Assessing Adoption Viability of Energy Reduction Efforts: A Case Study of Environmental Sustainability in AI Development

Tilburg University

Master thesis in Information Management for MSc Information Management

Author:

Victor Franklin Selgert

SNR: 2106986

v.f.selgert@tilburguniversity.edu

Supervisor Tilburg University:

drs. ing. Kenny Meesters

June 2024

Abstract

With the rise in complexity and ubiquity of Artificial Intelligence (AI) in everyday life, also comes a rise in energy consumption to train and use AI. Research into minimizing energy consumption of AI has been underway for a while but has not reached the industry practice. This is hindering organizations who seek to minimize energy consumption of AI products. This research aims to fill this gap by answering the question “How can the adoption viability of energy reduction techniques in AI development be assessed?”. To answer this question, this research identifies barriers to adoption of sustainability practices based on existing literature. These barriers and the influence they have on adoption of energy reduction techniques in AI development are then qualitatively analyzed and validated through a case study within an organization active in AI development. Based on these insights this research creates a framework which allows organizations to assess the viability of energy reduction techniques in AI development.

This research finds that using a combination of organizational and technological barriers, the adoption viability of energy reduction techniques in AI can be assessed and that the usage of the framework creates awareness of novel techniques. The assessment framework facilitates the transfer of knowledge between academia and industry and provides decision support to organizations looking to minimize energy consumption of AI products.

This research also finds that there is a high demand for quantitative decision support regarding the business case of environmental sustainability in software. Future research should aim to fill current gaps in the literature surrounding identifying and quantifying energy optimization opportunities and their costs and benefits to facilitate the creation of a robust business case.

Contents

Table of Tables

Table of Figures

1. Introduction

Artificial Intelligence (AI) has risen to prominence in the last few years, kickstarted by the release of the Large Language Model (LLM) ChatGPT 3. Since then, in an arms race to create the best AI, models have become larger and more complex. AI has been heralded as an important tool to tackle numerous challenges, from climate change to quicker vaccine development. However, the ever-increasing size and complexity of these models has also raised serious environmental concerns due to the substantial energy consumption required for their use and development. Research has shown that reducing the energy consumption of the development and usage of AI models is possible, yet the practical application of these techniques has lagged behind (Verdecchia et al., 2023). The goal of this thesis is to support businesses in assessing the adoption viability of these theoretical techniques. This enables businesses to take the most appropriate actions and facilitates the transfer of knowledge between academia and industry. This research will be conducted at Prime Vision, a medium sized enterprise located in the Netherlands. Prime Vision is a global leader in computer vision integration and robotics for logistics and fulfillment.

1.1 Defining Sustainability

Sustainability is a term that is widely associated with care for the environment and people, it is often used interchangeably with "green" or "renewable". The public discourse on climate change has propelled the term into the mainstream, it has also become a buzzword for many businesses that want to convey their corporate responsibility. But how sustainability is defined and what it entails, is not always clear.

The first popular use of the term sustainability and sustainable development stems from the 1987 United Nations (UN) report "Our common future". In this report sustainable development is defined as "*Meeting the needs of the present without compromising the ability of future generations to meet their own needs.*" (World Commission on Environment and Development (WCED), 1987). The report focuses mainly on environmental issues, but the meaning of sustainable development according to the UN would broaden significantly in the coming decades as evidenced by the creation of the Millenium Development Goals (MDG) and the Sustainable Development Goals (SDG). The effects on the environment and citizens are often the result of economic activity by businesses, thus they are an integral part of achieving sustainable development. In the early 90's, businesses and researchers were beginning to see the advantages of incorporating sustainability into their business practices due to mounting pressure from regulations and public opinion (Elkington, 1994). This pressure spawned a new way of thinking for businesses that now had to try to balance profits with their social and environmental image. This inspired the now famous "Triple Bottom Line" (TBL) commonly referred to as the three P's "People, Planet, Profit". The TBL defines the three dimensions of sustainability as social, environmental, and economic. (Elkington, 1997). Since the conceptualization of the TBL, there has been increasing research interest in measuring and improving sustainability performance (De Oliveira et al., 2023). In this thesis the focus will be on the environmental dimension of sustainability, this dimension encompasses the effect that certain products or activities have on the natural environment around them. This can manifest as emission of greenhouse gases or energy consumption.

There has been more attention for environmental sustainability in information communication technology (ICT) (Singh and Sahu, 2020). This increased attention is for a good reason, ICT is currently estimated to account for 1.8 – 2.8% of all Green House Gas (GHG) emissions, but the percentage could be as high as 2.1 to 3.9% (Freitag et al., 2021). For reference, the aviation sector is estimated to produce around 2.2% of all GHG emissions.

1.2 Problem Indication

Prime Vision has been receiving an increasing number of questions from customers regarding its sustainability efforts. This puts Prime Vision under pressure to improve upon the sustainability of its products and services. Currently, the willingness to develop more sustainable products and services is present among management. However, identifying opportunities for improvement and justifying these improvements is still a difficult hurdle to overcome. This means that it is very difficult to enact initiatives that address this topic.

Part of the problem is the lack of insights into the existence of optimization opportunities and the lack of methods to assess the viability of opportunities. The information needed to be able to govern sustainability in improvements in development is either not available or not structured in a way that facilitates use.

1.2.1 Academic Relevance

As AI models get larger and larger and the number of use cases for AI continues to expand, the energy consumption of AI is expanding with it. Energy usage of training and deploying increasingly large neural networks was most notably researched by Strubell et al. (2019). In this paper, the authors analyzed the energy consumption, computation cost and CO_2 emissions of training large Natural Language Processing (NLP) models. It was found that training and fine tuning a single large NLP model could emit as much as 36 tons of CO_2. The authors call for research into the design of more efficient algorithms and training practices to minimize the emissions of AI models. Schwartz et al. (2020) builds on the analysis of Strubell et al. (2019) and conceptualizes the terms “Red AI” and “Green AI”. The term red AI is used to describe AI models which solely focus on the accuracy and performance of the model without accounting for the skyrocketing computational requirements. The term green AI is defined as AI models and research which considers resource usage and achieve favorable performance/efficiency trade-offs. Since the publication of Schwartz et al. 2020) the number of papers studying green AI has grown significantly (Verdecchia et al., 2023).

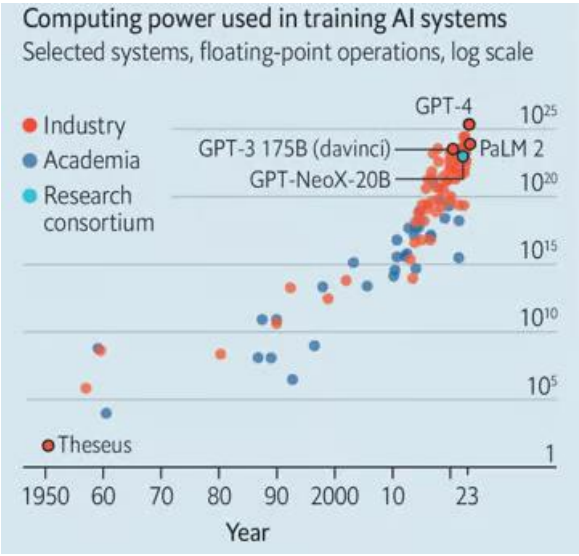

Figure 1: Sevilla et al 2023.

