing and removing computational overhead, while losing almost no accuracy.

Pruning can be combined with other techniques, like quantization, for even greater effect, which is examined next.

[64] Liang et al. 2021, p. 6.
[65] Ibid., p. 6.
[66] Ibid., p. 6.
[67] Ibid., p. 10.
[68] Ibid., p. 10.
[69] Ibid., p. 10.
[70] Ibid., p. 10.
[71] Ibid., p. 7.
[72] Ibid., p. 7.
[73] Ibid., p. 12.
[74] Ibid., p. 9.

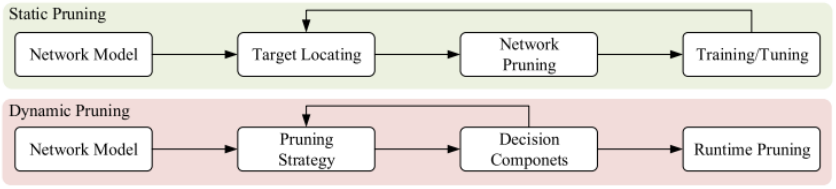

Fig. 7: Pruning steps can be run before inference, which refers to Static Pruning, while Dynamic Pruning completes its steps at runtime. Source: Liang et al. 2021, p. 6

3.3.2 Quantization

Quantization is another method of compressing an AI model, but instead of removing irrelevant parameters like pruning, it focuses on reducing the numerical precision of a network's parameters and computations. It describes the process of approximating a continuous signal by a set of discrete symbols or integer values[75]. Quantization methods can vaguely be categorised into weight clustering and low-bit numeric representations. The first approach works by restricting weights into small sets of clusters, e.g. by using k-nearest neighbours to find common weights, which are then stored in a compressed file instead of the full precision numbers[76]. These weights are then decompressed, using a lookup table or linear transformation, during inference time; however, such partial quantization mainly reduces storage and has to decompress weights at runtime, so it doesn't speed up computation by itself[77]. The more common approach, and the focus in most literature, is low-bit quantization of weights and activations for actual computation. This practice has been proposed as far back as the 1990s and translates to lowering the precision values of weights, biases, and activations[78]. For example, weights that were previously 32-bit are converted to 8-bit integer values, thus reducing memory overhead and in turn improving inference time. Typically, networks have been trained using a 32-bit floating point representation of their parameters, but research has shown that 8-bit values not only have faster inference times, but also significantly less storage requirements, while still having similar accuracy[79]. These findings piqued the interest in quantization, and nowadays, 8-bit quantization is seen as the sweet spot for accuracy and compression, as it provides roughly a 4× reduction in model size (8-bit numbers are 4× smaller than 32-bit) and is supported by GPUs, CPUs, and other specialised hardware, due to significantly faster 8-bit arithmetic operations[80]. By using proper techniques, 8-bit quantization or other quantization methods can still achieve high accuracy with minimal loss.

Quantization Techniques describe how quantization can be applied to networks. The main work flows include:

- **Post-Training Quantization** takes a fully trained model and quantizes the weights. For example, a usual 32-bit trained model is taken, its values are mapped to a lower-bit presentation (such as 8-bit), and often a calibration step is involved to adjust for quantization error[81].
- **Quantization-Aware Training (QAT)** integrates the quantization of parameters into the training process itself. A model is either trained or fine-tuned while simulating downscaled

[75] Ibid., p. 12.
[76] Ibid., p. 12.
[77] Ibid., p. 12.
[78] Ibid., p. 12.
[79] Ibid., p. 12.
[80] Ibid., p. 23.
[81] Ibid., p. 12.

low-precision weights, such as 8-bit, so that a network may tolerate the quantization effects[82]. By doing so, the model adapts to the quantization noise and typically yields higher accuracy after quantization compared to a post-training approach[83].

When applying quantization, it is important to consider the extent to which the network components are quantized. Quantizing only the weights can yield moderate reductions in memory usage and computation time. However, for more substantial performance gains, it is common practice to quantize both weights and activations to 8-bit integers. This broader quantization enables more efficient use of hardware accelerators that are optimised for low-precision arithmetic. In contrast, network biases are typically left in higher precision, such as 32-bit floating point. This is because they occupy a negligible portion of the model's overall memory, and preserving their precision helps avoid numerical instability, particularly during accumulation operations in inference[84]. Lower bit quantization has been researched, but it introduces re-training difficulties of networks, as well as problems for inference[85]. Choosing the proper method and adequate quantization techniques is key to reducing memory usage and improving computation times, while losing little accuracy.

3.3.3 Knowledge Distillation

The Section 3.2.2 has already discussed how knowledge distillation can be utilised to compress a model, to reduce size and improve its efficiency. This section will dive deeper into the methods and techniques which have been applied to the natural language model BERT to produce its distilled version, DistilBERT. To recap, DistilBERT is a compressed or distilled version of BERT, which managed to retain 97% of BERT's accuracy, but only using 60% of its size[86]. The key motivation was to bring complex models to edge devices. To achieve this, it used knowledge distillation, where a smaller student model learns from a more complex teacher model. The student model learns from the probabilities of class predictions, also called soft probabilities, which have been made by the teacher model, while minimising cross-entropy between both distributions[87]. This results in the student model mimicking the teacher's behaviour and predictions. DistilBERT employs a triple loss function during its knowledge distillation process to effectively transfer knowledge from the teacher model to the smaller student model. Specifically, DistilBERT's triple loss function consists of:

1. **Masked language modelling loss** is a technique to learn language patterns, where a word is replaced with a special character, which has to be then predicted by the model. The cross-entropy of the student's and teacher's predictions is calculated, and the student's parameters are updated using backpropagation[88].

2. **Distillation loss** is important to help the student capture patterns learned by the teacher. The probability distributions of the student model are matched to those generated by the teacher model[89].

3. **Cosine Embedding Loss** encourages the alignment of the hidden-layer embeddings between the teacher and the student, to ensure the student can replicate the teacher's internal semantics[90].

[82] Liang et al. 2021, p. 12.
[83] Ibid., p. 23.
[84] Ibid., p. 14.
[85] Ibid., p. 28.
[86] Sajid 2024.
[87] Ibid.
[88] Ibid.
[89] Ibid.
[90] Ibid.

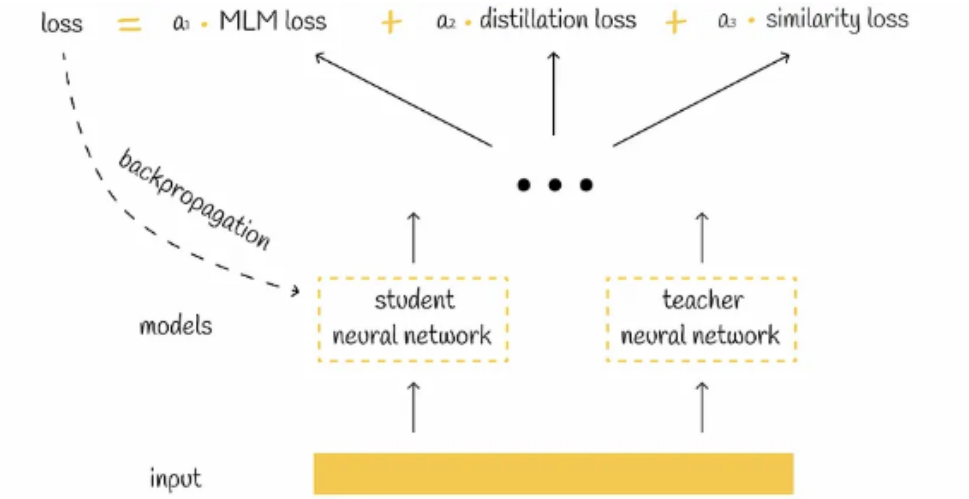

Fig. 8: In the knowledge distillation process of DistilBERT, the three loss functions—distillation loss, masked language modelling loss, and cosine embedding loss—were linearly combined into a single loss value. The resulting aggregation was subsequently used during the backpropagation phase to adjust the weights of the student model, thereby effectively guiding it toward the behaviour and representations of the teacher model. Source: Sajid 2024

When calculating the distillation loss, a softmax temperature of greater than one was used. This results in more distributed output probabilities, so skewed labels can be reduced due to the teacher's certain probabilities. This helps the student learn how to imitate the teacher more closely[91]. Figure 8 illustrates how the three loss functions were used to adjust the student's weights.

3.3.4 Real-World Impact

Section 3.3 has discussed different models and architectures and how model compression can impact the size and speed of AI systems. Models, like DistilBERT, have been discussed in depth; however, not many people may interact with such a network directly. As industries are continuing to deploy AI at scale, this section will briefly discuss real-world case studies on how model compression can be used on systems that are used daily by a multitude of people.

YouTube, the most well-known video-sharing platform, had an estimated 210 million viewers in the United States alone in 2022[92]. This metric underscores the immense volume of videos consumed daily. The content can be accessed in different ways, be it the search function, a shared video, or by browsing subscriptions. While all of these are still actively used, research has shown that recommendations are the most influential driver of views, with related video recommendations accounting for approximately 30% of the overall views[93]. AI models, which are behind recommender systems, can use vast amounts of data to process and recommend content. Which is why there have been challenges, like the YouTube-8M Video Understanding Challenge, in constructing a constrained system that can handle the massive amounts of content, although compressed for efficiency. Using quantization and replacing the 32-bit precision parameters with 16-bit precision, researchers have managed a 48.5% compression, along with having no loss of model accuracy[94].

[91] Ibid.
[92] Ceci 2023.
[93] R. Zhou et al. 2010, p. 6.
[94] T. Liu and B. Liu 2018, p. 1.

To further demonstrate examples of how model compression can be beneficial, this section will take another look at CNNs (see Section 3.1.3). The previously mentioned database ImageNet enables comparisons of various CNNs by their performance in classifying images. Image Recognition itself supports a wide range of practical applications, but most of society uses the smartphone extensively. CNNs are used for common image recognition use cases in smartphones, ranging from everyday convenience to more advanced features:

- **Face Unlock** is an everyday feature in modern smartphones, allowing for quick, convenient, and secure device access through facial recognition.
- **Augmented Reality (AR)** leverages image recognition to track surfaces and features, seen in applications like Pokémon GO, where virtual elements are anchored to the physical environment.
- **Photo Search** connects text to visual elements, enabling quick locating of specific pictures.
- **Magic Eraser** is a fairly new Google Pixel device feature, which has already been adopted by many devices, allowing users to remove unwanted objects from photos seamlessly through the use of image recognition and inpainting[95].

These features require models which are capable of recognising visuals, while being small due to the restricted storage space on phones. As it has already been examined, EfficientNet, an extraordinarily small model, managed state-of-the-art accuracy while being 8.4× smaller, as well as 6.1× faster, than competitors[96]. SqueezeNet, a different CNN, managed to compress its original model's size by a factor of 510, compared to AlexNet, to only 0.47MB, while still having the same Top-1 ImageNet Accuracy[97].

In summary, model compression techniques are already enabling AI at scale (YouTube) and on resource-constrained devices (smartphones) without sacrificing capability. However, efficiency isn't just impacted by model design – it also depends on the hardware it runs on and the software optimisations in use, which will be addressed in the following section.

3.4 Hardware and Software Optimisations

After the discussion of energy-efficient AI models and how to construct them, this section will go even deeper into the AI lifecycle. It will examine how additional efficiency can be squeezed out by deploying specialised hardware, implementing software optimisations, and how intelligent data-centre designs can further improve performance.

3.4.1 Specialised Hardware

Choosing the correct hardware impacts the performance of AI systems. Graphic Processing Units (GPUs) and Tensor Processing Units (TPUs) are far more efficient for AI workloads, due to their ability to perform massively parallel arithmetic operations. Google's first TPU managed to gain 30×-80× performance-per-watt compared to its counterpart GPUs and CPUs[98]. This section will examine components and hardware which have been specifically built for AI computation to improve energy efficiency. Due to their nature, CPUs are inherently inefficient at processing vast amounts of sequential data. While they do offer more resources than e.g. GPUs, their parallelism capabilities are limited and their power consumption is high[99]. As section 3.1.1 has mentioned,

[95] w.A. 2023.
[96] Tan and Le 2020, p. 1.
[97] Iandola et al. 2016, p. 7.
[98] Kaz Sato 2017.
[99] Dhilleswararao et al. 2022, p. 2.

computing inference in DNNs can be broken down into matrix computation. The Arithmetic Logic Unit (ALU) and reduced instruction set computer (RISC) architectures of CPUs are built for common operations, not large matrix multiplications (as they require multiple sequential steps)[100]. Some specialised hardware takes advantage of this fact and specifically builds upon this.

Graphics Processing Units (GPUs) based accelerators have been commonly used for training AI systems. By nature, their design is highly compatible with running parallel computation in terms of the number of cores and computation speed[101]. By leveraging large parallel cores and the Single Instruction Multiple Thread (SIMT) execution models, they are highly favoured for performing deep learning algorithms[102]. Especially NVIDIA GPUs and their CUDA technology manage up to 50×-150× speed-up when compared to an equal CPU-based implementation[103]. Although powerful and fast GPUs are still not specifically tailored for DNN applications, a major drawback being their high power consumption.

Application Specific Integrated Circuits (ASIC) are specifically tailored for DNN applications and offer high energy efficiency and computational performance[104]. Industries have started building custom ASICs to accelerate AI computation, like Eyeriss and Google Tensor Processing Unit (TPU).

- **Google's Tensor Processing Units (TPU)** is a type of ASIC which manages 15×-30× higher performance in DNN applications[105]. As has been mentioned, computing inference can be broken down into numerous matrix calculations. TPUs are customised to be outstanding at these operations. One of the reasons is that their instruction set is based on the complex instruction set computer (CISC), which allows for high-level instructions to complete complex tasks[106]. It includes the Matrix Multiplier Unit (MU), the Unified Buffer (UB) and Activation Unit (AC), which are all controlled by this instruction set. The MU in particular is designed to handle hundreds of thousands of operations in a single clock cycle[107]. Keeping the design minimalistic, straightforward and specific to DNN operations is what increases efficiency.

- **Eyeriss** is another ASIC accelerator useful for increasing CNN efficiency. It uses a row-stationary dataflow, which is proficient in data reusing, thus effectively minimising energy consumption[108]. Eyeriss consists of an array of 168 processing elements (12×14), on-chip feature map compression units, and a global buffer.[109]. The global buffer is the reason for data reuse. Its successor, Eyeriss v2, further enhances efficiency and throughput by using a hierarchical mesh Network-on-Chip (NoC) for better hardware utilisation and support for sparse neural networks[110]. This setup shows support for various CNNs (e.g AlexNet) and flexible configurations, as well as experimental results showing an 11.3× increase in energy efficiency and a 42.5× increase in throughput, compared to its v1 predecessor[111].

Field Programmable Gate Arrays (FPGAs) offer a more flexible alternative to ASICs. Due to their fixed nature, ASICs are extremely efficient for certain tasks but not easily adaptable,

[100] Kaz Sato 2017.
[101] Wang et al. 2020, p. 2.
[102] Dhilleswararao et al. 2022, p. 24.
[103] Ibid., p. 24.
[104] Ibid., p. 19.
[105] w.A. 2023.
[106] Ibid.
[107] Kaz Sato 2017.
[108] Dhilleswararao et al. 2022, p. 21.
[109] Ibid., p. 21.
[110] Ibid., p. 22.
[111] Ibid., p. 22.

whereas FPGAs trade some efficiency for flexibility. This is possible because FPGAs give hardware architects the choice to implement only the required logic in the hardware, as needed for the target application[112]. Their speed-up in computing is achieved by mapping to parallel hardware, i.e. several DNN models run in parallel[113]. FPGAs can be categorised into three types: for specific applications, specific algorithms, and accelerator frameworks with hardware templates[114]. While FPGAs usually cost less and are faster in production, due to their flexibility, they are significantly less performant and efficient than ASICs[115].

3.4.2 Software Optimisations

Software optimisations in AI frameworks and compilers have also contributed significantly to efficiency improvements. AI frameworks (e.g. TensorFlow, PyTorch), which offer building blocks for AI systems, have started implementing graph optimisations to remove redundant computations[116]. Techniques like mixed-precision training, where a lower precision (e.g. like quantization 3.3.2) can accelerate training and inference while using less energy, with negligible accuracy impact[117]. Using more efficient algorithms for matrix operations, caching intermediate results, and more are all contributing to increased efficiency. Choosing the proper hardware and software setup is what yields the best energy efficiency. This section will concisely review NVIDIA's TensorRT, which can optimise trained networks for faster and more efficient inference on specific hardware.

TensorRT is a proprietary software inference engine provided by NVIDIA designed to optimise neural network (NN) inference on edge devices[118]. The key components used in TensorRT's software optimisations are model compression, which uses discussed techniques like quantization (see 3.3.2) and parameter pruning (see 3.3.1) to achieve higher efficiency[119]. Following the model compression, it is then mapped onto hardware-specific kernels, specifically CUDA kernels in NVIDIA GPUs, to maximise computational efficiency and minimise inference latency[120]. With its optimisation steps, it managed a 23×-27× higher classification throughput, while maintaining or sometimes even gaining model accuracy[121]. However, these optimisations introduced some unexpected behaviours. Sometimes varying inference outputs across multiple compilations of the same NN model are produced, leading to inconsistent predictions for identical inputs[122]. Additionally, latency anomalies were observed where execution on more powerful hardware was slower compared to less powerful hardware[123]. This was attributed to differences in CUDA memory copying speeds and kernel execution times[124].

3.4.3 Energy-Efficient Data Centre Design

Beyond deploying efficient software and hardware, designing infrastructures with efficiency in mind is the next step in optimising AI. Innovations like advanced cooling or intelligent power management can be applied in modern data centres. For instance, raising the inlet temperature makes air cooling more efficient, a practice which is now mandatory by law in Germany due to

[112] Dhilleswararao et al. 2022, p. 8.
[113] Ibid., p. 8.
[114] Ibid., p. 8.
[115] Ibid., p. 7.
[116] w.A. 2024g.
[117] Micikevicius et al. 2018, p. 1.
[118] Ho et al. 2024.
[119] Shafi et al. 2021, p. 2.
[120] Ibid., p. 1.
[121] Ibid., p. 2.
[122] Ibid., p. 4.
[123] Ibid., p. 6.
[124] Ibid., p. 7.

the Energy Efficiency Act (see Section 4.1.4). Integrating energy reuse, such as capturing and reusing server heat, and making use of renewable energy will be important for reducing carbon footprint. Software can help too, by scheduling tasks efficiently to avoid peak thermal loads, and splitting the workload to less busy servers. Data Centre efficiency matters as they usually power AI operations and due to their high energy consumption. In 2009 they accounted for an estimated 2% of global energy consumption[125].

To counter the rising energy consumption, there are a few design innovations, namely a centralised energy-efficiency controller. Usually, a data centre controller splits resource control among various resource controllers. More often than not, these resource controllers in data centres can work against each other and sometimes counter their effects[126]. A centralised controller, by contrast, receives workload conditions from various resource controllers, makes a global decision by coordinating resources across the data centre, and then provides feedback to the individual controllers[127]. This feedback enables resource controllers to make clearer and more energy-efficient decisions[128].

Deploying specialised hardware for AI systems is an additional step towards increasing efficiency. Section 3.4.1 discussed TPUs and how their specific architecture can be used for DNN to increase performance. After the issue of energy consumption became apparent in 2013, Google designed, tested and deployed TPUs to their data centres in just 15 months[129].

Cooling is an important aspect to consider when designing data centres, as it can account for 38% of power consumption[130]. Implementing efficient cooling systems, like direct-to-chip cooling and evaporative cooling, is key, but recently, a new norm has emerged. Free Cooling, a fairly popular concept, holds the air inside data centres until it's hot and replaces it with fresh and cool outside air[131]. Depending on the region, this cooling method could completely cut energy costs, due to outside air always being cooler than the data centres[132]. Free Cooling is split into two techniques, direct and indirect free cooling. Direct free cooling draws outside air directly into the data centre and exhausts hot air back out[133]. However, concerns are air quality as usually filtration and mixing with recirculated warm air may be needed. A safer approach is indirect cooling, as the air is exchanged via a heat exchanger, which keeps indoor quality stable and avoids contamination[134]. Although the more prominent approach, the trade-off is slightly less efficiency, due to the added heat exchange step.

Efficiency in data centres can only be achieved through holistic consideration, encompassing hardware utilisation, data centre location, architectural design, and the integration of systems aimed at improving overall efficiency.

3.5 Training Efficiency and Resource Optimisation

Efficient deep neural network (DNN) training is crucial for reducing environmental impact and operational costs associated with AI systems. This section will discuss techniques which have emerged over the past years to improve training efficiency. They include techniques like transfer learning, which describes the process of starting from a pre-trained model instead of training from scratch, or early stopping, which stops training when sufficient accuracy is achieved, to avoid overtraining. Subsequently, a discussion on how to handle and allocate resources, where

[125] Shuja, Bilal, Madani, Othman, et al. 2016, p. 1.
[126] Ibid., p. 2.
[127] Ibid., p. 2.
[128] Ibid., p. 2.
[129] Kaz Sato 2017.
[130] Shuja, Bilal, Madani, Othman, et al. 2016, p. 6.
[131] Mukherjee et al. 2020, p. 2.
[132] Ibid., p. 2.
[133] Ibid., p. 13.
[134] Ibid., p. 14.

scheduling can yield efficiency gains, will follow. Finally, optimisation frameworks, such as Zeus, will be discussed, which address the environmental impact of training by balancing energy consumption and training performance through automatic configuration of training parameters.

3.5.1 Efficient Training Algorithms

Throughout this thesis, it has been examined how training different models can have extraordinary energy consumption. Particularly, section 2.1 mentioned how training the GPT-3 model required the same amount of energy as roughly 300 Austrian households in a year. As AI capabilities grew, researchers developed new training strategies to reduce the computational energy required.

- **Transfer learning** is a technique used to increase the target efficiency of models by transferring information from different domains[135]. It stems from the human capabilities of learning new hobbies more quickly than others, by being proficient in a similar domain[136]. An active handball goalkeeper may be more quickly accustomed to playing as a football goalkeeper, due to their muscle memory being used to catching, in comparison to someone with not sports background. This effect can be used in machine learning, as a classification algorithm for alike domains (such as digital camera reviews and food reviews), can boost the target inference accuracy by using less data[137]. This reduction in data consumption and training time can lead to a more efficient training of AI models.
- **Early stopping** is another training algorithm which has multiple upsides, such as the reduction of overfitting and increased training efficiency. Usually, during training, parameters and hyper-parameters are tuned to acquire the best performance, but researchers have found that performance and accuracy may plateau, or worse, even decrease, due to overfitting[138]. A simple, but effective way to combat overfitting is early stopping. By utilising different techniques to choose a fitting stopping criterion, one cannot only reduce overfitting, but also shorten training time, and effectively increase efficiency[139].

3.5.2 Resource Scheduling and Allocation in Date Centres

Large-scale training usually happens in data centres, where intelligent scheduling can improve efficiency and reduce the carbon footprint. In section 3.4.3, energy-efficient data centre designs were discussed. Task runs can be efficiently scheduled where renewable energy supply is high or when the grid carbon intensity is low. This process is called resource scheduling. While resource scheduling is also a common project management term, which dates back as early as 1960[140], it is also used to name the process of scheduling resources used in e.g. data centres. This can be very challenging, as there are multiple dimensions affecting operations, making it difficult to properly account for everything. Frameworks, like Data Centre-wide Energy-Efficient Resource (DCEERS), have been proposed, which schedule resources depending on throughput demand[141]. Data centres are modelled as a multi-commodity flow network, which enables the framework to calculate the minimal resource requirement for the current workflow[142]. Applying such intelligent resource scheduling increased the energy efficiency of data centres, thus reducing costs as well as environmental impact.

[135] Weiss et al. 2016, p. 1.
[136] Ibid., p. 2.
[137] Ibid., p. 2.
[138] Prechelt 1998, p. 1.
[139] Ibid., p. 1.
[140] Gordon and Tulip 1997, p. 1.
[141] Shuja, Bilal, Madani, and S. U. Khan 2014, p. 1.
[142] Ibid., p. 11.

Another technique, which hyperscale data centres use (i.e. large-scale, highly optimised and efficient data centres, see 4.3.1) is geographic load shifting. To reduce carbon emissions, hyperscale data centres can shift their workload between geographic locations, where one region has surplus green energy[143]. To achieve this, there are some considerations that have to be made, as to how flexible a workload is, i.e. if the input data is location locked, or if the hardware is specialised for such a workload, as well as if there are latency requirements[144]. Locational marginal carbon emissions have been proposed as an optimal measure for data centre load shifting. Researchers used this metric as a guide to develop an improved model that shifts loads independently of ISO collaboration, leading to a significant reduction in carbon emissions[145]. The model also shows that shifting greedily, such as only calculating for the next step and dismissing future steps, is not the best approach, as the current workload may have an impact on following calculations[146].

While scheduling can optimise the where and when of training, frameworks like Zeus optimise how training is executed on a given hardware, as discussed in the following section.

3.5.3 Optimisation Framework - Zeus

Section 3.4.2 introduced a proprietary Optimisation Framework, TensorRT, whose key features are increasing the efficiency of inference for AI models. In contrast, Zeus, an open-source optimisation framework, tries to uncover the best trade-off between energy consumption and performance for DNN training, by finding the best GPU configurations[147]. Zeus identifies key inefficiencies in common performance-oriented practices, such as using overly large batch sizes or running GPUs at maximum power, which often lead to disproportionately high energy consumption with diminishing returns in performance improvement[148]. It is an online, plug-in optimisation framework, which is specifically designed to enable users to find and navigate the Pareto frontier between energy and training time[149]. At its core, it includes:

- Dynamic adjusting of two primary parameters:
  - **Batch size** controls the number of samples processed per training iteration[150]
  - **GPU power limit** controls the energy usage trough dynamic voltage and frequency scaling[151]
- **Just-in-Time (JIT) Profiling** profiles energy and throughput characteristics in real-time without the need for extensive offline measurements, significantly reducing overhead and enhancing adaptability to workload changes[152].
- **Online Exploration-Exploitation Approach** utilises a Multi-Armed Bandit (MAB) algorithm (specifically Thompson Sampling) to systematically and efficiently explore optimal configurations, adapting to changing data conditions (e.g., data drift) and performance trade-offs[153].

By combining these strategies, Zeus reduces energy consumption by approximately 15.3%–75.8% across various workloads (see Figure 9), compared to traditional approaches that always maximise

[143]Lindberg et al. 2022, p. 1.
[144]Ibid., p. 2.
[145]Ibid., p. 6.
[146]Ibid., p. 6.
[147]You et al. 2023, p. 2.
[148]Ibid., p. 14.
[149]Ibid., p. 14.
[150]Ibid., p. 3.
[151]Ibid., p. 4.
[152]Ibid., p. 6.
[153]Ibid., p. 5.

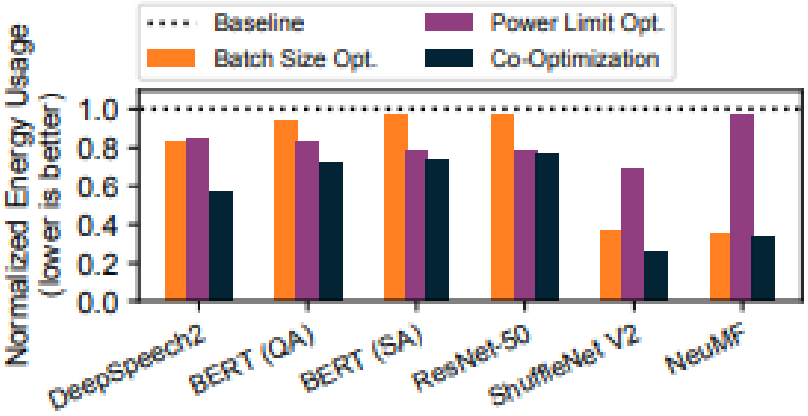

Fig. 9: This illustration shows us a baseline training energy consumption and how combining both batch size and power limit optimisations (as done by Zeus) can lower the consumption significantly in some cases. Source: You et al. 2023

throughput and GPU power usage[154]. Additionally, Zeus reveals a non-linear relationship between energy savings and training time increases, emphasising the nuanced balance between energy efficiency and computational performance[155]. In other words, Zeus finds sweet spots where a slight reduction in training speed dramatically cuts energy use, providing options for sustainable AI model training.

3.6 Discussion and Overview

This Chapter explores different techniques, frameworks, designs and optimisations, which can all be applied to achieve a more efficient AI lifecycle, to lessen the burden on the environment. To accurately compare the efficiency, it is essential to consider metrics which are consistent in different environments, including different configurations, which is why Green AI advocates for Floating Point Operations (FLOPs), as they are the least impacted by external factors[156]. Often, getting the best accuracy means disproportionally more computation, as Figure 4 shows. Section 3.2.2 examines the significantly smaller model, DistilBERT and how it manages to retain most of its accuracy through a technique called Knowledge Distillation (see Section 3.3.3). The goal is to find the sweet spot, as in most use cases, sacrificing too much efficiency for accuracy, or vice versa, is typically undesirable. Examining the environment, selecting the proper techniques and frameworks, depending on the context, is important to achieve truly sustainable AI. The technical strategies enable policies to require AI system developers and operators to design more efficient systems and data centres and limit their energy consumption. In the next part, the legal perspective will be examined, and a deep dive into the regulatory framework surrounding AI and incentives designed to support its sustainability will be taken.

[154] You et al. 2023, p. 3.
[155] Ibid., p. 2.
[156] Schwartz et al. 2019, p. 6.

4 Legal and Policy Frameworks for Sustainable AI

Following the exploration of technical innovations and environmental imperatives in the previous chapters, this chapter now turns to the legal and regulatory dimensions of sustainable AI. As mentioned in Section 1.2, sustainable AI cannot be achieved from one direction. It is a multidisciplinary approach, which must be supported by a framework of laws specifically tailored to it. Effective governance is crucial to ensure that AI development aligns with sustainability goals, both in terms of minimising environmental impact and upholding ethical standards.

This chapter explores the regulatory landscape surrounding AI with a focus on sustainable and ethical considerations. It will cover international and EU-level policies aimed at improving energy efficiency and reducing carbon emissions in the AI and data centre sector. Furthermore, existing and emerging regulations for AI will be surveyed and global regulatory approaches compared.

Furthermore, a deep dive into ethical considerations with legal implications will follow, addressing topics such as data privacy and protection, fairness, transparency, and accountability.

The focus then shifts to political incentives for local AI and small data centres. These centres, distributed across regions, can harness local renewable energy sources and feed their excess heat into district heating networks, reducing carbon footprint. Regulatory practices will be examined, which enable partnerships between data centres and sustainable energy providers, supported by targeted incentives and funding to establish a resilient, low-carbon AI ecosystem. The section concludes with a case study demonstrating successful waste heat reuse.

The final part highlights policy gaps which must be addressed to truly achieve sustainable AI. As noted in Section 3.2.1, a standardised metric is crucial for assessing sustainable AI, and this thesis advocates for its implementation. Furthermore, Section 4.4.3 will examine the challenges that come with developing a legal framework. It then moves to discuss effective enforcement, which must be applied to ensure compliance. Lastly, this chapter will plead for global coordination and equity concerning sustainability and AI governance.

4.1 Existing and Emerging Regulations for Sustainable AI

Despite the rapid growth of AI surprising lawmakers and some hurdles, a number of regulatory incentives have been introduced or proposed to guide AI toward sustainable and responsible outcomes. This section will cover international commitments that have been made and then progressively narrow the focus - first to the EU, and subsequently to Austria's and Germany's regulatory practices addressing sustainable AI.

4.1.1 International Commitments

Section 2.2 has introduced the Paris Agreement, a global pledge to combat climate change. To the 195 member states, it is a legally binding treaty to limit the temperature increase to 1.5°C[157]. While this treaty pressures the ICT sector to address its 5-9% of global electricity use and 2% of all emissions[158], there is no internationally binding treaty addressing sustainable AI. On the other side, leaders have acknowledged AI and its potential, including the dangers it can bring, and have presented the first-ever international legally binding ethics treaty. The Framework Convention on Artificial Intelligence aims to ensure that activities within the lifecycle of AI are aligned with human rights, democracy and the rule of law, while not halting its technological progress and innovation[159]. It requires members to implement principles which AI systems must follow:

[157] w.A. 2024h.
[158] Butler 2023.
[159] Rotenberg 2025, p. 5.

- Each member must implement measures that require AI systems to respect **human dignity and individual autonomy**[160].
- Each member must implement measures for **transparency and oversight** in AI systems[161].
- Each member must implement measures that ensure **accountability and responsibility** for wrongdoings[162].
- Each member must implement measures that require AI Systems to respect **equality and non-discrimination**[163].
- Each member must implement measures for **privacy and personal data protection** in AI systems[164].
- Each member must implement measures which promote **reliability** from AI systems[165].
- Each member must implement measures which foster **safe innovation** for AI systems[166].

In addition, a few soft laws have emerged, a notable one being the UNESCO Recommendation on the Ethics of AI (2021), applicable to all 194 member states[167]. Section 4.2 will go more in depth about legal frameworks regarding ethical considerations.