These studies regularly claim to achieve energy consumption reduction of 50% or more. These are promising results not only for the research field of green AI but also for industry and practitioners that want to reduce their energy consumption. However, the number of papers that include practical adoption is very limited, most papers are performed as laboratory experiments.

Furthermore, the number of papers where the intended audiences are industry and practitioners are marginal compared to the number of papers aimed at academics (Verdecchia et al., 2023).

A similar trend can be seen in the research field of conventional software energy optimization. Research contributions peaked in 2015 but steadily declined since, while industry involvement and practical adoption of promising energy optimization techniques is minimal (Balanza-Martinez et al., 2024; Pang et al., 2016). The drift into obscurity due to lack of adoption and industry involvement is a threat that now faces Green AI research. It is therefore important to study what is inhibiting the adoption of promising energy reduction techniques identified in green AI research.

Research in adoption of sustainability practices in several sectors has shown that businesses often lack a clear direction on where to start with initiatives. This makes it difficult to identify and choose relevant strategies and opportunities to improve sustainability (Saqib & Zhang, 2021; Johnson & Schaltegger, 2016). Furthermore, a combination of managers and practitioners needs to be involved in this process since consensus is needed on both sides to ensure adoption and use of the relevant techniques. Currently there are no ways to structurally evaluate the adoption viability of energy reduction techniques for AI models and their development.

The academic relevance and contribution of this thesis comes from the complementary nature of this research into the practical adoption of existing theories and techniques in energy reduction of AI.

1.2.2 Business Relevance

For businesses whose value creating activities involves developing AI enabled systems and applications it will be imperative that the sustainability performance of these activities and the systems themselves can be improved. Businesses that want to improve the sustainability of their software should be able to focus their efforts on the most effective areas for optimization to facilitate the creation of more energy aware software and a less energy consuming development process. The business relevance is clear, how can businesses assess opportunities that will have the most effect considering the existing barriers to adoption. This way the environmental sustainability of products can be meaningfully improved through targeted initiatives which have the best chance of success.

The result of this thesis should enable businesses to prioritize techniques and methods which can reduce the energy consumption of AI enabled applications and their development.

1.3 Research Questions

Based on the problem definition and the research objectives, the following research question and sub questions were formulated.

Research Question: “How can the adoption viability of energy reduction techniques in AI development be assessed?”

Sub Questions:

1. What is the current state-of-the-art in energy efficient AI development techniques research?
2. What are the barriers to adoption of energy efficient AI development techniques?
3. What influence do the barriers of adoption have on the adoption of energy efficient AI development techniques?

1.4 Research Method

This thesis consists of a literature review and a single mechanism case study, the literature review serves to uncover existing techniques for the energy reduction of AI systems and the identification of the barriers to adoption of these technologies in practice. The literature review thereby answers the two sub questions:

SQ1: What is the current state-of-the-art in energy efficient AI development techniques research?
SQ2: What are the barriers to adoption of energy efficient AI development techniques?

Once the relevant techniques and barriers are identified, these will inform the creation of an assessment method of techniques for practical adoption. The single mechanism case study serves to validate the barriers found in the literature and the adaptation of the theoretical assessment method through qualitative interviews and usability assessments with industry practitioners.

1.4.1 Scope

The goal of this thesis is the creation of a method to assess the adoption viability of techniques and methods that reduce the energy consumption of training and using AI models. This thesis focuses on assessing the viability of techniques from the perspective of organizational barriers to adoption. Since these techniques have been shown to be theoretically possible and effective, this thesis will only discuss technical barriers in relation to the ability of organizations to implement these techniques and not their technical feasibility. This research is conducted from the 1st of February until the 6th of June 2024.

It is important to define what is meant by AI in this thesis. AI is a term that is used to describe any program or situation where a computer mimics human intelligence. AI encompasses multiple techniques that are often used interchangeably with AI such as Machine Learning (ML), Neural Networks (NN), and Deep Learning (DL) (see figure 2). This thesis will discuss techniques which are applicable to different levels of AI. Some energy reduction techniques might be specific to DL while others can affect other facets of AI and AI development.

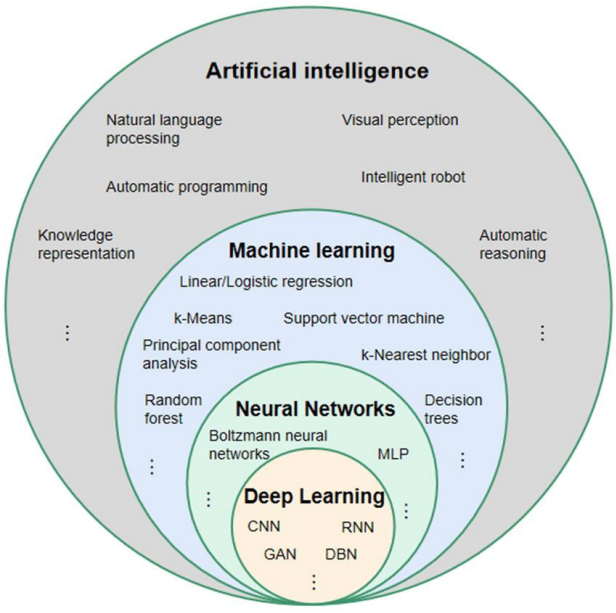

Figure 2: AI Onion Chart (Song et al., 2021)

1.5 Case study Description

Prime Vision was officially founded in 2003, the core products were based on innovative Optical Character Recognition technology (OCR). OCR is a computer vision technology which allows computers to recognize and process hand-written and machine-written text from an image. At the time this technology was used mainly for the purposes of sorting mail and handling bank statements.

Since 2003 Prime Vision has solidified its position in the postal and logistics market and has expanded its capabilities significantly to include sorting decision systems and even robotic sorting among other solutions. Prime Vision designs and integrates solutions using the latest recognition, identification, and robotics techniques to optimize the automation of mail and package sorting processes.

Prime Vision is owned by two major shareholders, PostNL and First Dutch Innovations. Currently the company consists of roughly 155 FTE making Prime Vision a Small to Medium Enterprise (SME). The portfolio of products is divided into three main sections:

- Intelligence
- Sight
- Movement

Intelligence

The intelligence portion of the portfolio focuses on the operating logic needed for sorting parcels and letters. The products in the Intelligence portfolio ensure that packages and letters are sorted to the correct chutes at the correct times. These software solutions are instrumental in processing millions of parcels and letters worldwide.

Sight

The Sight portfolio is aimed at providing computer vision solutions to improve sorting processes. These solutions include reading hand and machine written text on parcels and letters, recognizing hazardous goods symbols, and object and scene recognition.

Movement

The Movement portfolio focuses on the more portable part of the solutions. Among the solutions are sorting robots which sort parcels to the correct locations autonomously, and projectors which use computer vision to project sorting locations on parcels to improve manual sorting efficiency. The key driver behind the movement portfolio is enabling flexibility in logistics operations.