In summary, globally there is convergence in high-level principles, e.g. ethics and sustainability highlighted by UNESCO and such, but divergence in implementation, especially regarding sustainability. No global authority exists, but international pressure and examples, like the EU, are influencing national policies.

4.1.2 European Union: AI Act, Energy Efficiency and Green Initiatives

Having discussed global principles, which still miss an established international treaty that addresses both ethical and environmental concerns, the EU is a pioneer and has taken a leading role in regulating AI.

Its most significant development is the **Artificial Intelligence Act** (AI Act), introduced in 2024, which aims to establish a comprehensive regulatory framework for the development and deployment of AI in the European Union. Its purpose is the upholding of the EU's values, by providing a number of requirements including transparency and energy consumption, while supporting innovation[168]. The AI Act defines AI systems as machine-based systems that can operate with varying levels of autonomy[169] and general-purpose AI (GPAI) models as AI models, which have been trained using a large amount of data combined with self-supervision at scale, thus being capable of performing a wide range of general tasks[170]. These systems are classified by their risk, the highest being unacceptable risk (e.g. manipulative, deceptive, etc.), followed by high-risk, then limited risk and lastly minimal risk. Most of the Act regulates high-risk AI systems, which refers to systems deployed in critical areas such as education, employment, and infrastructure[171]. A provider, referring to a body developing and offering AI systems or

[160] *Framework Convention on AI* 2024, Article 7.
[161] Ibid., Article 8.
[162] Ibid., Article 9.
[163] Ibid., Article 10.
[164] Ibid., Article 11.
[165] Ibid., Article 12.
[166] Ibid., Article 13.
[167] w.A. 2024d.
[168] AI Act, Recital 1.
[169] AI Act, Article 3(1).
[170] AI Act, Article 3(66).
[171] AI Act, Annex III.

GPAI[172], is obliged to several rules, whereas high-risk systems have additional rules. They must include technical documentation for the AI system, which must include a breakdown of its energy consumption. If the energy consumption is not known, then it may be based on computational resources used[173]. This documentation is to be kept up to date, as the European Artificial Intelligence Office (AI Office) may request it[174]; however, GPAI models launched before the 2nd of August 2024 have a 2-year grace period and are currently exempt from this rule[175]. Additionally, some AI systems are classified as 'systemic risk', referring to GPAI models with broader reach or potential negative impact on society[176]. One such impact is energy consumption, creating an incentive for providers to minimise it in order to avoid additional obligations[177]. Article 40 directs EU standardisation bodies to produce technical standards on reducing energy and other resource consumption of high-risk AI systems and on energy-efficient design of AI models[178]. However, these are not yet implemented and may only be voluntary.

Beyond AI specific laws, the European Green Deal, launched in 2019, aims to transform the EU into a modern, resource-efficient and competitive economy[179]. It legally binds the 2050 carbon neutrality and promises to cut at least 50-55% of emissions[180]. It has more ambitious plans for climate neutrality, which include investing in clean technology and green infrastructure[181]. While it does not directly address AI, the investments may be used as an incentive for implementing sustainable AI practices or developing more efficient algorithms.

Furthermore, the 2023 revised **Energy Efficiency Directive** (EED) addresses the broader ICT sector, which operates AI. In 2018, the total energy consumption of data centres was 76.8 TWh, a figure expected to rise by 28% to 98.5 TWh by 2030. To counter this trend, the EED urges member states to mandate the collection and publication of data relevant to energy performance, particularly from data centres with a significant footprint, where a design upgrade can increase efficiency[182].

In summary, the EU is building a multi-layered framework to regulate AI. The AI Act shows us the first approach to oversee and govern systems. Although it does acknowledge the environmental impact of AI and introduces obligations for providers to be mindful of its efficiency, it still lacks technical requirements and standardisations for AI. To properly regulate and require sustainability in AI, these frameworks must introduce obligatory conditions which address these. This, however, might be a hurdle to innovation and create an incentive for providers to move to a less stringent environment, slowing AI progress within the EU. Hardships and Policy Gaps such as these will be discussed more in depth in Section 4.4. While the AI Act is not yet fully developed, mostly lacking in mandatory requirements regarding sustainability, it represents a crucial stepping stone for member states to follow.

4.1.3 Austria's Regulatory Framework for AI

Building on the EU framework, Austria has implemented its own efficiency act. The country stands out with a very clean electricity grid, with about 87% of Austria's electricity coming from renewable sources[183]. Clean Energy can thus be used for powering AI, making operations low-carbon, but there's still the issue of using energy efficiently. Austria's climate and energy

[172] AI Act, Article 3(3).
[173] AI Act, Annex XI, Section 1(2)(e).
[174] AI Act, Recital 101.
[175] AI Act, Article 111(3).
[176] AI Act, Article 3(65).
[177] AI Act, Article 51(2).
[178] AI Act, Article 40.
[179] *European Green Deal* 2019, p. 2.
[180] Ibid., p. 4.
[181] Ibid., pp. 5–6.
[182] EED, Recital 85.
[183] Burger 2023.

strategy aims for carbon neutrality by 2040, ahead of the EU's 2050 goal, meaning all sectors need to optimise energy use. This section will dive deeper into Austrian laws and strategies in the context of AI and how it tackles energy efficiency.

In the year 2021, Austria launched the *AIM AT 2030* mission, a national strategy addressing AI, shaping the path it will take[184]. At its core, it revolves around three strategic goals:

1. A deployment of AI for the common good is pursued, based on fundamentals and human rights, the current and forthcoming European values[185].
2. Austria shall position itself as a hub for artificial intelligence research[186].
3. Through the development and deployment of AI, the competitiveness of Austria's technology and business shall be secured[187].

The mission is designed to be agile, allowing for ongoing changes and refinements. At present, 13 fields of action (see Figure 10) and additionally 11 specific application areas (see Figure 11) have been defined, forming the two foundational pillars: **Trustworthy AI** and **AI ecosystem**. A total of 91 measures are already in planning or put in action[188]. One of the focus points is climate neutrality and sustainability through AI. The mission addresses current world climate concerns and acknowledges that AI is part of it. To achieve a climate-neutral Austria by 2040, it proposes a *Twin Transition*, the idea being digitisation and sustainability working in tandem, instead of as opposing forces[189]. The implementation plan for the years 2024-2026 includes 47 measures, ranging from funding initiatives such as AI for Tech, AI for Green, and AI for Transformation, to raising the potential of AI getting directly involved in the federal ministry, as well as the development of a guideline on Green AI[190].

Furthermore, Austria's Energy Efficiency Act regulates energy efficiency in diverse areas, one of them being data centres. It does not enforce any efficiency regulations, such as power usage effectiveness (PUE) or waste heat reuse requirements; however, data centre operators are required to provide thorough information, including efficiency in energy, electricity usage, and temperature[191].

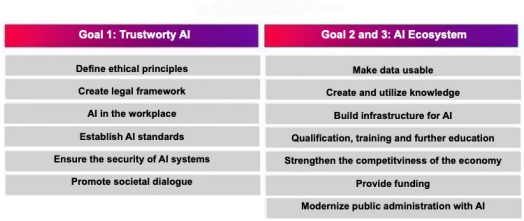

Fig. 10: The thirteen fields of action forming the foundational pillars, Trustworthy AI and AI Ecosystem, of Austria's AIM AT 2030 Mission. Source: w.A. 2024e, p. 10 – translated from German.

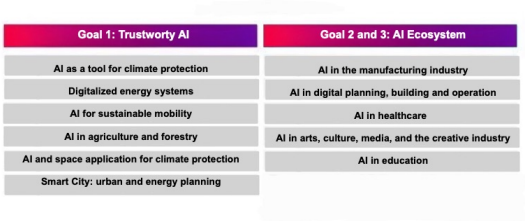

Fig. 11: The eleven application areas forming the foundational pillars, Trustworthy AI and AI Ecosystem, of Austria's AIM AT 2030 Mission. Source: w.A. 2024e, p. 10 – translated from German.

[184] w.A. 2024e, p. 4.
[185] Ibid., p. 10.
[186] Ibid., p. 10.
[187] Ibid., p. 10.
[188] Ibid., p. 4.
[189] Ibid., pp. 36–37.
[190] Ibid., p. 45.
[191] *EEffG* 2024, Section 72a.

4.1.4 Germany's Energy Efficiency Act (Energieeffizienzgesetz - EnEfG)

In contrast to the Austrian Energy Efficiency Act, Germany's Energy Efficiency Act (EnEfG) has tighter policies, directly targeting data centres and their climate neutrality. Existing data centres have to improve their Power Usage Effectiveness (PUE) over time, with 1.0 being the ideal[192]. In July 2027 it must be ≤ 1.5 PUE[193], tightening to ≤ 1.3 PUE in July 2030[194]. Data centres which are to be deployed after July 2026 must present a PUE of ≤ 1.2. Furthermore, they must prove an energy reuse of at least 10%, increasing to 15% for data centres deployed in 2027 and 20% for the same in 2028. These targets essentially force data centres to improve their energy efficiency, implement better cooling designs or reuse their wasted heat. Starting in 2024, at least 50% of a data centre's electricity must come from renewable sources[195], and by 2027 it must be 100% renewable[196]. It is important to note that this can also be achieved through direct purchases of green electricity, renewable energy certificates or by generating renewable energy themselves. Additionally, data centre operators must implement an energy and environment management system by July 2025[197]. This includes the continuous monitoring of energy consumption[198] and seeking of energy efficiency improvements[199]. Additionally, they must report key metrics to the government[200], which will be stored in an energy efficiency registry and made public[201]. They even have to inform customers about their energy consumption if they use services of a data centre[202].

Compared to Austria's Energy Efficiency Act, these policies are much stricter and proactive. However, through the enforcement of efficiency, they push for sustainable data centres and innovation to help achieve these targets. The transparency policies create accountability and let customers and regulators see which facilities are efficient, thereby further promoting sustainability. Any AI service running in a German data centre will indirectly be subject to these efficiency and green energy rules. Such measures often drive technical innovation and adaptation, as they must meet sustainability requirements. Technical adaptation is often intertwined with ethical questions. The following section takes a closer look at ethical considerations in AI and examines how the regulatory framework, particularly the AI Act, seeks to address them.

4.2 Ethical Considerations with Legal Implications

Ensuring AI is sustainable goes beyond environmental efficiency, as ethical considerations are just as important. Many ethical issues surrounding AI have direct legal implications. In fact, the push for *Trustworthy AI* and *Responsible AI* intersects with law. This section discusses these issues and how they are being addressed through legal and regulatory measures. First, it will emphasise the importance of data privacy and protection and examine regulations addressing them. Afterwards, it will examine ethical questions and regulations of fairness and bias in AI. Finally, the importance of transparency is addressed and the section concludes by examining accountability regulations.

[192] *EnEfG* 2023, Article 3(15).
[193] Ibid., Article 11(1)(1.)
[194] Ibid., Article 11(1)(2.)
[195] Ibid., Article 11(5)(1.)
[196] Ibid., Article 11(5)(2.)
[197] Ibid., Article 12(1).
[198] Ibid., Article 12(2)(1.)
[199] Ibid., Article 12(2)(2.)
[200] Ibid., Article 13.
[201] Ibid., Article 14.
[202] Ibid., Article 15.

4.2.1 Data Privacy and Protection

The century-old fight for privacy, illustrated in 12, dates to 1890, when Samuel D. Warren II and Louis Brandeis introduced the concept of the **Right to Privacy**, defining it as "the right to be left alone"[203]. More than a hundred years later, this principle faces new challenges, as our personal data is easily collected through our browsing habits and device tracking, creating an online fingerprint. This can and will be used for targeted advertisement, location-based offers or more malicious intent, like identity theft. A countermeasure is the concept of data privacy, restricting which data can be collected and data protection describing security measures that are to be taken in order to restrict important data access. Section 2.1 outlines how AI consumes an increasing amount of data, which usually includes personal data and habits, raising once again the question of privacy and protection.

One of the fundamental principles of The Framework Convention on Artificial Intelligence, is the respect for personal privacy and data protection. Each signing party must thus implement or maintain measures, which ensure that the privacy rights of users of AI systems and their data are protected[204]. Within the EU, the AI Act works alongside the General Data Protection Regulation (GDPR) to regulate and protect personal data. The GDPR mandates how organisations and businesses have to handle personal data or any information which may identify a person[205]. The principles of data minimisation and storage limitation also promote sustainability, as they permit data processing only when necessary for a defined purpose and require that data be stored in a limited, often compressed form[206]. Additionally, AI designers are guided towards privacy-preserving techniques, as the processing of personal data is only allowed with consent[207]. This also means respecting fundamental rights such as the right to data access[208] and right of erasure[209]. A notable law enforcement concerning privacy was Italy banning ChatGPT in 2023 over privacy concerns, underscoring the need for privacy consideration for AI development[210]. In a sustainable AI context, strong privacy laws ensure that the drive for more data is balanced against fundamental rights, encouraging more efficient and ethical data practices.

4.2.2 Fairness, Bias and Non-Discrimination

To fully achieve sustainable and ethical AI, it must be avoided replicating or amplifying biases that lead to unfair outcomes for certain groups. When using AI for tasks such as hiring, it must not make decisions on attributes such as race or gender. An unfair AI affects the sustainable adaptation of AI. This is not only an ethical issue, but also a legal one, as many jurisdictions have anti-discrimination laws. This section will discuss the existing legal framework, but also dive deeper into the issue of fairness in AI.

Again, the Framework Convention on AI addresses this, by requiring each signing party to develop measurements for equality and against discrimination, within the lifecycle of AI[211]. The EU AI Act requires that training, validation, and testing datasets used for high-risk AI systems must be examined for biases that could affect fundamental rights, and that appropriate measures be taken to detect, prevent, and mitigate such biases[212]. A regulatory foundation is built to address the problem of fairness, but the real challenge is that bias in AI is subtle and hard to detect.

[203] Warren and Brandeis 1890.
[204] *Framework Convention on AI* 2024, Article 11.
[205] GDPR, Article 4(1).
[206] GDPR, Article 5(1)(c) and (e).
[207] GDPR, Article 6(1)(a).
[208] GDPR, Article 15.
[209] GDPR, Article 17.
[210] McCall 2023.
[211] *Framework Convention on AI* 2024, Article 10.
[212] AI Act, Article 10(2)(f) and (g).

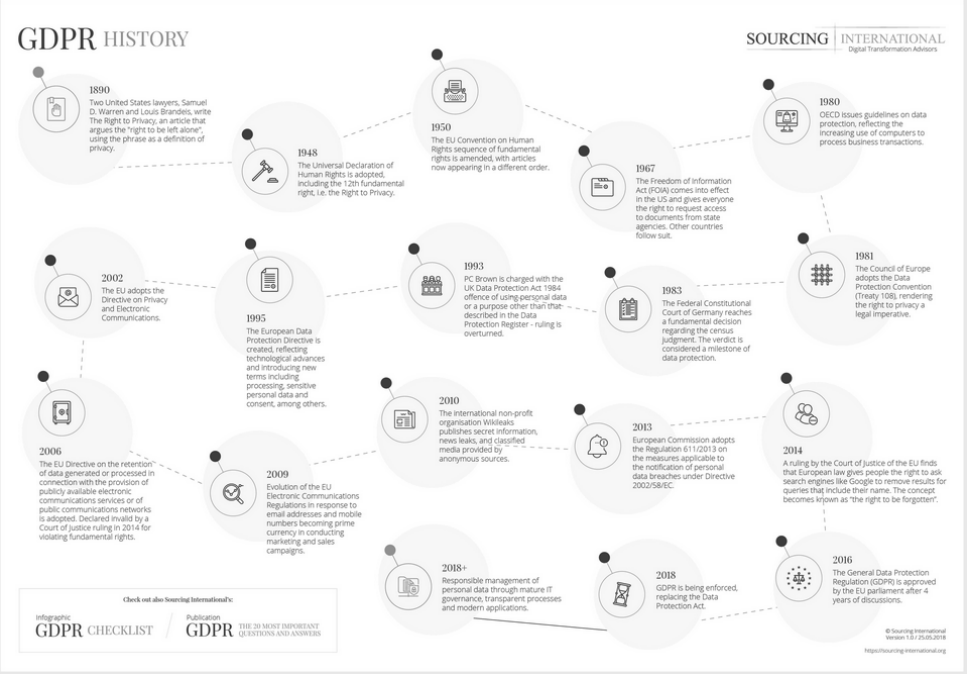

Fig. 12: This figure illustrates how data privacy and protection has evolved over time. It started in 1890 with the *the right to be left alone*, leading to the creation of the GDPR in 2016. Source: w.A. 2018

Current "fairness" algorithms in AI are critiqued as they rely on institutional categories (e.g. race, gender and age) and fairness metrics (e.g. demographic parity and equal opportunity)[213]. While these are useful, they also have limitations, as they only address data, which in turn might be biased and ignore real-world issues. As this data is no more than human construction, it also integrates existing power dynamics. Furthermore, these metrics risk oversimplification as they are usually too broad to capture real diversity, missing social and contextual issues. Additionally, the problem of intra- and inter-category inequality persists, as correcting unfairness within protected groups might disadvantage less privileged individuals in these groups[214]. To combat this, researchers have proposed technical fixes such as fair representations. The model is supposedly missing the context of sensitive variables (e.g. race, gender and age), while still being highly predictive of the target attribute[215]. Other scholars suggest more radical approaches. For example, by reshaping reality tests, procedures (e.g., hiring or scoring) that grant or deny opportunities, bias can be challenged at its root. Yet this increases opacity, as only experts can redefine these tests, shifting power away from society[216]. Instead of yet another algorithm, researchers propose creating critical spaces, where fairness categories themselves can be questioned and redesigned, giving power to society[217].

Another big issue with fairness and AI is when situations arise that have no correct answer. This has been a big debate in the self-driving cars department. Imagine a scenario in which a vehicle suddenly cuts in front of an autonomous vehicle, leaving the only option to avoid a collision being to swerve into a wall and potentially endangering the AV's own driver. Which action should the AV take? This comes down to the infamous trolley problem, where you either harm

[213] John-Mathews et al. 2022, p. 6.
[214] Ibid., p. 7.
[215] Ibid., p. 8.
[216] Ibid., p. 9.
[217] Ibid., p. 12.

one or the other party. Current legal frameworks provide no definitive guidance, highlighting that fairness is not always a technical question but ultimately a societal and legislative one.

4.2.3 Transparency and Explainability

Over the years, the word transparency has evolved to multiple meanings. At first, it was known as the physical property of objects which allows light to pass through them. Law and regulation have given it a new meaning in the 1970s as they gave the public more access to governmental information, requiring a more open and transparent governance[218]. It has been a recurring ethical demand for AI to be transparent and explainable. This meant that users must understand "when" they are affected by AI and the right to know the "why" behind the outcome.

The EU AI Act sets the regulatory basis and addresses transparency thoroughly and requires numerous pieces of information to be accessible and interpretable. High-risk AI systems must have a sufficiently transparent output, as to ensure that their information output is interpretable and usable by deployers[219]. These systems are also required to include a complete, correct and concise instruction sheet[220], which contains information such as the system's capabilities[221] and the resources needed for operation[222]. Article 50 goes in depth about transparency obligations surrounding the interaction of users with AI systems. Providers must ensure natural persons are informed that they are interacting with an AI system if it is not already obvious[223]. The generation of audio, image, text, and video content by AI must be marked in a machine-readable format and detectable as AI-generated[224]. This addresses a big concern of AI videos or text flooding the Internet, and fueling the spread of misinformation that is difficult to distinguish from authentic content. In addition to transparency regulations in Article 50, any affected person by the decision from a high-risk AI system has the *right to explanation*, as to why this decision was made[225].

The ethical goal is to avoid "black box" AI, where no one understands the decisions made, which is unsustainable from a trust perspective. It is also important to highlight the trade-off of too much transparency, as it can backfire in terms of privacy. By embedding explainability into law, developers are required to lower the complexity barrier, not only making AI decisions more interpretable, but sometimes leaner. Less complexity also indirectly translates to resource intensity (see Section 2.4), thus to more energy consumption. This means the push for explainability promotes more efficient AI architectures, aligning ethical and environmental sustainability goals.

4.2.4 Accountability and Governance

Ethically, deploying AI requires governance structures that oversee its use, mitigate harm, and hold systems accountable. Legally, a regulatory basis is emerging in the form of compliance and oversight. In this section, the legal ground will again be explored, following AI-specific challenges that complicate the regulation. Before examining the legal mechanisms for AI accountability, it is necessary to define what accountability means in this context.

At its core, accountability means that you, the agent, must explain and justify your actions to someone, the forum, who has the right to judge you. The forum then decides if you have acted properly. Neither the agent, nor the forum, must be a single person - they could be groups,

[218] Ball 2009, p. 1.
[219] AI Act, Article 13(1).
[220] AI Act, Article 13(2).
[221] AI Act, Article 13 (3)(b).
[222] AI Act, 13 (3)(e).
[223] AI Act, Article 50(1).
[224] AI Act, Article 50(2).
[225] AI Act, Article 86.

organisations, or governments. This usually happens when someone, the principal, gives the agent a certain responsibility or task[226]. The accountability given has three key parts:

1. **Authority Recognitions:** The agent accepts the task they have been given, by the principal. The principal accepts that the agent has authority. The agent accepts that the forum can judge them[227].

2. **Interrogation:** The forum must be able to question and examine the Agents actions. Otherwise, it collapses to self judgement[228].

3. **Limitation of power:** The forum monitors, evaluates and constrains the agent, meaning it cannot do whatever they want[229].

This defines accountability as more than just being responsible, but as a structured relationship that keeps powers in check and ensures transparency. Only when all three conditions are met, does accountability work.

A key challenge of accountability in AI is that its outcomes are often opaque and inherently unpredictable. If an AI system were to infringe on a fundamental right, pinpointing individual responsibility becomes difficult, as there are numerous points in the process where things could go wrong, similar to the many hands problem. This infringement could arise from biased training data, bugs, programmer errors or misuse, making it almost impossible to tell where exactly it went wrong. Additionally, multiple forums with conflicting judgements raise the symmetrical problem of many eyes, where either no one is accountable or there is an accountability surplus[230].

Legally, the rule of authority recognition is already at the core of democracy. Society, the principal, has given the EU (or the AI Office), a forum, the right to judge providers (or AI operators), an agent. To satisfy the second condition of accountability, the AI Act requires a certain transparency to be upheld by AI systems, as discussed in the previous Section 4.2.3. To be able to question and examine an AI system's actions, the Act requires high-risk AI to allow for automatic record-keeping of relevant events[231]. Additionally, AI systems must be developed in a way that allows human controllers to oversee its lifecycle[232]. Human oversight must be able to understand the system's capabilities and detect anomalies[233], as well as have the ability to override the output[234] or completely stop a high-risk AI system[235]. The Act establishes the AI Office[236] to supervise AI providers, alongside national competent authorities tasked with market surveillance and conformity assessments. This structure constrains providers by designating oversight actors with enforcement capacity, satisfying the third and last condition.

While the AI Act addresses the many eyes problem by centralising oversight in the AI Office (supported by national authorities), it still primarily makes providers and deployers liable for wrongdoings. What it does not yet resolve is the question of civil liability — who compensates victims if an AI system causes harm, or the many hands problem. This gap has prompted the EU to propose the AI Liability Directive, which is still being developed. Only with such frameworks in place can accountability be considered complete, balancing compliance duties with enforceable avenues for redress.

[226] Novelli et al. 2024, p. 2.
[227] Ibid., pp. 2–3.
[228] Ibid., p. 3.
[229] Ibid., p. 3.
[230] Ibid., p. 5.
[231] AI Act, Article 12.
[232] AI Act, Article 14.
[233] AI Act, Article 14(4)(b).
[234] AI Act, Article 14(4)(d).
[235] AI Act, Article 14(4)(e).
[236] AI Act, Article 64.

4.3 Local AI and Small Data Centres

A notable policy consideration is whether to consider moving data centres from large hyper-scaled data centres to smaller decentralised data centres hosting local AI. Section 3.5.2 discussed how energy-efficient data centre design leverages renewable energy to reduce its carbon footprint. They can be relocated and distributed across multiple decentralised sites powered by renewables, while additionally making use of their wasted heat. This section examines the pros and cons of decentralised and hyper-scaled data centres, then explores EU and national incentives and funding to foster local AI ecosystems. Furthermore, laws which enforce renewable grid integration and foster decentralisation of data centres will be looked into, and finally, a successful case study is examined.

4.3.1 Decentralisation vs. Hyperscale

Hyperscale data centres are the primary infrastructure powering AI, and globally there are over 1000 currently deployed[237]. It refers to massive data centres run by cloud providers, such as AWS, Microsoft, or Google, designed to scale to thousands of servers. They can handle intensive global operations, e.g. AI training, with enormous compute power and storage capacity. Their biggest advantage is their central nature, as resource handling is highly optimised, thus resulting in efficient and robust processing[238]. With abundant resources at their disposal, hyperscale data centres are both highly scalable and reliably resilient. On the downside, their disadvantages include latency delays caused by constant cloud round-trips and a concentrated energy demand[239].

In contrast, decentralised or local data centres distribute resources across different smaller sites. Their key benefits include:

- **Real-Time AI Inference at the Edge:** AI infrastructure can be deployed in locations, where milliseconds matter and latency can have devastating effects[240].
- **Distributed Model Training:** To improve cost efficiency training can be distributed across sites, which reduced bottlenecks of the central cloud hub[241].
- **Built-In Data Privacy:** Sensitive data can be processed on location, increasing security and privacy[242].
- **Resiliency & Redundancy:** As infrastructure is spread across multiple locations, having a single site fallout is not a problem any more, as fault tolerance and geographic redundancy is gained[243].

Decentralised data centres, especially those equipped with on-site energy generation and local energy storage, can reduce stress on the electrical grid. By operating on renewable sources, they can be located in regions such as Austria, where 87% of electricity already comes from renewables[244]. However, they also bring disadvantages, like operational complexity, as managing a diverse fleet of servers across locations and different environments is challenging, especially at scale. Security risks are another issue as there are multiple nodes of entry, and processing large resources can be quite inefficient, which defeats their purpose[245]. Indeed, hyperscale centres often

[237] w.A. 2025k.
[238] Cao 2022, p. 8.
[239] w.A. 2025d.
[240] Ibid.
[241] Ibid.
[242] Ibid.
[243] Ibid.
[244] Burger 2023.
[245] w.A. 2025d.

Aspect	CeAI	DeAI
Methodology	Top–down, holism, authority, and autocracy	P2P, bottom–up, reductionism, autonomy, and democracy
Objective	Global objectives	Local objectives
Intelligence	General and strong intelligence	Local, edge and weak intelligence
Task	Global and central task	Local and distributed task
Data/repository	Central and single resourcing and storage	Local, distributed, multiple, end resourcing and storage
Model	Central and global model	End, and local model
Architecture	Central, vertical and hierarchical control, mediation, matchmaking, gateway, and server/client structure	Horizontal, P2P, D2D, distributed, chain, and flat structure
Process	Consensus-building, aggregation, and orchestration	Partition, decomposition and splitting
Mechanism	Predefined, vertical, design-time alliance, coordination, cooperation, normalization, and standardization	*Ad hoc*, horizontal, run-time self-motivation and organization, and negotiation
Computation	Central, global computing infrastructure	Distributed, local computing systems
Communication	Broadcasting and hierarchical	P2P, D2D
Decision	Global goal-driven authority, and strategic minority	Local goal-driven personalization and majority voting
Output	Global, aggregated, and integrative	Local, individual, and personalized
Privacy	Weak protection	Strong protection
Security	Central authorization, monitoring, risk, and governance	Local or distributed authorization, monitoring, and risk
Pros	Unification, capacity, order, efficiency, stability, strategic, concentrated resources, computing, and data	Personalization, flexibility, adaptivity, resilience, transparency, autonomy, expandability, high fault tolerance, scalability, throughout, low cost, and risk
Cons	Low flexibility, adaptivity, autonomy, robustness, reliability, accountability, fault tolerance, high vulnerability, risk, cost, and catastrophic mistake	Low capacity, energy, resource, computation, and stability, high latency, and chaos

Fig. 13: This table shows us a direct comparison of centralised AI and decentralised AI, highlighting their strengths and weaknesses. Source: Cao 2022, p. 8

achieve very low PUE ratios and scale very well in energy usage, which local centres may struggle to match. This illustrates that sustainable AI infrastructure may require a hybrid approach: leveraging hyperscale efficiencies where appropriate, while using local centres for latency-critical or energy-sharing opportunities. However, realising these benefits of decentralisation requires supportive policies and incentives, as the next sections discuss.

4.3.2 Incentives and Funding

The previous section has examined how decentralised and local data centres are beneficial for certain applications. Deploying and running a data centre requires a lot of knowledge and resources, which is why this section will discuss incentives and funding for data centres and AI.

EuroCloud's mission, a pan-European innovation hub, is a knowledge-sharing network between European countries to ease the entry to cloud computing, research centres, or start-ups. Their communication between partners is open, in order to bring IT and businesses together, thus fostering a *European Digital Single Market*. By partnering with the EU and local government, they encourage the development and growth of cloud infrastructure. EuroCloud's orientation, guidance, and best practice delivery helps create secure, standards-compliant services, thus creating incentives for data centre projects[246]. European countries have developed their own incentives for cloud services, such as Austria's O-Cloud Initiative, which is supported by EuroCloud.

The EU's Digital Europe Programme **DIGITAL** is a funding programme aimed at supporting the green and digital transformation of infrastructures. The programme offers support in areas such as supercomputing, cybersecurity and artificial intelligence, helping the development of digital technologies in the EU. However, DIGITAL is not the only funding programme focused on

[246] w.A. n.d.

digital innovation. It is complemented by several others that aim to enhance the EU's industrial competitiveness and reinforce its sovereignty. Together, they provide potential funding pathways for AI and data centre projects focusing on sustainability and innovation[247].

Austrian AI Infrastructure **AI:AT** is a nationwide incentive that aims to strengthen Austria's digital competitiveness and sustainable AI ecosystem. It aims to build a nationwide AI hub by linking research, business, and public institutions, with state-of-the-art infrastructure targeting local businesses. This initiative provides institutional support and could serve as a catalyst for decentralised AI infrastructure development[248].

4.3.3 Energy Law and Grid Support

Energy law and smart grid infrastructure policies that enable data centres to reuse their wasted heat or enable better integration of renewable energy play a decisive role. This section will discuss how Austria and Germany envision a sustainable grid infrastructure and examine their existing policies and energy law supporting this.

The Austrian Elektrizitätswirtschaftsgesetz (ElWG) aims to modernise its electricity market by supporting decentralised, sustainable and consumer-centric energy systems. It intends to implement the EU's internal Electricity Market Directive, Renewable Energy Directive and Energy Efficiency Directive into Austrian legislation. The draft recognises peer-to-peer contracts, allowing data centres to buy renewable energy directly from a producer[249]. Additionally, producers are allowed to construct direct electricity transmission lines, so data centres can directly leverage clean energy[250]. In combination with the recognition of renewable energy communities[251], these policies empower decentralised actors such as data centre operators and enable more flexibility to go green. Furthermore, providers may refuse grid access due to lack of capacity or disturbances in the grid[252]. If full access cannot be granted, the operator must offer flexible or limited access[253]. This creates a legal framework for versatile grid access, promoting once again renewable energy integration.

While the ElWG does address decentralisation and data centres, it is still a draft, whereas the German EnEfG, examined in Section 4.1.4, has already entered into force. It does not directly address decentralised data centres and local AI, but it fosters the development of such through various policies. Companies, including data centres, must avoid waste heat as far as reasonably possible[254]. This includes reusing it inside their premises or piping it to third parties[255]. Data centres in particular are required to utilise a portion of their energy/wasted heat. As it has been mentioned in Section 4.1.4, newer data centres must reuse at least 10% of their energy, rising to 20% for data centres deployed in 2028[256]. Furthermore, data centres are required to implement an energy and environmental management system, forcing them to:

1. continuously measure the performance and energy requirements of its components[257].
2. take action to continuously improve its energy efficiency[258].

[247] w.A. 2025b.
[248] w.A. 2024e.
[249] *ElWG (Ministerialentwurf)* 2025, Article 62.
[250] Ibid., Article 59.
[251] Ibid., Article 64.
[252] Ibid., Article 95.
[253] Ibid., Article 96.
[254] *EnEfG* 2023, Article 16(1).
[255] Ibid., Article 16(1).
[256] Ibid., Article 11(2)(2.)
[257] Ibid., Article 12(2)(1.)
[258] Ibid., Article 12(2)(2.)

However, data centres are exempt from implementing this system if they show 50% of their energy reuse integrated into a local grid system[259]. These policies encourage operators to capture the heat their servers generate and supply it to nearby homes or businesses, indirectly fostering decentralised data centres. By integrating data centres into local energy systems, policies support a symbiotic relationship: the data centre gets clean power and the community gets excess heat/energy or grid stability services. Following this section will be a successful case study that shows waste heat reuse in practice.