1.5.1 Operating Markets

Prime Vision operates in a few key sectors:

- Postal
- Courier, Express, Parcel (CEP)
- E-Commerce

The most mature of these sectors is the postal sector which is dominated by large national postal operators. This sector has traditionally been the main source of revenue for Prime Vision, but the mail volume is now declining.

CEP has been a fast-growing market, which Prime Vision has capitalized on and wants to continue expanding in. This market saw explosive growth during the COVID pandemic and continues to show steady growth.

E-Commerce is a newer market into which Prime Vision is entering, more and more E-Commerce operations are integrating logistics into the core business activities which causes competition with the established players. Prime Vision wants to use their expertise in improving sorting operations to capitalize on this competition.

Prime Vision operates worldwide and has an office in the Netherlands and an office in the United States. These locations also reflect the main customer base of the business, western/northern Europe and the USA.

1.5.2 Strategy

Prime Vision has traditionally been a software company with close ties and good relationships with its customers. This has manifested itself in a way of working that is project based instead of product based. High customization of solutions to fit the customer's situation was the norm for a long time. However, this way of working has limited scalability and is burdened by complicated support for many custom solutions.

Prime Vision is currently in a transitioning phase from project-based work to more standard products and solutions. This approach brings a new level of ambiguity to product development, Prime Vision has to be more proactive in predicting what the customer will value in the future instead of customizing the solution on the fly to fit with customer needs.

It is imperative for Prime Vision that the relationship with the customer is preserved and strengthened, there are only a limited number of large players that have a need for Prime Vision's products and expanding within an existing customer's operations is a large driver of revenue for Prime Vision. Furthermore, contracts are often multi-year and integration within the customer processes increases the chances of winning future contracts.

The focus on future value of products is also applied to sustainability in product development, Prime Vision predicts that customers will have a greater interest in the sustainability of solutions going forward. Therefore, the strategy calls for products that minimize energy consumption, maximize reusability of software and hardware, and maximize the recyclability of hardware.

1.5.3 Case study focus

This case study will focus on the department responsible for the sight portfolio of the organization. This department consists of three subdivisions:

- Product
- Research & Development
- Solutions

The product team works on guarding and further standardizing the development process. They provide the tools and standards for the research and solutions teams to do their work. For example, the product team creates the pipelines for training deep learning models in a way that works with the organizations' storage systems.

The Research and Development team works on finding new optimizations and techniques to improve the products. This includes finding research papers and experimenting with the newest techniques.

The Solutions team is at the end of the development line, they work to finalize models and products for the customer to ensure final performance. This team trains the models on the appropriate data and optimizes for specific customer requirements and often works with stricter deadlines.

1.5.4 Description of the OCR process

The largest application of machine learning models at Prime Vision has the purpose of recognizing and reading machine- and handwritten characters on packages and letters, this can be in the form of addresses and names but also images and characters that indicate warnings about the contents of a package. To achieve this, it utilizes a pipeline of data transformations and machine/deep learning models. Depending on the specific application the steps may vary, but this pipeline can be summarized with four generalized steps:

Pre-processing

To improve image quality and facilitate the best model results, certain transformations are made to the image before being fed into the model. The pre-processing step serves to standardize the input data as much as possible to ensure consistent results. Techniques like noise reduction and binarization are used for this purpose.

Text/object detection

After this data is transformed, it is fed to an object recognition model, often a Convolutional Neural Network (CNN), to detect regions of interest (ROI) where the text is located in the image. In the context of packages and letters, ROIs could be postal code, street address and name.

Text /object recognition

Once the ROIs are identified, the characters within these regions must be identified. To achieve this another Neural Network is used, this can be a transformer model which leverages attention mechanisms to recognize characters or a more conventional CNN or Recurrent Neural Network (RNN) model depending on the requirements and practical application.

Post-processing

Once the characters have been read, the next and final step is post-processing. This step can improve the accuracy of the text recognition by matching the output of the text recognition step to expected outputs and formats. An example of a technique used in this step is format parsing which structures the output of the text recognition step into the required format.

1.5.5 Process of model selection and fine-tuning

When choosing a machine learning model, developers look at the requirements for accuracy and latency. Then a suitable existing model architecture is found that fulfills the requirements. These models can be pre-trained to achieve many different accuracy and latency tradeoffs. After the model has been selected, it is trained on the relevant image data and fine-tuned for accuracy and speed by improving the data set and loss functions. The selection of a new model only happens occasionally when a completely new model architecture is needed. In most cases there are standard models for a specific task that can be retrained on new images. The cycle of finding, researching and fine tuning a model can take 6-12 months and many training sessions. While smaller optimizations can take 2-4 weeks.

1.6 Thesis Structure

This thesis will start by exploring the existing literature on energy reduction techniques in AI development to get an overview of the current state of the art. Next, current literature on barriers to adoption of sustainable practices will be reviewed and consolidated into the most influential barriers. In chapter 4 the preliminary assessment framework based on the literature will be created, this framework will be used as the basis on which iterations and revisions will take place based on the results of the case study. Chapter 4 will discuss the research methods and design chosen for this thesis and will expand on the argumentation for each of those. Chapter 5 will present the results of the interviews and the operationalization and subsequent revisions and validation of the assessment framework.

In chapter 6 the results and the implications of the assessment framework will be analyzed and discussed, this chapter will also review additional insights encountered during the research process. Finally in chapter 7 the conclusions, recommendations, and future research directions will be discussed. In this chapter the limitations of this thesis will also be laid out and reviewed.

2. Literature Review

In this chapter the current state of the literature will be explored. This literature review aims to find answers to the following sub questions defined in section 1.5:

SQ1: What is the current state-of-the-art in energy efficient AI development techniques research?
SQ2: What are the barriers to adoption of energy efficient AI development techniques?

This chapter will be structured as follows: First the current literature on energy efficient development techniques will be discussed, this serves to give an overview of the landscape and the possible techniques and measurements which are available to businesses that want to improve the energy efficiency of their AI models and development. Next, current research on barriers to adoption of sustainable practices will be reviewed and subsequently adapted to the context of AI and AI development.

2.1 Energy reduction techniques in AI

It is important to review the landscape of existing energy optimization techniques in AI to get a better understanding of their uses and feasibility. In recent years, research into the energy and resource efficiency of AI has become more frequent as the size of AI models has exploded (Verdecchia et al., 2023). The energy optimization measures in the relevant research can be divided into three main categories:

- Hardware-level optimizations
- Model-level optimizations

2.1.1 Hardware-level optimizations

AI models used to be constrained to the usage of CPUs, this limited the size and performance of larger AI models. This is because CPUs are not optimized for massive parallel processing but for higher clock speeds which prioritizes complex sequential serial processing. This is an issue for AI models since the number of computations required is high but the computations themselves are not complex. GPUs on the other hand can better parallelize the computations required for AI models which consist mostly of matrix multiplications and vector operations. To draw a comparison, modern high-end CPUs can have up to 128 cores while modern high-end GPUs have up to 15.000 cores. This shows a clear advantage to the usage of GPUs for training and running AI models (Shahid & Mushtaq, 2020). Moreover, GPUs are more energy efficient for training and running AI models compared to CPUs (Kang et al., 2020).