4.3.4 Case Study - Digital Reality

Digital Reality is a sustainability project in Vienna focusing on converting its wasted heat into usable energy. They partnered with a clinic to reuse their data centre's wasted heat and supply the building with it. In combination with the reduced heating costs of the clinic and the reused heat of the data centre, they managed to save 4,000 tons of CO_2, yielding sustainable profit for both[260]. Their project consists of four steps:

1. Capture the wasted heat of their server[261].
2. Lead the heat through a heat exchanger, bringing it to a useable temperature[262].
3. Integrate the heat into a local energy grid[263].
4. Reuse the heat, such as heating for a building[264].

While they acknowledge the difficulty, they list advantages such as energy efficiency, cost reduction, carbon footprint reduction and sustainability, and appeal for similar laws, as presented in the EnEfG, forcing such a practice by law[265]. This case study shows us a successful implementation of waste heat reusing and additionally highlights missing policies in Austrian law, such as the missing heat reuse policy in Germany's EnEfG (see 4.1.4). The next section delves deeper into policy gaps, focusing on shortcomings in enforcement and compliance, while also examining possible future directions and challenges for legal frameworks.

4.4 Policy Gaps and Future Directions

The previous sections discussed the regulatory framework addressing AI and data centres, with a focus on sustainability and efficiency. It then moved to ethical questions and policy considerations and finished with exploring decentralised data centres and local AI. Throughout the section and this thesis, missing policies were highlighted, which this section dives deeper into. Current regulations only partially address the challenges outlined, and new strategies will be needed to fill these gaps. This section identifies the major shortcomings in today's legal frameworks and discusses potential directions for future policy and research. It will start by addressing these policy gaps in AI regulation and further advocate for the need for standardised metrics. Additionally, it will emphasise how enforcement is important to achieve compliance. It then turns to the challenges inherent in legal frameworks and concludes by highlighting the risks posed by lacking global coordination.

[259] Ibid., Article 12(4).
[260] w.A. 2025a.
[261] Ibid.
[262] Ibid.
[263] Ibid.
[264] Ibid.
[265] Ibid.

4.4.1 Policy Gaps in AI Regulation

As discussed in section 4.1.2, the EU AI Act introduces transparency requirements, but overall it lacks enforcement of sustainability. It opened the doors for codes of conduct, making use of different actors in AI development. However, they lack the incentive for compliance, rendering them voluntary in practice. Other jurisdictions have even less in place, highlighting an imbalance in regulation. Governments are pushing for AI regulations but miss explicitly incorporating sustainability or environmental considerations[266]. A clear gap is the missing policy for AI developers requiring them to minimise carbon footprint, resource use, or efficiency requirements. Policymakers should consider setting enforceable measures, e.g. setting energy efficiency benchmarks, perhaps similar to Corporate Average Fuel Economy (CAFE) regulating how far a vehicle must travel on a gallon of fuel[267]. Another policy could mandate that companies must report carbon emissions or energy consumption of training large models, providing the option to choose sustainability. Training of models could also be restricted to achieve a sustainable percentage, e.g. only training when enough renewable energy is available[268]. Scholars suggest a more innovative approach, like *sustainability by design*, stemming from the GDPR's privacy by design. Just as the GDPR requires companies to embed privacy into system architecture, AI regulation should require sustainability considerations to be built into AI models and infrastructures from the outset[269]. Sustainability Impact Assessments (SIAs) would require developers to assess and disclose the environmental footprint of their AI system and firms to incorporate efficiency and resource-saving measures into model architecture, training processes, and deployment[270]. While similar suggestions are still emerging, the critical challenge lies in correctly assessing and crafting policies that do not hinder AI development or drive companies to relocate to more lenient jurisdictions. Section 4.4.3 will analyse and discuss these challenges more in depth. The following section will advocate for the importance of metrics and the need for its standardisation.

4.4.2 Standardised Metrics and Disclosure

A recurring theme is the lack of standardised metrics for measuring AI's sustainability. Companies may use different methods and different standards to calculate, e.g. CO_2 emissions for training an AI model, making it difficult to compare. Section 4.1.4 discusses the PUE requirement for data centres in Germany, which lowers the difficulty of comparing their sustainability. The absence of standardised reporting means that even when disclosure is encouraged or required, compliance might be inconsistent.

Experts have been advocating for standardised metrics in various areas, as it prevents miscommunication when different departments or international teams use different units of measurement, enhances accuracy by using consistent measures, or saves money by avoiding mistakes due to wrong metrics[271]. A famous mix-up is the 1999 NASA's Mars Climate Orbiter, which was unsuccessful as the spacecraft got lost on arrival, which was due to a mismatch of units on the ground and on board[272].

Standardised AI sustainability measurements could be achieved through international standards such as ISO/IEC or with regulations. The challenge lies in finding a consistent method and metric for measurement that can be applied across all AI systems. Section 3.2.1 discusses the Green AI approach and compares different metrics. This approach advocates for FLOP(s) as a primary unit of measurement. While it offers many advantages, it is important to not overlook complementary

[266] w.A. 2024a.
[267] w.A. 2024b.
[268] Hacker 2024, p. 27.
[269] Ibid., p. 22.
[270] Ibid., pp. 25–26.
[271] w.A. 2025j.
[272] w.A. 2025g.

metrics such as CO_2 emissions. Its disadvantages include the unpredictable outcomes, varying with factors such as time, location, and equipment. Yet, with proper standardisation of these parameters, it could demonstrate consistency comparable to FLOP(s). Future regulations might combine both FLOP(s) and CO_2 emissions to capture not only how much computation is performed but also what its environmental cost is. Once they are defined, regulation can require the transparent disclosure of those metrics, e.g. in the AI Act. Regulations could follow a similar approach as Germany's EnEfG (see Section 4.1.4) for data centres, where laws could mandate that AI models publish their metrics in a public registry. This creates market pressure for efficiency and sustainability as investors and consumers can compare AI services on these metrics. Through clearer metrics and stronger standardisation, AI regulations are coming closer to truly sustainable practices. However, regulations and compliance are only as strong as the enforcement of them. Subsequently, the next section will discuss how enforcement plays an important role in regulation and other challenges, such as finding a balance between regulation and freedom.

4.4.3 Challenges for Legal Frameworks

Throughout this thesis, the rapid growth of AI has been examined and discussed. This growth challenges legal frameworks, as, due to its nature, law can lag behind. This section summarises the challenges discussed so far and introduces new ones.

Even the best-crafted policies mean little without effective enforcement. It is complicated to ensure compliance, as regulators need technical expertise and resources to audit AI systems for energy use and ethical compliance. Future efforts might include establishing specialised AI sustainability audit bodies to conduct energy and ethical checks on AI systems. The EU's AI Office could be one such enforcement body, but it would need sufficient expertise for sustainability checks. However, positive incentives could complement penalties to encourage AI developers and deployers to push for efficiency. Governments can drive sustainable AI through funding and procurement, by e.g. providing grants for energy-efficient AI algorithms or easing regulation when sustainability goals are met. By aligning economic incentives, such as electricity pricing, carbon taxes or credits, with AI usage, companies can be encouraged to optimise. It is important to find a key balance between what can be done and what can't, posing a challenge for legal frameworks. These incentives will be discussed more in depth in Chapter 5.

Section 4.1 shows how currently, regulation mostly focuses on data centres powering AI and targeting their energy efficiency and resources. AI models themselves are mostly addressed ethically. Finding the balance between regulation and freedom will be challenging for lawmakers. The absence of sustainable AI regulations could carry serious repercussions for the environment. However, too strict regulations could hamper the development of AI, wasting its potential. It is important to note that stricter regulations can lead AI developers and businesses to shy away to areas with laxer regulations. Companies seeking faster growth or higher profits may simply withdraw from tightly regulated markets and shift operations to these more non-restrictive environments. This is a general rule as the very recent relocation from Austria of Ryanair and Wizz Air shows. They argued that the cost of operation in Austria, particularly the aviation tax, is too high and leaves no room for growth[273]. In this case it might be in some ways beneficial for the environment, but not beneficial for Austria in every scenario. Future law could encourage firms with, the already mentioned, economic incentives when obliging law and meeting goals.

Another problem arises when inspecting AI itself. Section 4.2.4 discusses how AI, by nature, is unpredictable, raising the many hands problem. The many hands problem is described as many different actors contributing in many different ways to outcomes, making it difficult to blame any actor[274]. Traditional law assumes a clear agent to hold accountable. AI's design

[273] Öser 2025.
[274] Thompson 1980, p. 2.

and operation, however, are spread across developers, data curators, regulators and users[275]. This shows that our current regulatory frameworks do not yet translate well to AI's opaqueness. Imagine a patient is diagnosed by a doctor who uses and is allowed to use AI, but it is faulty, raising the question of who is responsible. Some scholars suggest that legal frameworks must move beyond pinpointing an individual at fault and instead support mechanisms of challenge, redress and systemic oversight, accounting for AI's unpredictable behaviour[276].

In summary, the rapid growth of AI and its unpredictability brings great challenges for legislatives. As it has been made clear, the biggest challenge lies in balancing restrictions to not hinder the development of AI. Following this analysis of challenges for legal frameworks is another challenge nations have to overcome to achieve sustainable AI.

o

4.4.4 Global Coordination, Equity and Conclusion

AI is a global industry, which is dominated by a few nations, but its environmental and social impacts are distributed worldwide, often unequally. Current regulations are split, with only a few international obligations, whereas most regulatory efforts are led by the EU. There is a risk of lacking regulations where enforcement is not developed, drawing in bad actors to exploit a lenient regime. Regimes with lacking or no regulations at all could become a dumping ground for inefficient hardware, out-dated and energy-hungry AI models. Future law should lead with global coordination and an equal distribution of responsibility for sustainable AI, potentially through the G20 or United Nations, much like the Paris Agreement (see Section 2.2). Engaging the US, China, and other leading AI nations will be essential, as otherwise, efforts may remain fragmented. While non-binding agreements (e.g. UNESCO) are a start, eventually, agreements or national policies may be needed. Ensuring equity is also crucial. Most of the resource extraction and e-waste of AI burden communities usually far from AI's beneficiaries[277]. Their voices must be included in shaping regulations that hold AI producers accountable for their impacts.

In conclusion, there are many gaps in policies which are still being worked on or yet to be addressed. Despite the progress achieved with instruments such as the EU AI Act, significant gaps to govern sustainable AI still remain. Most frameworks still focus on compliance and innovation while overlooking standardised sustainability metrics, binding efficiency targets, and the integration of AI into broader climate and energy policies. Globally, there is still a considerable amount of work for regulators to do. Closing these gaps requires moving from soft law to enforceable rules, backed by incentives for greener AI design and stronger international cooperation. Only then can AI governance evolve from emerging regulations into a solid framework that truly aligns technological progress with sustainable development. Bridging the policy gaps identified above with the technical innovations discussed in Chapter 3 is crucial. However, providing economic incentives for companies developing AI will be key to achieve its sustainability. Only a synergy of technical efficiency measures, robust legal frameworks and economic incentives will truly align AI development with sustainable development goals. The next chapter turns to a closer examination of such incentives.

[275] Slota et al. 2023, p. 2.
[276] Ibid., pp. 10–11.
[277] Pereira 2024.

5 Aligning Economic Incentives with Sustainable AI

AI's rapid growth has brought unexpected social, environmental, and legal challenges, but it has also opened up opportunities for businesses across many sectors. This spans from developing proprietary systems to simply implementing ready-made solutions into daily work routines. Many companies have also expanded their products and services by embedding AI features, whether through intelligent assistance, predictive analytics, or automated decision-making tools.

Following the core analysis from a technical and legal perspective of sustainable AI, this chapter evaluates it through an economic lens. Addressing the challenges of sustainability requires aligning economic incentives at both macro and micro levels with suitable and appropriate goals. To achieve these goals, governments, industries, and companies must work towards increasing the efficiency of AI and reducing its carbon footprint.

This chapter starts by analysing how companies are increasingly implementing sustainability goals in their firm-level decision making. Many have recognised that energy efficiency and carbon reduction directly translate to cost savings. By optimising energy efficiency, choosing appropriate hardware, and utilising smart cooling, companies can considerably lower the operational costs of AI while simultaneously reducing their environmental impact. Section 5.2 builds on the findings from Chapter 3, linking technical strategies to economic incentives for businesses.

The chapter continues by investigating how AI development, combined with sustainability initiatives, can spur innovation. Inherently, these initiatives often spark technical innovation. Furthermore, strategic innovation to gain a competitive advantage over other firms is another powerful driver of sustainability efforts. An analysis of business opportunities that have emerged due to sustainable efforts and incentives for AI will conclude the section.

Finally, a discussion on how the alignment of macro and micro incentives is crucial to achieving sustainable AI will follow. The discussion will examine how regulations can help shape incentives for sustainable AI business practices. The idea of carbon prices and energy taxes, as well as subsidies and R&D support, will be explored.

5.1 Sustainable Decision-Making

The adoption of AI has become increasingly important for companies, as 61% of investors believe that its incorporation into their lifecycles is significant for the future, despite its risks[278]. Furthermore, at the firm level, incorporating sustainability into their strategies is seen as vital for long-term value. Thus, operations that are energy efficient and low-carbon have become prevalent, as companies are starting to embed environmental objectives to remain competitive. A 2023 investors survey showed that 75% of investors consider how firms manage sustainability risks and opportunities[279]. Crucially, they are interested in how sustainability is integrated into their business model. They argue that companies should invest in sustainability, even if it entails a short-term reduction in profits[280]. In the context of AI adoption, this means sustainable practices yield dual benefits: cost savings through energy efficiency and stronger stakeholder relationships via improved sustainability. In practice, many companies are already prioritising AI implementations when making capital investments. However, a rising interest in sustainability practices can be observed, as approximately 40% of CEOs are focusing on them[281]. By treating sustainability as a core strategy, firms can future-proof their AI deployments and avoid later expenses. The next section will examine the operational cost of energy efficiency and how AI is tied to it in more depth.

[278] Chalmers and Picard 2023.
[279] Ibid.
[280] Ibid.
[281] w.A. 2025c.

5.2 Operational Efficiency and Cost Savings

A primary economic incentive for sustainable AI is improved operational efficiency, which directly translates to cost savings. As section 2 discusses, AI greatly impacts the environment due to its massive energy consumption, both in training and operation. However, a recent large-scale study of Chinese firms shows strong empirical evidence that AI adoption is strongly connected to reduced energy consumption. For every 1% increase in AI use, overall energy consumption drops by about 0.48%, primarily through green innovation[282]. This suggests that AI, when deployed with efficiency in mind, can optimise operations and cut resource waste. It can dynamically tune processes to save energy, a task that would otherwise be tedious and time-consuming for human operators. For example, idle servers can be turned off, cooling can be optimised, or supply chains streamlined. This section will examine how operational costs can differ when operating AI with efficiency in mind.

5.2.1 Sustainable Data Centre Practices

The operational cost of data centres largely consists of their electricity consumption, as discussed in Section 3.4.3. Since data centres typically power AI systems, their operational costs can be reduced through sustainable AI practices.

The MIT Lincoln Laboratory's Supercomputing Centre demonstrated that pairing sustainability measures with AI workloads can cut both energy use and expenses. They tested a range of measures to cut energy costs and found that, when implemented, data centre emissions can be reduced by 10%-20%; most importantly, without significant capital investment[283]. In other words, by reducing AI's energy use, operational expenses can directly improve the bottom line. A simple change includes opting for more efficient hardware, as discussed in Section 3.4.1. If this is not possible, firms can experiment with *power capping* or limit the power available, depending on the tasks. The lab managed to reduce the overall energy consumption of AI workloads, thus lowering cooling demand[284]. Similarly, rethinking model training routines yielded significant savings. By developing a tool that allows for early stopping once convergence is predicted, 80% of the computation could be eliminated, with no loss in accuracy, translating to substantial energy and cost savings. In summary, these sustainable and efficient practices for AI demonstrate how they directly cut operational costs and lower the environmental impact at the data centre level[285].

5.2.2 Data Centre Infrastructure

Apart from sustainable data centre practices, data centre infrastructure optimisation yields further economic benefits. Section 3.4.3 discusses how smart resource distribution and modern cooling systems can reduce energy consumption. Sometimes, there are too many dimensions for human operators to evaluate consistently. AI systems can assist by processing this complexity and supporting more effective decision-making. This section will outline the successful integration of smart infrastructure, where AI was used efficiently to reduce energy consumption in resource management and cooling.

A famous example is Google leveraging its DeepMind AI to manage data centre cooling, achieving up to a 40% reduction in energy usage[286]. They accomplished this by collecting historic data, such as temperature and power, and furthermore applying machine learning, utilising the data to optimise PUE. Additionally, two other models predicted the temperature and pressure of

[282] Y. Zhou and Bu 2025, p. 1.
[283] Stackpole 2025.
[284] Ibid.
[285] Ibid.
[286] Evans and Gao 2016.

their data centre, to avoid overshooting any operating constraints. This setup was deployed in a live data centre, consistently achieving a 40% reduction in energy consumption for cooling[287]. This translated to a 15% increase in overall PUE, which is essentially an enormous cut in the electricity bill for an already very optimised data centre[288]. DeepMind continued to learn and adjust cooling parameters, effectively outperforming human operators in squeezing out efficiency.