The increasing computational requirements of large AI models, in particular neural networks, and the need to deploy these models on embedded systems has driven the creation of more specialized hardware to improve the throughput and latency of AI models while using fewer resources (Shahid & Mushtaq, 2020). Two of the most important hardware innovations in this space are the Tensor Processing Unit (TPU) by Google, and the use of Field Programmable Gate Arrays (FPGA) for machine learning workloads.

The TPU is a type of Application Specific Integrated Circuit (ASIC) which is specifically designed to process Tensor multiplications and additions. Tensors are vectors with N-dimensions and the most common calculations in neural networks are Tensor multiplication and addition. This optimization causes a TPU to be more power efficient in both the training and inference phases for most neural network-based models. This specialization does make them less versatile for other workloads, so the choice of using TPUs should be carefully considered.

The FPGA is a processing unit of which the purpose and architecture can be changed after production to fit the required workload using Hardware description language (HDL). The energy efficiency of FPGAs is better than that of CPUs and GPUs while its performance outpaces that of a CPU but not a GPU, this makes FPGAs ideal for usage in resource constrained devices. The drawback of FPGAs is that the on-chip memory is quite low, which inhibits its use with very large models (Mittal, 2018).

Committing to a certain hardware architecture to accelerate machine learning applications or improve energy efficiency can be an expensive undertaking which severely impacts the development process of neural networks. However, there are also possibilities to manipulate the existing hardware infrastructure to reduce energy consumption.

An example of such a method can be found in You et al. (2022), in this paper the authors outline a tool which co-optimizes the power limit on a GPU with the batch size used during training to reduce the overall energy consumption of training a deep neural network (DNN). This paper finds that a reduction of 15.3% - 75.8% in energy consumption can be achieved during the training phase of a DNN. It is also found that the manipulation of the power limit and batch size can have a negative impact on training time, however the highest power limit and batch size possible will still result in a longer training time than many feasible power/batch size configurations (see figure 3). This result is also found in Krzywaniak et al. (2022), this study looks at the impact of different power limits on the time required for training. In this study it was found that the energy consumption of training will always be lower when a power limit is set on the GPU and the time required will always be higher. The achieved energy reduction ranges from -16.9% to -32.5% and the increase in time ranges from +4.5% to +35.8%.

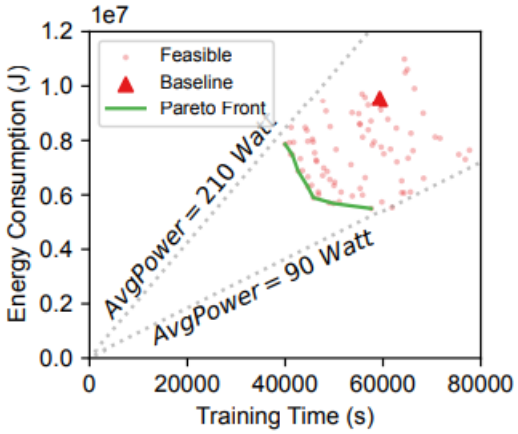

Figure 3: Energy-Time trade-off (You et al., 2022)

The current landscape of energy optimization on the hardware-level consists mostly of specialized hardware for running neural networks which require significant changes in model development processes. This makes changing to specialized hardware platforms an unlikely choice for businesses that solely want to improve their energy efficiency. However, the possibility of optimizing usage of existing hardware can be an attractive option with less risk and investment involved while still achieving significant results.

2.1.2 Model-level Optimizations

Energy consumption reduction can also be attained by altering the AI models themselves, performing these optimizations for the purpose of energy reduction is currently mostly a research activity and has only been scarcely adopted in practice (Verdecchia et al., 2023).

In deep learning, Hyper-parameters are the parameters of a deep learning model which cannot be changed by training the model, and thus must be set independently of the training phase (Yu & Zhu, 2020). Common hyperparameters include the Learning Rate, the Stochastic Gradient Descent (SGD) and the Batch Size. These hyperparameters can have an influence on the performance and accuracy of the model and are traditionally optimized for these two factors. However, the hyperparameters can also have an influence on the energy consumption of a model. Researchers have studied this effect, and some have included energy consumption or a proxy for energy consumption in the criteria with which the hyperparameters are optimized (De Chavannes et al., 2021).

For the search of the best hyperparameter values there are different methods, ranging from manual selection and brute force grid search to statistical search methods over the hyperparameter space. Optimizations can also be found in this task, for every combination of hyperparameters tested, the model needs to be trained at least partly, which in turn consumes energy. This stage of experimentation in model development is often resource-intensive (Wu et al., 2021). If the optimal hyperparameters are found more quickly it can reduce the number of training cycles required and thus decrease energy consumption. According to research by Strubell et al. (2019) and Turner et al. (2021) the optimal hyperparameter search algorithm is Bayesian optimization.

Another way to decrease the computational requirements of AI models is model compression. Model compression in deep learning is the operation of reducing the overall size of the model by removing weights (pruning) or reducing the size required to represent the weights in the model (quantization). The goal of model compression is to reduce the size of the model while maintaining the required accuracy and performance as much as possible. In the domain of energy consumption, these techniques can deliver reduced energy requirements for models since they are effectively reducing the number of computations required to run the model which in turn reduces the resource usage of the model.

Hubara et al. (2016) explored energy reduction of CNN-training trough quantization-aware training. The research concludes that quantization-aware training can lead to drastic reductions in power consumption and computation speed while minimizing accuracy loss. In quantization aware training the activations and weights of a model are quantized to a lower precision during training, this allows for quantization of the model with reduced accuracy loss due to the model being trained using lower precision.

However, Quantization aware training is more complex and requires extra training cycles to train the model with the reduced precision weights or activations. A simpler alternative to quantization aware training is post-training quantization. Post training quantization negates the need for retraining with the quantized weights, but consequently can cause lower model accuracy since the model is not trained on the possible errors induced by the quantization. Strategies for post-training quantization for energy efficiency improvement are explored in Bai et al. (2021) and Guo (2018). However, Guo (2018) also warns of the complexity of implementation of some of these strategies, and that they might not be practical for widespread use.

Another model compression technique which can be used to reduce the energy consumption of a neural network is pruning. Pruning reduces the size of the model by removing weights that are not

as useful in determining the outcome of inference. There is a trade-off in pruning a neural network, if too many weights are pruned from the model, it can have an impact on the accuracy (Hoefler et al., 2021). With AI becoming more and more power-hungry, pruning techniques that focus specifically on reducing the energy consumption of a model have been identified by researchers. Yang et al. (2017) defines a pruning algorithm which prunes weights based on their estimated energy consumption. With this approach Yang et al. (2017) were able to reduce the energy consumption select CNN's by up to 3.6 times.