Other firms replicated this approach by using AI for energy management. Hilton and ei3 developed LightStay, an AI-driven energy management platform used to monitor and reduce energy, water, and waste across thousands of hotels. This initiative led to over 1 billion dollars in cost savings in about a decade, over a 20% reduction in water and energy usage, and a 30% reduction in carbon emissions and waste output[289]. They achieved these results through an automated alert system that triggers when performance falls below expected levels, a comprehensive impact tracking system allowing for the monitoring of vital metrics, and a global footprint assessment of various events[290]. These features fostered a sustainable environment while allowing for ongoing high performance.

In summary, these case studies illustrate that businesses can employ AI-driven optimisation even in already well-run operations to save money and future-proof their facilities. They show how AI-based efficiency measures scale across industries to deliver cost savings at the firm level.

5.2.3 Hardware Choices

Hardware choices also play a crucial role in the cost structure. Section 3.4.1 analyses different hardware components and their effectiveness for AI operations. Figure 14 illustrates that by choosing accelerated computing hardware, such as TPUs or specialised FPGAs, AI and high performance computing (HPC) tasks can be performed far more efficiently than with legacy CPUs. In other words, by investing in modern hardware, firms can yield long-term gains. NVIDIA reports that by transitioning to accelerated hardware for their servers, they can save over 40 TWh of energy annually for HPC and AI workloads, equivalent to what nearly 5 million U.S. homes need for electricity[291]. Another example comes from a financial services company, Murex, which found that integrating NVIDIA's specialised Grace Hopper Superchip yielded a 4× reduction in energy consumption and a 7× faster computation speed compared to traditional CPU-based systems[292]. In data analytics, GPU acceleration of Apache Spark was shown to cut power use by 5× and infrastructure costs by 4×, saving a typical enterprise nearly 125 million dollars and 10 GWh over a period of use[293].

At the firm-level, these efficiency gains translate to powering AI workloads with fewer servers, thereby reducing operational costs. Even if efficient hardware costs more upfront, the energy savings it delivers reduce operating expenses and ultimately increase profits, making the investment pay off. Firms that aggressively adopt efficient algorithms and hardware can achieve a structural cost advantage over competitors still using energy-hungry setups.

In summary, sustainable AI practices contribute to leaner operations, meaning fewer resources and fewer redundant processes. By slashing energy and resource waste, firms can reduce operating costs, indirectly recouping their investments in green technology. Energy-efficient AI aligns with profit motives, as a lower carbon footprint frequently also reduces electricity consumption and hardware requirements. It improves ROI for AI and can free up the budget for further innovation. This highlights cost efficiency as one of the strongest drivers for companies to pursue sustainability in their decision-making, especially when adopting AI.

[287] Ibid.
[288] Ibid.
[289] w.A. 2008.
[290] Ibid.
[291] Harris 2024.
[292] Ibid.
[293] w.A. 2024f.

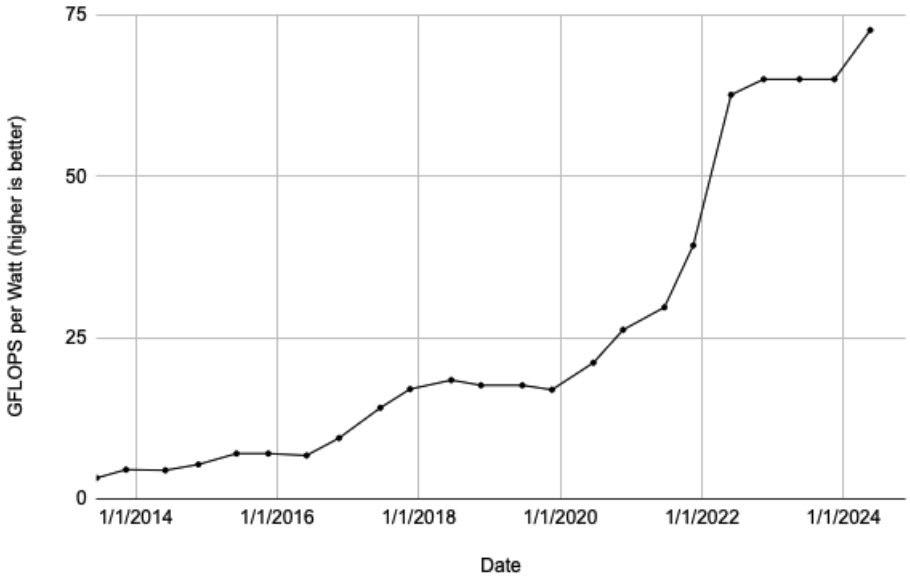

Fig. 14: This figure illustrates the energy-efficiency gains for the most efficient supercomputer. This was possible due to the use of specialised hardware. Source: Harris 2024

5.3 Sustainable Innovation and Competitiveness

Innovation has always propelled society forward, enabling technological evolution. It spans from the creation of tools in the Stone Age to ENIAC, the first digital computer, in 1940[294]. Today, AI might stand as the next great turning point. Sustainable AI, in particular, sparks new forms of innovation by demanding greater efficiency, both in technical design and in strategic approaches. At the same time, it creates new business opportunities as firms experiment with new products, services, and models to align profitability with sustainability. This section explores how sustainable AI acts as a catalyst for innovation in these three dimensions.

5.3.1 Technical Innovation

The need to make AI more efficient often drives technical innovation. Section 3.3 analyses different model compression techniques, which were mainly accelerated by the desire to maintain model accuracy while cutting computational costs. A recent study shows that by applying model compression techniques to large NLP models, like BERT (see Section 3.2.2), a 32% reduction in energy consumption can be achieved while maintaining high performance[295]. Similarly, its distilled student model, DistillBERT, achieved comparable performance. By combining two compression techniques, knowledge distillation and pruning, it achieved a 6.7% energy reduction, despite already being highly compact and efficient[296]. These innovations mean that companies can offer AI services that are cheaper to run but still high-performing. For example, a startup that provides AI predictions but uses a highly optimised model might offer a lower price or a sustainability guaranty to customers, undercutting competitors who require more compute resources. In this way, sustainable AI methods can become a source of competitive differentiation. This underscores the tight interplay between technical innovation and economics, as competitiveness sparks further innovation, and those innovations enable better offerings and cost savings.

[294] McCartney 1999, p. 1.
[295] Paula et al. 2025, p. 1.
[296] Ibid., p. 1.

5.3.2 Strategic Innovation

The **Porter Hypothesis** in environmental economics posits that environmental challenges can stimulate innovation that eventually improves both environmental and business performance, essentially translating to the idea that environmental regulations can spur innovation and efficiency[297]. This hypothesis is mirrored through sustainable AI, as limited energy budgets and carbon targets push companies to innovate AI algorithms and infrastructure.

Microsoft's case provides evidence, as the pledge to be carbon negative by 2030 has forced new innovation. Their engineers are re-architecting how AI workloads are scheduled, introducing AI-driven global schedulers and power harvesting techniques that didn't exist in traditional cloud operations. Project Forge is an AI scheduler that achieved 80-90% GPU utilisation by intelligently scheduling AI jobs during times when resources are idle[298]. The result is not only an environmental win, but also a competitive one, allowing their cloud platform, Azure, to handle more demand without proportionally expanding its servers. Microsoft also implemented power harvesting in their data centres, reallocating unused power from over-provisioned workloads to other tasks, which reclaimed 800 MW of capacity since 2019, a technique discussed in Section 4.3[299].

Likewise, Google pledged to be carbon neutral by 2030 and developed carbon intelligent computing, which moves flexible computing tasks across time and geographies to use cleaner energy when available[300]. This required new software and forecasting innovations, such as models that predict hourly grid carbon intensity. Google can now offer cloud development customers the ability to run their tasks when renewable energy is abundant, effectively lowering or negating their carbon footprint[301]. This capability may attract environmentally conscious clients, thereby becoming a market differentiator.

Through their own carbon pledges, these firms have pushed themselves to innovate sustainable practices for their data centres. Their advancements were not only beneficial for the environment but also proved to bring a competitive advantage, as they can either attract environmentally conscious customers or offer services and products at a lower price. Furthermore, these innovations foster new development, which, in turn, offers new business opportunities, discussed further in the following section.

5.3.3 Business Opportunities

As sustainable AI continues to advance, it demands greater innovation, thereby creating new products and opening new market opportunities. The pressure to be carbon neutral rises while the timeline to achieve it is shrinking. Companies have realised the importance of AI in creating a carbon-neutral space and have started seeking solutions to optimise energy usage, design greener products, or manage carbon emissions. Section 5.1 discusses how investors and AI buyers are primarily interested in sustainable practices and companies AI adoption efforts. Proficiency in such practices can be leveraged to offer new services. For instance, NVIDIA has positioned itself as a leader in sustainability through their push for energy efficiency in their chips and now markets their hardware for enabling sustainability efforts[302]. As their GPUs and software explicitly promise both performance and energy efficiency, they appeal to customers with environmental, social, and governance (ESG) goals. Another example is the opportunities that efficient AI brings to firms focusing on energy management. DeepMind's success with their AI cooling management not only saved money but also created a template that could be applied

[297] Y. Zhou and Bu 2025, p. 5.
[298] Russinovich 2024.
[299] Ibid.
[300] Koningstein 2021.
[301] Ibid.
[302] S. Khan 2024.

in other locations. Indeed, the team behind DeepMind noted that these methods could help any data centre improve efficiency[303]. Similar opportunities arise in the automotive AI space, as the push for energy efficiency to prolong battery life could be achieved through low-power AI accelerators, once again giving companies with sustainable know-how an advantage.

From a microeconomic standpoint, innovation driven by sustainability can upgrade a firm's competitive standing. It can yield unique capabilities, such as efficient algorithms and tools for optimisation, which rivals may lack. Furthermore, by operating and developing AI with efficiency in mind, one can also future-proof their company against evolving regulations or resource constraints and additionally appeal to sustainable customers. A firm capable of handling more computation with less power will face lower carbon taxes and energy costs than one burdened by inefficient, power-hungry AI systems. Moreover, by achieving efficiency first, firms can set industry benchmarks and shape standards. They can be placed in a leadership position, as their practices become gold standards, forcing others to follow in their footsteps.

Section 5.2 discusses empirical evidence on how the adoption of AI can facilitate innovation in sustainability. AI can function as a *green-enabling* dynamic capability, as companies adopting AI also tend to boost their green innovation[304]. It helps firms design cleaner processes and products, aligning with the idea that it can unlock new ways to meet environmental challenges[305]. A study found that green innovation was a key mechanism by which AI adoption led to reduced energy consumption at firms[306]. It can again be concluded that AI and sustainability are not only beneficial for the environment but also for the electricity bill.

In summary, sustainable AI can be a catalyst for innovation. Firms that embrace the challenge of using and developing AI in a sustainable way often end up with more efficient technology, processes, and know-how. It strengthens competitiveness, lowers costs, and improves resilience. As public concern for the environment grows and more customers demand sustainable products, adopting sustainable AI practices becomes a natural necessity. This means that doing good for the environment can also mean doing well financially at the firm level. By setting industry standards, companies may benefit from lower policy-related costs, such as carbon or energy taxes—which are discussed as incentives for sustainability in the following section.

5.4 Integrating Legal and Economic Approaches

Throughout this chapter, it has been illustrated how technological innovation strongly pushes economic incentives. Crucially, macro and micro incentives reinforce each other when properly coordinated. Legal frameworks provide external pressure and support, e.g. through standards, penalties, or public investment, while economic self-interest embraces innovation. For instance, a policy mandating the reporting of energy consumption, such as Germany's EnEfG (see Section 4.1.4), establishes accountability. Due to the competitive nature of markets, each firm tries to outperform its rivals on those metrics in order to attract investors. Likewise, if governments set a carbon price or efficiency standard, companies have the certainty needed to invest in long-term energy-saving innovations. In effect, policy *pulls* and market *pushes* can align toward the same goal. Chapter 4 has thoroughly discussed the legal frameworks surrounding AI and sustainability, whereas Section 4.4 examines policy gaps and features some incentives for businesses. This section will further examine how policymakers can encourage sustainable AI practices for businesses through taxes, subsidies, and regulations that alter the cost of operating AI sustainably.

[303] Evans and Gao 2016.
[304] Y. Zhou and Bu 2025, p. 2.
[305] Ibid., p. 5.
[306] Ibid., p. 2.

5.4.1 Carbon Pricing and Energy Taxes

A recent effort to reduce emissions has been the taxation of greenhouse gas emissions. The **Carbon Tax** is a law-mandated tax that targets operations emitting greenhouse gases, essentially placing a price on carbon emissions. In 1990, Finland was the first country to introduce such a tax, and since then, 23 European countries have implemented some form of carbon tax[307]. The scope of the carbon tax usually differs between countries, as some only target a small percentage of their carbon emissions, while others cover more than half[308]. To combat the variety, the EU decided to implement a unified system called the Emissions Trading System (ETS). It is based on *cap and trade*, where the cap refers to a limit on greenhouse gases that can be emitted. Companies affected by the ETS are therefore encouraged to implement sustainability measures, as they must pay heavy fines if they cannot account for their emissions[309]. Currently, only the sectors of electricity and heat generation, industrial manufacturing, aviation, and maritime transport are covered by the EU ETS[310]. Given that most new policies require data centres or AI systems to disclose their emissions (see Section 4.1.2), it could be considered to bring these sectors into the scope of the ETS. This would urge data centre operators and AI companies to implement sustainability practices throughout their lifecycle in order to avoid heavy fines. However, data centres and AI systems primarily consume electricity and do not burn fuel, making them low direct carbon emitters.

Some economists thus advocate that a broad climate price is one of the most effective tools for curbing emissions, as it encourages reduced fossil-fuel consumption. For data centres, the International Monetary Fund (IMF) estimates that a targeted electricity tax of $0.03 per kWh could significantly cut their carbon footprint[311]. This would internalise the environmental cost of running power-hungry AI workloads, nudging companies to optimise their compute for efficiency. However, this does not represent the status quo of today. Many jurisdictions offer tax breaks and incentives despite their high energy usage to lure data centre investments.

For example, the U.S. state of Indiana passed a bill in 2019 that offered a 7% tax cut for data centre operators when purchasing resources[312]. Today, nearly all U.S. states have implemented a similar tax break for data centres, and these investments are projected to reach one trillion dollars by 2027[313]. These incentives are being re-examined as their public benefits come under question and are not tied to any goals. Data centres have growing concerns, such as their environmental impact through energy consumption, while offering very few long-term jobs[314]. In 2024, lawmakers from Georgia proposed a bill that would pause tax breaks for data centres in order to evaluate their impact on the environment. Though the proposal was vetoed in the end to maintain investment flow, it underscores a shifting attitude toward more conditional, sustainability-oriented incentive policies[315]. Instead of providing tax breaks without any requirements, policies could consider offering subsidies that are tied to sustainability goals.

5.4.2 Subsidies and R&D Support

On the flip side of using penalties to encourage sustainable practices, governments could use financial incentives to accelerate the development and adoption of greener AI technologies. This might include subsidies for companies that invest in energy-efficient hardware, algorithms, or infrastructure. For example, public policies can provide R&D funding and innovation prices

[307] Alex Mengden 2025.
[308] Ibid.
[309] w.A. 2005a.
[310] w.A. 2005b.
[311] Shafik Hebous 2024.
[312] Tolockaite et al. 2025.
[313] Ibid.
[314] Hughett 2025.
[315] Ibid.

for new techniques that reduce AI's energy consumption. Policymakers can thus lower the cost barrier for firms to pursue efficiency.

The EU is once again a pioneer, as it has initiated the European Green Deal, discussed in Section 4.1.2, which has pledged to invest at least 1 trillion euros into sustainable investment[316]. As part of its agenda, there are investments in energy-efficient AI R&D called **GREEN.DAT.AI**. The EU is funding the project, which is a consortium of 17 organisations developing novel AI techniques that use less energy and reduce environmental impact[317]. It explicitly aims to channel AI towards the goals of the European Green Deal by creating large-scale data analysis services that are more energy-efficient for industry, thereby cutting the carbon footprint of AI and data processing[318].

Similarly, the U.S. has started backing research on sustainable AI. In 2024, the U.S. Department of Energy announced a $68 million funding program for AI in scientific research[319]. This includes the development of energy-efficient AI algorithms and hardware that use fewer resources.

The intervention of law-makers is particularly important, as letting the market determine the shape of AI may be too risky. If there are no incentives for the research and development of sustainable AI, then AI might not take a turn towards the public good[320]. In conclusion, fostering sustainable AI is as much an economic and legal challenge as it is a technical one. Well-designed legal incentives set the stage by rewarding low carbon operations while penalising wasteful ones, and economic incentives encourage businesses to adopt greener practices. Efficiency is the key connection, as it reduces energy consumption, lowers carbon emissions, and serves as a source of cost savings. To achieve sustainable AI, it must be ensured that technical advances make it possible; subsequently, economic factors and legal frameworks must make it standard practice. This multidisciplinary approach will be crucial to guiding AI development onto a more sustainable trajectory for the future.

[316] w.A. 2021.
[317] w.A. 2025e.
[318] Ibid.
[319] w.A. 2024c.
[320] Stern et al. 2025, p. 6.

6 Discussion and Conclusion

The technical advancements in sustainable AI discussed in Chapter 3 allow for stringent regulations that focus on efficiency instead of accuracy. Conversely, sustainable policies promote technical innovation by creating requirements and standards. For example, Germany's Energy Efficiency Act mandating PUE requirements for data centres forces operators to invest in smart cooling strategies and sustainable measures, which were further explored in Chapter 3. These investments also make economic sense, as they reduce long-term operating costs. A company might adopt model compression or more efficient hardware primarily to save money. This illustrates the synergy between technology, policy, and the economy, as regulation drives technical innovation while technological advances make it possible to uphold regulatory standards, thus providing cost benefits for businesses. An integrated view helps balance the trade-offs, which is critical given the rapid development of AI. Chapter 4 stresses the importance of not hindering the development of AI. Sometimes, the highest energy efficiency might come at a higher upfront cost due to investments in development and a sacrifice in performance. Economic analysis helps determine whether the long-term savings justify the expenditure. Legal frameworks can tip this balance by either offering subsidies, such as reducing the cost of sustainability, or by making inefficiencies costlier through penalties or energy prices. Technologically, there might be trade-offs in performance, presenting the question: how much performance are we willing to trade for sustainability? This again cycles back to political and ethical questions, e.g. should larger models be restricted according to their environmental footprint? This thesis strongly advocates a sustainable approach, presenting different techniques and suggestions throughout the technical, legal, and economic chapters.

6.1 Synergy Between Technology, Policy and Economics

6.2 Need for Collaboration

This thesis has consistently underscored the importance of collaboration among policymakers, developers, and business managers in advancing sustainable AI. Policymakers should consult experts on realistic but ambitious efficiency targets, while business managers should advocate for and help shape regulations. Decentralised data centres, discussed in Section 4.3, open the room for public–private collaboration. Integrating excess heat from data centres into public infrastructure, while offering benefits in return, provides businesses with incentives to pursue such sustainable practices. Imagine a scenario where a business developing a new AI product considers using either a massive AI model or a smaller custom model. The technical perspective showed us that smaller models can be nearly as good but far more efficient. If the legal framework becomes stricter toward models with higher environmental impact while granting benefits to those that are more efficient and environmentally friendly, the choice for businesses may become obvious. This hypothetical illustrates the interlink between technical and legal aspects and how they affect economic standards.

6.3 Conclusion

This thesis has explored sustainable AI from technical, legal, and economic perspectives. It began by examining the growing concern over AI's environmental impact, driven by its rapid expansion. By defining the concept of sustainable AI, it established the central focus of this work: the pursuit of efficient, environmentally responsible, and economically viable artificial intelligence.

Technically, it has examined multiple advancements towards sustainability, ranging from energy-efficient AI model design and hardware optimisations to model compression techniques and smart training regimes, drastically reducing the energy consumption of AI systems. Real-

world examples like DistilBERT and the Zeus framework demonstrate how efficiency gains are achievable without severely compromising performance. Energy-efficient data centre designs have also been inspected, as they usually power AI.

Legally, existing and emerging frameworks have been explored, with the EU AI Act being a pioneer and setting expectations for sustainable AI to become standard practice. A deep dive into national regulatory frameworks, such as those of Austria and Germany, followed, whereas the EnEfG in Section 4.1.4 demonstrated how a country can lead with sustainability measures. Mandating PUE targets, waste heat reuse, and renewable energy puts it at the forefront of sustainable AI infrastructure. The case study showed the interlink between perspectives: legal requirements push technical innovation, and in turn, these innovations make compliance feasible. As the practice of sustainable AI extends beyond just energy efficiency, ethical considerations have been surveyed, and the ways in which regulators have addressed them have been examined. However, even through the recent EU AI Act, policies, especially internationally, are still not complete, as discussed in Section 4.4. A standardised metric for comparing the efficiency of AI models is still missing, making it difficult for users to differentiate between them in terms of sustainability. Furthermore, the lack of global coordination of legislation is addressed, as policy gaps can lead to unsafe and power-hungry AI models.

Economically, incentives have been aligned with sustainable AI. The demand for AI systems integrated into business services continues to rise, and with it, sustainable decisions are becoming increasingly important to investors. Companies have started adopting environmentally conscious practices, as they are not only beneficial for the environment but also offer future-proof infrastructure for reduced long-term costs. Section 5.2 goes in depth about how sustainable AI operations can both reduce carbon emissions and operational costs. Sustainable data centre practices, smart infrastructure, and modern hardware all play a role in reducing power consumption, shrinking carbon emissions, and, at the same time, saving money. Furthermore, the chapter explores how sustainable AI fosters technical and strategic innovations, which open the door to new business opportunities. The final section of the chapter is a macroeconomic overview of business incentives that policymakers use (or could employ) to guide AI firms towards sustainability. They play a crucial role in shaping the future of AI, as businesses must comply to avoid fines. Striking the right balance between penalties and rewards is essential to avoid discouraging competition.

Looking ahead, AI will only increase its presence, and its sustainability will become more important. It has already undergone enormous development over the course of its lifecycle, with trends such as generative AI, autonomous driving, smart assistance, and more. If properly guided by sustainability principles, its environmental impact can either be equalised or at least lead to efficient growth. Technologically, more efficient solutions can be expected as the Green AI approach continues to rise and specialised hardware and software are developed. Legally, more nations will likely implement AI policies, using the AI Act as a basis. The AI Act itself is still in development, but future generations will possibly already require sustainability requirements. Internationally, legal standards still need to emerge, but standardised metrics on AI sustainability measures could be the first step. Economically, businesses will continue to integrate AI into their operations and services. The more efficiently they use AI, the more cost-effective it will become.

The journey towards sustainable AI is a collective one. This thesis underscores the importance of a multidisciplinary approach to sustainable AI. Through innovation and efficiency in mind from technical engineers, guided by appropriate laws from legislation, and fostered by economic incentives, environmentally friendly and sustainable AI can be achieved. This thesis has shown that it is not just about mitigating the negatives. It is an opportunity to improve AI technology, strengthen its efficiency, and responsibly shape the future of AI in society. In conclusion, pursuing sustainable AI is a necessity for the future of technology, the well-being of humanity, and the preservation of our planet.

List of Figures

References

Alex Mengden, A. N. (2025). *Carbon Taxes in Europe, 2025*. Accessed: 2025-09-28. URL: https://taxfoundation.org/data/all/eu/carbon-taxes-europe/.

Amodei, D. and D. Hernandez (2018). *AI and compute*. Accessed: 2025-04-23. URL: https://openai.com/index/ai-and-compute/.

Ball, C. (2009). "What Is Transparency?" In: *Public Integrity* 11.4, pp. 293–308. DOI: 10.2753/PIN1099-9922110400. eprint: https://www.tandfonline.com/doi/pdf/10.2753/PIN1099-9922110400. URL: https://www.tandfonline.com/doi/abs/10.2753/PIN1099-9922110400.

Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei (2020). *Language Models are Few-Shot Learners*. arXiv: 2005.14165 [cs.CL]. URL: https://arxiv.org/abs/2005.14165.

EEffG (Apr. 18, 2024). *Bundesgesetz über die Verbesserung der Energieeffizienz bei Haushalten, Unternehmen und dem Bund sowie Energieverbrauchserfassung und Monitoring (Bundes-Energieeffizienzgesetz – EEffG)*. URL: https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=20008914.

ElWG (Ministerialentwurf) (July 3, 2025). *Bundesgesetz zur Regelung der Elektrizitätswirtschaft (Elektrizitätswirtschaftsgesetz – ElWG)*. Ministerialentwurf; Begutachtung 7.7.–15.8.2025. URL: https://www.parlament.gv.at/gegenstand/XXVIII/ME/32.

Burger, C. (2023). *Austria achieves 87 percent of its electricity from renewable sources, second highest in the EU*. Accessed: 2025-08-29. URL: https://www.apg.at/en/news-press/apg-balance-sheet-record-figures-in-2023-confirm-challenging-overall-situation/.

Burmagina, K. (2025). *ChatGPT Usage: Statistics and Facts*. Accessed: 2025-04-22. URL: https://elfsight.com/blog/chatgpt-usage-statistics/#:~:text=A%20Record%2Dbreaking%20user%20base&text=Within%20just%20five%20days%20of,million%20ChatGPT%20weekly%20active%20users..

Butler, G. (2023). *A tidal wave of regulations*. Accessed: 2025-04-25. URL: https://www.datacenterdynamics.com/en/analysis/a-tidal-wave-of-regulations.

Cao, L. (2022). "Decentralized AI: Edge Intelligence and Smart Blockchain, Metaverse, Web3, and DeSci". In: *IEEE Intelligent Systems* 37.3, pp. 6–19. DOI: 10.1109/MIS.2022.3181504.

Ceci, L. (2023). *Number of YouTube viewers in the United States from 2018 to 2022*. Accessed: 2025-05-14. URL: https://www.statista.com/statistics/469152/number-youtube-viewers-united-states/#:~:text=The%20online%20video%20platform's%20audience,popular%20online%20video%20property%20worldwide..

Chalmers, J. and N. Picard (2023). *PwC's Global Investor Survey 2023: Trust, tech and transformation — Navigating investor priorities*. Accessed: 2025-09-24. URL: https://www.pwc.com/gx/en/issues/c-suite-insights/global-investor-survey-2023.html.

Dash, S. (2025). "Green AI: Enhancing Sustainability and Energy Efficiency in AI-Integrated Enterprise Systems". In: *IEEE Access* 13, pp. 21216–21228. DOI: 10.1109/ACCESS.2025.3532838.

Dhilleswararao, P., S. Boppu, M. S. Manikandan, and L. R. Cenkeramaddi (2022). "Efficient Hardware Architectures for Accelerating Deep Neural Networks: Survey". In: *IEEE Access* 10, pp. 131788–131828. DOI: 10.1109/ACCESS.2022.3229767.

Energy Efficiency Directive (Sept. 20, 2023). *Directive (EU) 2023/1791 of the European Parliament and of the Council of 13 September 2023 on energy efficiency and amending Regulation (EU) 2023/955 (recast)*. Directive. OJ L 231, 20.9.2023. ELI: http://data.europa.eu/eli/dir/2023/1791/oj.

European Parliament and the Council. URL: https://eur-lex.europa.eu/eli/dir/2023/1791/oj.

Evans, R. and J. Gao (2016). *DeepMind AI Reduces Google Data Centre Cooling Bill by 40 percent*. Accessed: 2025-09-25. URL: https://deepmind.google/discover/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-by-40/.

Framework Convention on AI (May 17, 2024). *Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law*. Convention. Adopted by the Committee of Ministers, opened for signature by Council of Europe member and non-member states. URL: https://rm.coe.int/1680afae3c.

EnEfG (Nov. 17, 2023). *Gesetz zur Steigerung der Energieeffizienz in Deutschland (Energieeffizienzgesetz – EnEfG)*. Bundesgesetz. In force since 2023-11-18. URL: https://www.recht.bund.de/bgbl/1/2023/309/VO.

Gordon, J. and A. Tulip (1997). "Resource scheduling". In: *International Journal of Project Management* 15.6, pp. 359–370. ISSN: 0263-7863. DOI: https://doi.org/10.1016/S0263-7863(96)00090-7. URL: https://www.sciencedirect.com/science/article/pii/S0263786396000907.

Hacker, P. (2024). "Sustainable AI Regulation". In: *Common Market Law Review* 61.2, pp. 345–386. ISSN: 0165-0750. URL: http://www.kluwerlawonline.com/api/Product/CitationPDFURL?file=Journals%5CCOLA%5CCOLA2024025.pdf.

Harris, D. (2024). *Sustainable Strides: How AI and Accelerated Computing Are Driving Energy Efficiency*. Accessed: 2025-09-25. URL: https://blogs.nvidia.com/blog/accelerated-ai-energy-efficiency/#:~:text=By%20transitioning%20from%20CPU,homes.

Ho, E., H. Mao, C. Luo, D. Wu, and A. Eassa (2024). *Accelerate Generative AI Inference Performance with NVIDIA TensorRT Model Optimizer, Now Publicly Available*. Accessed: 2025-05-17. URL: https://developer.nvidia.com/blog/accelerate-generative-ai-inference-performance-with-nvidia-tensorrt-model-optimizer-now-publicly-available/.

Hughett, E. (2025). *Why U.S. States Are Reconsidering Data Center Incentives: Navigating Shifting Policies and Sustainability Expectations*. Accessed: 2025-10-12. URL: https://info.siteselectiongroup.com/blog/states-rethink-data-center-incentives-amid-new-pressures#:~:text=%2A%20Few%20Long,which%20can%20cause%20delays%20or.

Iandola, F. N., S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer (2016). *SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size*. arXiv: 1602.07360 [cs.CV]. URL: https://arxiv.org/abs/1602.07360.

John-Mathews, J.-M., D. Cardon, and C. Balagué (2022). "From Reality to World. A Critical Perspective on AI Fairness". In: *Journal of Business Ethics* 178.4, pp. 945–959. ISSN: 1573-0697. DOI: 10.1007/s10551-022-05055-8. URL: https://doi.org/10.1007/s10551-022-05055-8.

Kaz Sato, C. Y. (2017). *An in-depth look at Google's first Tensor Processing Unit (TPU)*. URL: https://cloud.google.com/blog/products/ai-machine-learning/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu?hl=en.

Khan, S. (2024). *AI at COP29: Balancing Innovation and Sustainability*. Accessed: 2025-09-28. URL: https://blogs.nvidia.com/blog/cop29-energy-efficiency-panel/#:~:text=That%E2%80%99s%20why%20accelerated%20computing%20is,sustainable%20computing.

Koningstein, R. (2021). *We now do more computing where there's cleaner energy*. Accessed: 2025-09-25. URL: https://blog.google/outreach-initiatives/sustainability/carbon-aware-computing-location/#:~:text=And%20that%E2%80%99s%20exactly%20what%20our,at%20all%20times%2C%20by%202030.

Li, Z., F. Liu, W. Yang, S. Peng, and J. Zhou (2022). "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects". In: *IEEE Transactions on Neural Networks and Learning Systems* 33.12, pp. 6999–7019. DOI: 10.1109/TNNLS.2021.3084827.

Liang, T., J. Glossner, L. Wang, S. Shi, and X. Zhang (2021). "Pruning and quantization for deep neural network acceleration: A survey". In: *Neurocomputing* 461, pp. 370–403. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2021.07.045. URL: https://www.sciencedirect.com/science/article/pii/S0925231221010894.

Lin, T., Y. Wang, X. Liu, and X. Qiu (2022). "A survey of transformers". In: *AI Open* 3, pp. 111–132. ISSN: 2666-6510. DOI: https://doi.org/10.1016/j.aiopen.2022.10.001. URL: https://www.sciencedirect.com/science/article/pii/S2666651022000146.

Lindberg, J., B. C. Lesieutre, and L. A. Roald (2022). "Using geographic load shifting to reduce carbon emissions". In: *Electric Power Systems Research* 212, p. 108586. ISSN: 0378-7796. DOI: https://doi.org/10.1016/j.epsr.2022.108586. URL: https://www.sciencedirect.com/science/article/pii/S0378779622006757.

Liu, T. and B. Liu (2018). *Constrained-size Tensorflow Models for YouTube-8M Video Understanding Challenge*. arXiv: 1808.06739 [cs.CV]. URL: https://arxiv.org/abs/1808.06739.

Luccioni, A. S., S. Viguier, and A.-L. Ligozat (2022). *Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model*. arXiv: 2211.02001 [cs.LG]. URL: https://arxiv.org/abs/2211.02001.

McCall, S. (2023). *ChatGPT banned in Italy over privacy concerns*. Accessed: 2025-09-01. URL: https://www.bbc.com/news/technology-65139406.