2.2 Barriers to Adoption of Energy Optimization Techniques in AI Enabled Applications

After exploring the state-of-the art of energy optimization techniques and methods in AI, it is important to understand why these methods and techniques are not widely adopted in practice. What is inhibiting practitioners from applying the relevant techniques to reduce energy consumption? To answer this question, the literature on barriers to adoption of sustainable practices in a variety of sectors and contexts will be reviewed. Because of the novelty of AI and AI energy reduction techniques, literature into adoption of innovative technology will also be reviewed. The existing literature does not conform one to one with the barriers that might influence adoption of energy reduction techniques in AI development but will provide a better understanding of the factors. At the end of this chapter, the identified barriers will be interpreted and adapted to the context of AI development to form a clear catalogue of probable inhibiting factors. The adapted barriers are displayed in table 2. The structure of table 2 is based on the research of Deely et al. (2020) and research by Emmerloot (2020) into the barriers to information sharing.

Tornatzky et al. (1990) developed a framework called Technology, Organization, and Environment (TOE) explaining the process of technological innovation within organizations. Tornatzky et al. (1990) argues that these three factors also determine the adoption of innovative technology within an organization. Since its inception it has also been used to study green innovation adoption (Alraja et al., 2022; Hwang et al., 2016; Gohoungodji et al., 2020). TOE is a generic framework and is thus very adaptable to different contexts but is not as useful for assessment or evaluation without changes (Zhu & Kraemer, 2005; Baker, 2011). Consequently, this literature review will adapt the TOE framework to focus on the technological and organizational barriers with the purpose of informing an internal assessment method of energy reducing techniques in AI development.

2.2.1 Technological Barriers to Adoption of Sustainable Practices

Every sustainable practice and technology brings challenges and characteristics which can influence adoption. Cooremans (2012) argues that the characteristics of investments in energy efficiency like the complexity and the number of organizational changes they require also influence the adoption. An example in AI development is found in Guo (2018) where the author remarks on the complexity of a model compression technique and the extra effort that it requires which could inhibit its widespread use. Complexity is also noted as an important factor in green innovation adoption by Weng and Lin (2011) arguing that complexity can hinder information sharing and thus slow or prevent adoption. Weng and Lin (2011) also find that compatibility of an innovation plays a role in adoption due to an organization's tendency to prefer techniques that are already within, or close to, existing knowledge and skill levels.

Other examples of these practical barriers can be found in the literature on traditional barriers to sustainable practices. Caldera et al. (2019) finds that the time investment required to implement sustainable practices was seen as an inhibiting factor. This implies that the adoption of certain techniques for the purpose of sustainability is influenced by the effort required to implement them.

This is also remarked by Ervin et al. (2012) who finds that the significant upfront time investment discouraged corporate environmental management in SMEs.

The impact of sustainability practices on the production processes has also been found to form an obstruction to adoption. Olsthoorn (2015) states that businesses expressed serious concerns over the interruption and disruption of production processes when considering the adoption of sustainable practices. This effect was stated by Arvanitis and Ley (2012) to be a main barrier to adoption of energy saving technologies. Moreover, business owners feared that the adoption of these practices could impact the quality of the product (Fleiter et al., 2019; Olsthoorn et al., 2015).

2.2.2 Organizational barriers to adoption of sustainable practices

The focus of research into sustainability adoption has traditionally been on sustainability practices in heavy industry since this sector contains the heaviest polluters. Barriers defined in this early research were often financially motivated and focused on capital investment opportunities and risks. While these risks are not present in the same way for software and AI development, the need for businesses working in AI development to get a favorable return on investment remains relevant, regardless of whether that is through investment in hardware or investment of developer time and resources.

In Cooremans (2007), an example of perceived lack of return on investment in energy efficiency is showcased. In this research it is shown that low energy costs for businesses can mean that any energy reduction initiative will not be seen as strategic. Furthermore, Journeault et al. (2021) found that managers are often unaware of the impacts and benefits of sustainability practices. Even in situations where managers are aware of the costs and benefits, according to Friedman and Miles (2002) the knowledge and information to measure and assess the potential impacts and benefits is not present. This means that the underlying business case for sustainability practices is perceived as weak. This especially affects SMEs since they often do not have the same amount of available resources to invest in sustainability practices as larger corporations (Saqib & Zhang, 2021).

A shortcoming of existing research is the focus on the managers of businesses. Employees are rarely the target of questions regarding sustainability practices (Collins et al., 2010; Journeault et al., 2021; Cooremans, 2007; Friedman and Miles, 2002). This is understandable when the perspective is solely a financial one. However, when expertise and awareness of employees is an important factor, estimations made by managers are simply not a robust indication. Furthermore, factors pertaining to human resources such as perception and awareness were found to be important factors for the adoption of sustainability practices (Gómez-Bezares et al., 2019; Lueg et al., 2013).

In Lenox and Ehrenfeld (1997), the authors argued that the creation of environmental design capabilities is highly dependent on the organization's knowledge resources, the communication links with those resources, and the ability to effectively integrate this knowledge into the design processes. The importance of knowledge resources is echoed by Pereira et al. (2020) and Pang et al. (2016) who found that a heavy lack of knowledge and support for sustainable software design was a significant barrier to the adoption of energy reducing techniques. Other results from these studies showed that programmers are aware that their software influences energy consumption, but do not possess the knowledge to measure or reduce energy consumption. This lack of skill and expertise among management and employees is also noted as a barrier to adoption of sustainability practices in Journeault et al. (2021).

An issue raised in Pinto and Castor (2017) is the lack of tools available which support programmers in measuring and reducing the energy consumption of software. Finally, the lack of feedback from users and customers on the energy consumption of software also contributes to a lack of incentive for energy reduction among programmers (Pang et al., 2016).

Other key factors in sustainable practice adoption center around the motivation and awareness of sustainability by key stakeholders such as managers, employees, and customers (Journeault et al., 2021). Each of these stakeholders can hold a different perspective on the usefulness and value of implementing sustainable practices. Where employees can have intrinsic motivation to improve the sustainability of products, they cannot do so without the approval of management. In a similar vein, management can strive to improve sustainability of the products, but if the customer does not attach value to the sustainability of a product, management is less inclined to implement sustainability practices (Murillo & Lozano, 2006; Sen & Cowley, 2012).

This vicious cycle can lead to the stagnation of adoption of sustainability practices after the low-hanging fruit has been implemented. This stagnation is also noted in the work of Dooley (2017) where it was found that organizational inertia inhibits businesses from implementing sustainability initiatives in new areas, rather preferring focusing on continuous improvement of the current initiatives. It was found that this was the case even for low effort and high effect initiatives, showcasing that technological barriers alone cannot explain a lack of adoption.

To break this inertia, Journeault et al. (2021) identifies different roles that external stakeholders can have in kickstarting further progress in sustainable practice adoption. These roles are trainer, analyst, coordinator, specialist, and financial provider. The trainer promotes awareness, the analyst helps identify and assess new opportunities, the coordinator facilitates in implementing initiatives, the specialist provides technical support where necessary and the financial provider supports businesses with capital to fund initiatives. By collaborating with people or entities that can fulfill these roles, the adoption of sustainability practices can be promoted. These stakeholders do not have to be external, internal stakeholders can take on these roles if they are motivated to do so.