McCartney, S. (1999). *ENIAC: The Triumphs and Tragedies of the World's First Computer*. Walker & Company. ISBN: 0802713483.

Micikevicius, P., S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, and H. Wu (2018). *Mixed Precision Training*. arXiv: 1710.03740 [cs.AI]. URL: https://arxiv.org/abs/1710.03740.

Mukherjee, D., S. Chakraborty, I. Sarkar, A. Ghosh, and S. Roy (2020). "A detailed study on data centre energy efficiency and efficient cooling techniques". In: *International Journal* 9.5, pp. 9221–9242. URL: https://www.researchgate.net/publication/344100890_A_Detailed_Study_on_Data_Centre_Energy_Efficiency_and_Efficient_Cooling_Techniques.

Novelli, C., M. Taddeo, and L. Floridi (2024). "Accountability in artificial intelligence: what it is and how it works". In: *AI & SOCIETY* 39.4, pp. 1871–1882. ISSN: 1435-5655. DOI: 10.1007/s00146-023-01635-y. URL: https://doi.org/10.1007/s00146-023-01635-y.

Öser, C. (2025). *Nach Wizz Air: Ryanair dünnt Standort Wien-Schwechat aus*. Accessed: 2025-09-18. URL: https://orf.at/stories/3405737/.

Paula, E., J. Soni, H. Upadhyay, and L. Lagos (2025). "Comparative analysis of model compression techniques for achieving carbon efficient AI". In: *Scientific Reports* 15.1, p. 23461. ISSN: 2045-2322. DOI: 10.1038/s41598-025-07821-w. URL: https://doi.org/10.1038/s41598-025-07821-w.

Pereira, J. R. L. de (2024). *The EU AI Act and environmental protection: the case for a missed opportunity*. Accessed: 2025-09-18. URL: https://eu.boell.org/en/2024/04/08/eu-ai-act-missed-opportunity#:~:text=such%20as%20the%20Democratic%20Republic,Addressing%20these.

Prechelt, L. (1998). "Early Stopping - But When?" In: *Neural Networks: Tricks of the Trade*. Ed. by G. B. Orr and K.-R. Müller. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 55–69. ISBN: 978-3-540-49430-0. DOI: 10.1007/3-540-49430-8_3. URL: https://doi.org/10.1007/3-540-49430-8_3.

GDPR (May 27, 2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. Regulation. OJ L 119, 4.5.2016, p. 1–88; Corrigendum: OJ L 127, 23.5.2018. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

AI Act (EU) 2024/1689 (July 12, 2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 establishing harmonised rules on artificial intelligence.*

Regulation. ELI: http://data.europa.eu/eli/reg/2024/1689/oj. European Parliament and the Council. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

Rotenberg, M. (2025). "Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (Council Eur.)" In: *International Legal Materials* 64.3, pp. 859–902. DOI: 10.1017/ilm.2025.1.

Russell, S., P. Norvig, and A. Intelligence (1995). "A modern approach". In: *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs* 25.27, pp. 1–2, 27–28.

Russinovich, M. (2024). *Sustainable by design: Innovating for energy efficiency in AI, part 1*. Accessed: 2025-09-25. URL: https://www.microsoft.com/en-us/microsoft-cloud/blog/2024/09/12/sustainable-by-design-innovating-for-energy-efficiency-in-ai-part-1/#:~:text=Maximizing%20hardware%20utilization%20through%20smart,workload%20management.

Sajid, H. (2024). *DistilBERT: A Distilled Version of BERT*. Accessed: 2025-05-14. URL: https://zilliz.com/learn/distilbert-distilled-version-of-bert.

Sanh, V., L. Debut, J. Chaumond, and T. Wolf (2020). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. arXiv: 1910.01108 [cs.CL]. URL: https://arxiv.org/abs/1910.01108.

Schwartz, R., J. Dodge, N. A. Smith, and O. Etzioni (2019). *Green AI*. arXiv: 1907.10597 [cs.CY]. URL: https://arxiv.org/abs/1907.10597.

Shafi, O., C. Rai, R. Sen, and G. Ananthanarayanan (Nov. 2021). "Demystifying TensorRT: Characterizing Neural Network Inference Engine on Nvidia Edge Devices". In: *2021 IEEE International Symposium on Workload Characterization (IISWC)*, pp. 226–237. DOI: 10.1109/IISWC53511.2021.00030.

Shafik Hebous, N. V.-L. (2024). *Carbon Emissions from AI and Crypto Are Surging and Tax Policy Can Help*. Accessed: 2025-09-28. URL: https://www.imf.org/en/Blogs/Articles/2024/08/15/carbon-emissions-from-ai-and-crypto-are-surging-and-tax-policy-can-help#:~:text=For%20policymakers%2C%20a%20broad%20carbon,85%20per%20ton%20by%202030.

Shuja, J., K. Bilal, S. A. Madani, M. Othman, R. Ranjan, P. Balaji, and S. U. Khan (2016). "Survey of Techniques and Architectures for Designing Energy-Efficient Data Centers". In: *IEEE Systems Journal* 10.2, pp. 507–519. DOI: 10.1109/JSYST.2014.2315823.

Shuja, J., K. Bilal, S. A. Madani, and S. U. Khan (Dec. 2014). "Data center energy efficient resource scheduling". In: *Cluster Computing* 17.4, pp. 1265–1277. ISSN: 1573-7543. DOI: 10.1007/s10586-014-0365-0. URL: https://doi.org/10.1007/s10586-014-0365-0.

Slota, S. C., K. R. Fleischmann, S. Greenberg, N. Verma, B. Cummings, L. Li, and C. Shenefiel (2023). "Many hands make many fingers to point: challenges in creating accountable AI". In: *AI & Society* 38.4, pp. 1287–1299. ISSN: 1435-5655. DOI: 10.1007/s00146-021-01302-0. URL: https://doi.org/10.1007/s00146-021-01302-0.

Stackpole, B. (2025). *AI has high data center energy costs — but there are solutions*. Accessed: 2025-09-24. URL: https://mitsloan.mit.edu/ideas-made-to-matter/ai-has-high-data-center-energy-costs-there-are-solutions#:~:text=A%20playbook%20for%20reducing%20emissions.

Stern, N., M. Romani, R. Pierfederici, M. Braun, D. Barraclough, S. Lingeswaran, E. Weirich-Benet, and N. Niemann (2025). "Green and intelligent: the role of AI in the climate transition". In: *npj Climate Action* 4.1, p. 56. ISSN: 2731-9814. DOI: 10.1038/s44168-025-00252-3. URL: https://doi.org/10.1038/s44168-025-00252-3.

Tan, M. and Q. V. Le (2020). *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*. arXiv: 1905.11946 [cs.LG]. URL: https://arxiv.org/abs/1905.11946.

European Green Deal (Dec. 11, 2019). *The European Green Deal*. Communication from the Commission COM(2019) 640 final. European Commission. URL: https://eur-lex.europa.eu/resource.html?uri=cellar:b828d165-1c22-11ea-8c1f-01aa75ed71a1.0002.02/DOC_1&format=PDF.

Thompson, D. F. (1980). "Moral Responsibility of Public Officials: The Problem of Many Hands". In: *The American Political Science Review* 74.4, pp. 905–916. ISSN: 00030554, 15375943. URL: http://www.jstor.org/stable/1954312 (visited on 09/18/2025).

Tolockaite, A., P. Tortorelli, and P. Stevens (2025). *In race to attract data centers, states can forfeit hundreds of millions of dollars in tax revenue to tech companies*. Accessed: 2025-10-12. URL: https://www.cnbc.com/2025/06/20/tax-breaks-for-tech-giants-data-centers-mean-less-income-for-states.html.

Vincent, J. (2024). *How much electricity does AI consume?* Accessed: 2025-04-24. URL: https://www.theverge.com/24066646/ai-electricity-energy-watts-generative-consumption.

w.A. (2005a). *About the EU ETS*. Accessed: 2025-09-28. URL: https://climate.ec.europa.eu/eu-action/carbon-markets/eu-emissions-trading-system-eu-ets/about-eu-ets_en.

– (2005b). *EU ETS emissions cap*. Accessed: 2025-09-28. URL: https://climate.ec.europa.eu/eu-action/carbon-markets/eu-emissions-trading-system-eu-ets/eu-ets-emissions-cap_en.

– (2008). *Hilton's AI-driven energy-cost reduction journey: over 1 billion dollar achieved in savings*. Accessed: 2025-09-25. URL: https://ei3.com/case-studies/hilton-ai-energy-management-savings#:~:text=%241%20Billion.

– (2018). *A brief history of data protection: How did it all start?* Accessed: 2025-09-18. URL: https://www.sourcing-international.org/news/a-brief-history-of-data-protection-how-did-it-all-start.

– (2021). *Finance and the Green Deal*. Accessed: 2025-09-29. URL: https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal/finance-and-green-deal_en.

– (2023). *Magic Eraser*. Accessed: 2025-05-15. URL: https://pixel.withgoogle.com/Pixel_8_Pro/use-magic-eraser?hl=en&country=US.

– (2024a). *AI has an environmental problem. Here's what the world can do about that*. Accessed: 2025-09-13. URL: https://www.unep.org/news-and-stories/story/ai-has-environmental-problem-heres-what-world-can-do-about#:~:text=%E2%80%9CGovernments%20are%20racing%20to%20develop,related%20safeguards.%E2%80%9D.

– (2024b). *Corporate Average Fuel Economy*. Accessed: 2025-09-17. URL: https://www.nhtsa.gov/laws-regulations/corporate-average-fuel-economy.

– (2024c). *Department of Energy Announces $68 Million in Funding for Artificial Intelligence for Scientific Research*. Accessed: 2025-10-12. URL: https://www.energy.gov/science/articles/department-energy-announces-68-million-funding-artificial-intelligence-scientific#:~:text=laboratories%2C%20to%20accelerate%20scientific%20programming%2C,algorithms%20and%20hardware%20for%20science.

– (2024d). *Ethics of Artificial Intelligence*. Accessed: 2025-08-25. URL: https://www.unesco.org/en/artificial-intelligence/recommendation-ethics.

– (2024e). *Strategie der Bundesregierung für Künstliche Intelligenz Umsetzungsplan 2024*. Accessed: 2025-08-27. Wien. URL: https://www.digitalaustria.gv.at/dam/jcr:a293b8fe-0054-4eea-a47b-653f84e4cefb/KI-Umsetzungsplan_2024-final19.5.2025.pdf.

– (2024f). *Sustainable Computing*. Accessed: 2025-09-25. URL: https://www.nvidia.com/en-us/data-center/sustainable-computing/#:~:text=Data%20Analytics.

– (2024g). *TensorFlow graph optimization with Grappler*. Accessed: 2025-05-17. URL: https://www.tensorflow.org/guide/graph_optimization.

– (2024h). *The Paris Agreement*. Accessed: 2025-08-22. URL: https://unfccc.int/process-and-meetings/the-paris-agreement.

– (2025a). *Abwärme aus dem Rechenzentrum sinnvoll nutzen: Ein Schritt in eine nachhaltige Zukunft*. Accessed: 2025-09-10. URL: https://www.digitalrealty.at/resources/articles/making-sensible-use-of-waste-heat-from-the-data-center-a-step-towards-a-sustainable-future.

– (2025b). *AI:AT – Die AI Factory Austria*. Accessed: 2025-09-09. URL: https://ai-at.eu/.
– (2025c). *Artificial intelligence ESG stakes*. Accessed: 2025-09-24. URL: https://www.ey.com/content/dam/ey-unified-site/ey-com/es-es/insights/rethinking-sustainability/documents/ey-artificial-intelligence-esg-stakes-discussion-paper.pdf.
– (2025d). *Decentralized Infrastructure: The Future of AI Workloads*. Accessed: 2025-09-08. URL: https://www.datacenters.com/news/decentralized-infrastructure-the-future-of-ai-workloads.
– (2025e). *Energy-efficient AI-ready Data Spaces*. Accessed: 2025-10-12. URL: https://greendatai.eu/.
– (2025f). *Greenhouse gas*. Accessed: 2025-04-2. URL: https://commons.wikimedia.org/wiki/File:Greenhouse-effect-t2.svg#/media/File:Greenhouse-effect-t2.svg.
– (2025g). *Mars Climate Orbiter*. Accessed: 2025-09-17. URL: https://science.nasa.gov/mission/mars-climate-orbiter/.
– (2025h). *Paris Agreement on climate change*. Accessed: 2025-04-25. URL: https://www.consilium.europa.eu/en/policies/paris-agreement-climate/.
– (2025i). *The Climate Crisis – A Race We Can Win*. Accessed: 2025-04-30. URL: https://www.un.org/en/un75/climate-crisis-race-we-can-win.
– (2025j). *The Importance of Standardized Units in Complex Operations*. Accessed: 2025-09-17. URL: https://www.epsilon3.io/behind-the-console/importance-of-standardized-units.
– (2025k). *The World's Total Data Center Capacity is Shifting Rapidly to Hyperscale Operators*. Accessed: 2025-09-08. URL: https://www.srgresearch.com/articles/the-worlds-total-data-center-capacity-is-shifting-rapidly-to-hyperscale-operators.
– (n.d.). *EuroCloud: Trusted Digital Competence Platform*. Accessed: 2025-09-09. URL: https://eurocloud.org/about/.

Wang, Y., Q. Wang, S. Shi, X. He, Z. Tang, K. Zhao, and X. Chu (May 2020). "Benchmarking the Performance and Energy Efficiency of AI Accelerators for AI Training". In: *2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID)*, pp. 744–751. DOI: 10.1109/CCGrid49817.2020.00-15.

Warren, S. D. and L. D. Brandeis (1890). "The Right to Privacy". In: *Harvard Law Review* 4.5, pp. 193–220. ISSN: 0017811X. URL: http://www.jstor.org/stable/1321160 (visited on 09/01/2025).

Weiss, K., T. M. Khoshgoftaar, and D. Wang (2016). "A survey of transfer learning". In: *Journal of Big data* 3, pp. 1–40.

Wu, C.-J., R. Raghavendra, U. Gupta, B. Acun, N. Ardalani, K. Maeng, G. Chang, F. Aga, J. Huang, C. Bai, M. Gschwind, A. Gupta, M. Ott, A. Melnikov, S. Candido, D. Brooks, G. Chauhan, B. Lee, H.-H. Lee, B. Akyildiz, M. Balandat, J. Spisak, R. Jain, M. Rabbat, and K. Hazelwood (2022). "Sustainable AI: Environmental Implications, Challenges and Opportunities". In: *Proceedings of Machine Learning and Systems*. Ed. by D. Marculescu, Y. Chi, and C. Wu. Vol. 4, pp. 795–813. URL: https://proceedings.mlsys.org/paper_files/paper/2022/file/462211f67c7d858f663355eff93b745e-Paper.pdf.

Wulff, P. (2020). "THE CLIMATE LEGACY OF SVANTE ARRHENIUS". In: *Icon* 25.2, pp. 163–169. ISSN: 13618113. URL: https://www.jstor.org/stable/26983759 (visited on 04/25/2025).

You, J., J.-W. Chung, and M. Chowdhury (Apr. 2023). "Zeus: Understanding and Optimizing GPU Energy Consumption of DNN Training". In: *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*. Boston, MA: USENIX Association, pp. 119–139. ISBN: 978-1-939133-33-5. URL: https://www.usenix.org/conference/nsdi23/presentation/you.

Zambaldi, V., D. La, A. E. Chu, H. Patani, A. E. Danson, T. O. C. Kwan, T. Frerix, R. G. Schneider, D. Saxton, A. Thillaisundaram, Z. Wu, I. Moraes, O. Lange, E. Papa, G. Stanton, V. Martin, S. Singh, L. H. Wong, R. Bates, S. A. Kohl, J. Abramson, A. W. Senior, Y. Alguel,

M. Y. Wu, I. M. Aspalter, K. Bentley, D. L. V. Bauer, P. Cherepanov, D. Hassabis, P. Kohli, R. Fergus, and J. Wang (2024). *De novo design of high-affinity protein binders with AlphaProteo*. arXiv: 2409.08022 [q-bio.BM]. URL: https://arxiv.org/abs/2409.08022.

Zhou, R., S. Khemmarat, and L. Gao (2010). "The impact of YouTube recommendation system on video views". In: *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*. IMC '10. Melbourne, Australia: Association for Computing Machinery, pp. 404–410. ISBN: 9781450304832. DOI: 10.1145/1879141.1879193. URL: https://doi.org/10.1145/1879141.1879193.

Zhou, Y. and W. Bu (2025). "From Artificial Intelligence to Energy Reduction: How Green Innovation Channels Corporate Sustainability". In: *Systems* 13.9. ISSN: 2079-8954. DOI: 10.3390/systems13090757. URL: https://www.mdpi.com/2079-8954/13/9/757.