However, every business is different with regards to the attitude of stakeholders, which is why a generic solution to encourage adoption of sustainability practices is not identified in the current literature. Different businesses will experience different barriers and may perceive barriers differently based on the context of their sector and their stakeholders (Parker et al., 2009).

2.2.3 Construction of barrier table

To ensure the understandability of the different barriers and to account for variance in formulation by the authors of the original papers, the barriers are consolidated to represent the most significant barriers according to the reviewed literature. This consolidation can be found in table 1.

To construct the barrier adaptation table, both literature discussing the barriers to adoption and literature on energy efficient AI development techniques was reviewed. The literature is reviewed based on table 1, if the barriers or underlying factors are discussed in the paper as inhibiting adoption, the relevant sections of the literature will be cited in table 2 and adapted to the context of energy reduction in AI development.

Organizational Barriers	***Factors***
Lack of perceived benefit	Lack of perceived economic value Lack of perceived strategic value
Lack of awareness	No/low awareness of environmental impact No/low awareness of environmental practice
Lack of resources	Lack of human resources Lack of financial resources
Lack of information	Lack of metrics Lack of support
Technological Barriers	***Factors***
Complexity	Lack of experience Lack of expertise
Effort	-
Disruption	Loss of quality Process disruption
Energy Reduction	Low perceived effectiveness

Table 1: Consolidated Barriers

In this research *Lack of perceived benefit* is defined as the perception of an employee or decision maker that the adoption of energy efficient techniques in the AI development process do not provide the business with any upside. This lack of perceived benefit can be caused by other barriers including but not limited to *Lack of awareness* and *Lack of resources* (Journeault et al., 2021; Friedman and Miles, 2002). *Lack of awareness* is defined as the lack of awareness of environmental-impact, practices and/or goals in the line of work of the employee or decision maker. *Lack of resources* is defined as a situation where a business has a scarcity of resources and has difficulty devoting existing resources to new initiatives without losing business value. *Lack of information* is defined as the absence of information which is required to enable decision-making and tracking performance of energy efficient techniques.

Effort is defined as the amount of work required to implement the technique in question. *Complexity* is defined as the technical complexity of the technique; this refers to the rarity of the knowledge required to implement the technique. If highly specialized knowledge is needed, then the *complexity* of the technique will be high. *Disruption* is defined as the impact that the adoption of the technique would have on the existing processes, tools, and quality. *Energy reduction* refers to the theoretical energy savings achieved by this technique in the literature, this does not guarantee the same energy savings in practice but serves to give an estimation of the potential upside of the technique.

Literature	Relevant quotes	Adaptation to context of AI energy reduction	Barriers
(Pang et al., 2016)	"Programmers Lack Knowledge of Reducing Software Energy Consumption." "To reduce software energy consumption, programmers must start by measuring the energy consumption of their software. Only 12 respondents (10 percent) said they did this." "These results show that these programmers lacked knowledge of how to accurately measure software energy consumption." "Programmers Are Unaware of Software Energy Consumption's Causes" "Our survey results show that the programmers rarely addressed energy efficiency and that users rarely requested it. Only 22 respondents (18 percent) claimed to take energy consumption into account when developing software." "The fact that only 3 percent of the respondents received complaints about software energy consumption might suggest that users are unaware of it."	Not only is there a lack of knowledge in reducing energy consumption, but there is also a lack of insight into what causes energy consumption. Measuring software energy consumption is a rarity which reinforces the lack of insight into energy consumption drivers. There is a lack of incentive to reduce energy consumption due to customer apathy.	Complexity Lack of Information Lack of Perceived Benefits

(Pinto & Castor, 2017)	"Developers currently do not fully understand how to write, maintain, and evolve energy-efficient software systems. In this study we suggest this is primarily due to two problems: the lack of knowledge and the lack of tools." "Software developers currently have to rely on Q&A websites, blog posts, or YouTube videos when trying to optimize energy consumption, which are anecdotal, not supported by empirical evidence, or even incorrect."	There is a distinct lack of information resources available to programmers which can help them understand and reduce the energy consumption of software.	Lack of Information Complexity
(Olsthoorn et al., 2015)	"Concerns that energy efficiency measures may disrupt the production process and lead to revenue losses or affect product quality together with uncertainty about cost savings."	Reducing energy consumption of deep learning models can cause trade-offs with performance and may not integrate with existing processes.	Disruption
(Cooremans, 2007)	"Energy efficiency projects are not considered strategic due to the share of energy costs being rather low."	If reduction of energy consumption has a low potential for cost saving it will be a less attractive option.	Lack of Priority Energy reduction
(Weng and Lin, 2011)	"The difficulty in learning and sharing tacit technological knowledge makes it relatively difficult to adopt a complex technology. Therefore, the adoption of green innovations for SMEs is expected to be negatively associated with the perceived complexity of	If energy reduction techniques are perceived as too complex, it may discourage implementation and inhibit adoption. When energy reduction techniques can work within existing processes and pipelines, the chances of adoption are increased.	Complexity Disruption

	the innovations” “Green innovations that are more compatible to a company’s current technologies will be more easily to be diffused within the organization.”		
(Journeault et al., 2021)	“Lack of awareness of the impacts and benefits associated with sustainability.” “The low priority given to sustainable development issues within SMEs can be attributed to the fact that managers are often unaware of their firms’ social and environmental impacts.” “A number of studies report that lack of employee training in sustainable development and limited sustainability expertise among management staff are two significant barriers to the adoption of a sustainable development policy.” "SMEs suffer from a lack of time and resources [...] limited human and financial resources and time constraints are significant barriers to the implementation of sustainability initiatives within SMEs."	When awareness of environmental impact is low, the chance of implementation of energy reduction techniques is lowered. Lack of training and expertise among staff can also mount a serious barrier.	Lack of perceived benefit Complexity Lack of awareness Lack of resources

(Friedman and Miles, 2002)	"Where managers are aware of such costs and benefits, firms often lack the necessary information or knowledge to be able to accurately assess and measure them." "Time and resources were frequently cited as the major hurdles to implementation."	Without the ability to estimate and measure the effects on energy consumption of initiatives, the chance of adoption is lowered. When existing projects and activities take up all available time and/or compute power, the possibility for sustainability related activities is reduced.	Lack of Information Lack of Resources
(Arvanitis and Ley, 2012)	"Lack of compatibility with current product programme or current production technology seems to be the main barrier for firms that hinder them from adopting any kind of energy-saving technologies."	If integration within existing processes and frameworks is not possible or very difficult, adoption will be less likely.	Disruption
(Johnson & Schaltegger, 2016)	"The lack of awareness of sustainability issues is the first shortcoming frequently attributed to the reasons of limited implementation of tools by SMEs." "Given that SMEs have fewer employees, staff are usually responsible for, or at least involved in, more than one business function […] Because of this, they are usually required to focus on several different aspects of the organization simultaneously, making the addition of any new tasks or requirements difficult"	SMEs may not be fully informed about the environmental impacts of AI technologies. This lack of awareness can extend to the energy consumption of AI systems and the carbon footprint associated with training large models. Reduced availability of human resources due to scattered focus can inhibit time investment into new initiatives.	Lack of awareness Lack of human resources

(Brammer, Hoejmose, and Marchant 2012; Friedman and Miles 2002; Neamtu 2011)	"A second commonly discussed internal shortcoming is the absence of perceived benefits."	Management might not see a reason to invest time and resources into implementing energy reducing techniques due to the perceived lack of financial or strategic return.	Lack of perceived benefits
(Saqib & Zhang, 2021)	"It was explained that lack of the awareness of sustainable practices and its benefits are making the adoption more challenging for SMEs."	Practices identified in literature are not being diffused into industry processes due to uncertainty, causing reduced adoption. Benefits of energy reduction can either be unknown or can be seen as insufficient to warrant investment of time and resources.	Lack of awareness Lack of perceived benefits
(Lawrence et al., 2006)	"The perception they have little or no environmental impact compared with larger corporations."	When the perceived environmental impact is low, it lowers the willingness to implement improvements.	Lack of Awareness
(Fleiter et al., 2012)	"Technical risk of production interruption and product quality losses."	Reducing energy consumption of deep learning models and training can cause trade-offs with performance and speed.	Disruption
(Ournani et al., 2020)	"Need for a global score / KPI. This has been the most requested and discussed specification. Almost all the participants mentioned the need for a global score or Key Performance Indicators (KPI) for the total software energy consumption evolution."	It is hard for businesses to justify implementing energy reducing practices when the effects are not or cannot be measured.	Lack of Information
(Karita et al., 2021)	"When asked about the main barriers that hinder the adoption of sustainability actions and practices in the software development process of the corporate environment, 71% of the respondents stated that there is a lack of companies' awareness. Another 58%	With a lack of relevance and awareness comes a lack of motivation, if the motivation to reduce energy consumption is not present, no priority will be given to implementing energy reducing techniques.	Lack of Awareness Lack of perceived benefits

	understand that companies do not consider the subject as relevant."		
(Pereira et al., 2020)	"In fact, programmers many times seek help in resolving energy inefficiencies, showing that there are many misconceptions within the programming community as to what causes high-energy consumption, how to solve them, and a heavy lack of support and knowledge for energy-aware development."	when there is no knowledge base for programmers to draw from, implementation of energy reducing techniques are more difficult to implement due to lack of support.	Lack of information Effort
(Guo, 2018)	"Some of the methods need second-order information for updating the weights which leads to high computational complexity. From a practical perspective, it calls for more efforts to implement the proposed optimization algorithms which hinders their widespread use."	It can be very time and energy consuming to implement certain optimizations which can diminish their usefulness and usability.	Effort
(Parker et al., 2009)	"SMEs often have major problems with limited resources, limited knowledge, and limited technical capabilities to deal with their own negative environmental impact."	Due to the complex nature of AI development, a lack of expertise and knowledge regarding energy reduction techniques could increase the effort needed to implement and thus making implementation less likely.	Lack of Resources Complexity

Table 2: Adapted barriers to adoption of energy efficient techniques in AI development

2.3 Strength of Identified Barriers

In this section, the strength of the influence of the barriers according to existing research will be discussed.

Lack of Perceived Benefits

For the barrier lack of perceived benefits, the literature is clear on the severity of its impact on adoption. Especially Cooremans (2012) tackles this barrier in depth by comparing the perceived benefits of energy consumption reduction investments to their strategic value. In this study it is found that investments for the purpose of energy reduction face stricter requirements than other investments. Profitability was not seen as enough of a reason on its own to invest in energy reduction measures. Furthermore, the time to return on investment requirement was much shorter than other investments. De Groot et al. (2001) also finds that lack of priority is a barrier to adoption of readily available energy efficiency practices. The research finds that other investments were simply seen as more important to the core business. The study of De Groot et al. (2001) was conducted among Dutch companies in energy intensive sectors with a large opportunity for cost savings through energy efficiency improvements, highlighting the impact of the perceived lack of benefit barrier on adoption. These studies show that the lack of perceived benefits can have a large impact on the adoption of sustainable practices.

Lack of Awareness

Lack of awareness is attributed to inhibiting adoption of sustainable practices through to a lack of knowledge of environmental impact. If the knowledge of impact is low then businesses will not seek to remedy their environmental impact (Johnson & Schaltegger, 2016). General environmental awareness was also found to have a positive effect on sustainable practice adoption. However, positive attitude towards environmental sustainability alone was not necessarily found to be a contributing factor in improving sustainability in SMEs (Gadenne et al., 2008). Furthermore, the most significant inhibitor for adoption of sustainable practices found by Gadenne et al. (2008) was financial in nature. These studies show that a lack of awareness does indeed reduce sustainable practice adoption but is superseded by cost and benefit barriers.

Lack of Resources

Next to the lack of financial and strategic incentives defined by the lack of perceived benefit barrier, the lack of resources also plays a role in the adoption of sustainable practices. Businesses which lack the appropriate financial or human resources are constrained in implementing sustainable practices regardless of intention or attitude (Johnson & Schaltegger, 2016; Friedman and Miles, 2002). Especially within SMEs, human resources can be a large inhibitor to adoption. This is because employees within SMEs are more likely to perform multiple different roles (Johnson & Schaltegger, 2016). It is clear that a lack of resources can form a significant barrier to adoption. Here resources act as a facilitating factor, the required resources need to be present, but they do not directly affect priority or attitude towards sustainability practices.

Lack of Information

Lack of information is seen in research as an influential barrier to the adoption of sustainable practices, this barrier is especially noted in software development (Ournani et al., 2020). The need for key performance indicators (KPI) is reiterated in studies (Ournani et al, 2020; Pang et al, 2016). This information is necessary to track and evaluate any sustainable practices. This information would also aid in creating awareness. In Pereira et al. (2020) it is found that programmers do seek to fix

energy inefficiencies but simply do not have the correct information or support to tackle these problems. These studies show that the lack of information is a fundamental barrier to the adoption of sustainable practices in software development.

Complexity

Some literature implies that SMEs lack the technical expertise necessary to implement sustainable practices (Journeault et al., 2021). Complexity can also hinder knowledge sharing after implementation inhibiting widespread knowledge of the practice (Weng and Lin, 2011).

This means that the complexity of sustainable practices plays a role in the adoption of it. Furthermore, increasing complexity can increase the risk associated with the practice.

Effort

The effort required for the implementation and adoption of a sustainable practice can also constitute a barrier. This effect is noted by Guo (2019) where the implementation effort of a quantization technique could inhibit widespread use and adoption. Effort can also be related to cost in a development situation. Cooremans (2012) finds that costs are an influential factor in deciding to implement sustainable practices, by extension effort can thus act as a significant inhibitor.

Disruption

Disruption in the form of a loss in quality and process disruption that adoption of sustainable practices can cause can also form a barrier to their adoption (Fleiter et al., 2012; Olsthoorn et al., 2015). This can be of particular importance when quality requirements are strict or competition on quality is fierce. Disruption can thus be a major inhibitor to adoption of sustainable practices.

Energy Reduction

The effect that a sustainable practice will have on the actual energy consumption plays a role in the decision process. As discussed, perceived environmental impact can influence the willingness to implement remedies (Lawrence et al., 2006). If a practice offers significant energy reductions it can positively affect the decision for adoption.

2.4 Conclusion and Research Gap

The literature review has shown that there is a sizeable presence of techniques which can reduce the energy consumption of AI development and use, but that these techniques have not reached the mainstream industry practices. The research also shows that there is a need for tooling and support to incentivize further adoption of sustainable practices.

Existing research has delved into the barriers that organizations and especially SMEs experience when trying to implement sustainable practices. These barriers focus on the general organizational barriers that influence adoption. From the literature, a preliminary view of the strength of these barriers can be extracted. The lack of perceived benefits seems to have the strongest influence on adoption, the next strongest is the lack of resources, followed by the lack of information and lastly lack of awareness.

However, the identification of these barriers does not sufficiently enable organizations to mitigate them to improve adoption of sustainable practices. Technological characteristics of sustainable practices and their interaction with the existing barriers to adoption are also required for proper evaluation. This topic has not been researched explicitly which limits evaluation and by extension adoption of these practices.

This research creates a framework which bundles important barriers to adoption of sustainable practices and characteristics of energy efficient techniques in AI development. This framework allows for the evaluation of techniques based on these barriers and characteristics, thereby providing decision support to organizations, and enabling knowledge transfer from academia to industry.

3. Framework

Based on the literature review in the previous chapter a framework to assess the adoption viability of energy reducing techniques in AI development will be created. The framework will map the interactions between the identified barriers to adoption of energy reducing techniques in AI development and the characteristics of the techniques to come to an indication of the adoption viability of a technique for the target organization. The framework

3.1 Elaboration on the Framework

In this framework the most influential barriers to adoption of energy efficient techniques in AI are laid out in combination with the characteristics of the techniques which these barriers influence (see table 3). These barriers and technique characteristics are synthesized based on the literature review of known barriers and the interviews with practitioners and managers in the field of AI development.

In the top row of the framework, the barriers and characteristics are displayed. In the first column of the framework, the names of the techniques are displayed. The barriers comprise the first four columns after the technique column, the subsequent four columns represent the technique characteristics. The values of technique characteristics are variable between techniques, but the barriers are a property of the organization which is using the framework and will be static across the different techniques where an organization is concerned. This means that an organization can assess the barriers in a periodic manner while continuously evaluating different techniques based on these barriers. Barriers do not have to be re-evaluated for every technique.

The framework will be subject to evaluation and revision based on the feedback of the relevant stakeholders. This feedback will be incorporated iteratively, and the framework will then provide an answer to the research question: *"How can the adoption viability of energy reduction techniques in AI development be assessed?"*

	Barriers				Characteristics			
Barriers / **Techniques**	**Lack of Perceived Benefits**	**Lack of Environmental Awareness**	**Lack of Information**	**Lack of Resources**	**Complexity (L/M/H)**	**Effort (L/M/H)**	**Disruption (L/M/H)**	**-E%**

Table 3: Assessment framework for energy efficient development techniques in AI Development

3.1.1 Hypothesized relationships

The two main components that comprise the framework are the barriers to adoption and the characteristics of energy reduction techniques. The aim of this research is to map these two components and their interactions to create a method of assessing the feasibility of technique adoption for organizations. From the literature it is clear that the barriers to adoption have a negative influence on the adoption of sustainable practices, the barriers included in the framework represent the barriers that are mentioned in the literature most frequently and are found to have the biggest influence on adoption.

The second component of this framework comprises the technical characteristics of energy reducing techniques. This is a novel component that is mentioned in the literature only as implied factors that influence adoption. An example of this is found in Parker et al. (2009), where it is found that SMEs often lack the technical capability to implement sustainable practices. It is implied that the technical complexity of practices plays a role in the severity of the barrier, but it is not explicitly researched.

The relationships proposed in this research are as follows, (1) the characteristics influence the adoption of energy reducing techniques while (2) the barriers have a moderating effect on the effect produced by the characteristics. Here, the barriers are proposed to reinforce the negative effect created by the Effort, Complexity and disruption characteristics and decrease the positive effect created by the -E% characteristic. This means that the presence of barriers can exacerbate the negative effects of technique characteristics and minimize the positive effects. These proposed relationships are visualized in figure 4.

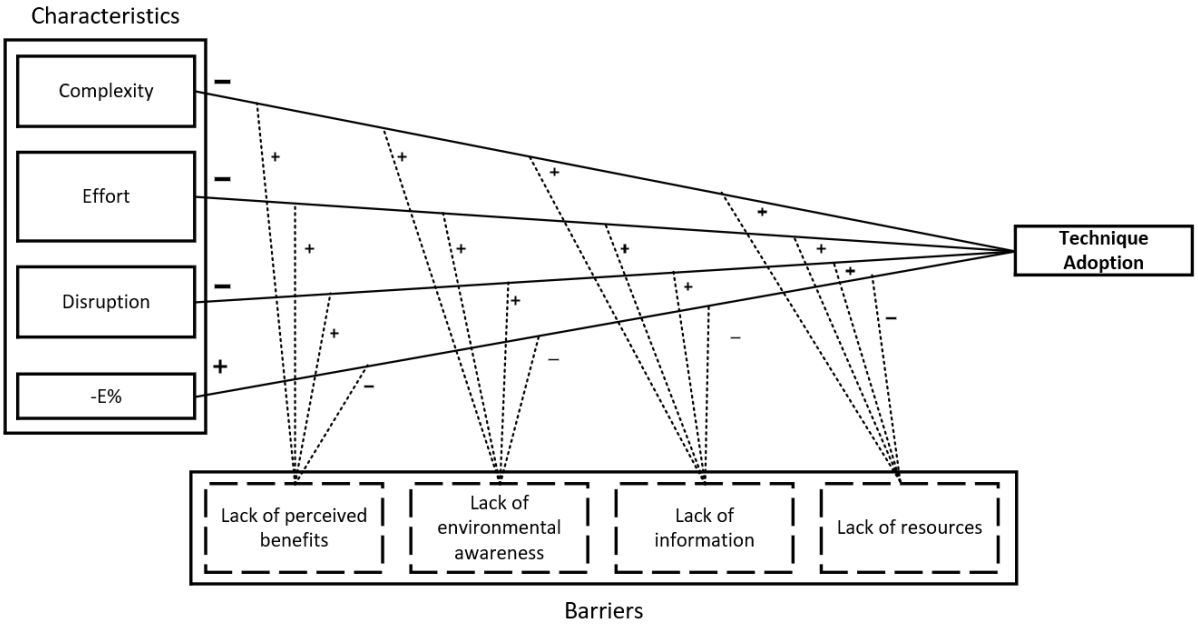

Figure 4: Proposed relationships between barriers to adoption and technique characteristics

Based on the existing literature and preliminary interviews, the interactions between the barriers to adoption and the technique characteristics will be assessed using table 4. The barriers and characteristics will be analyzed in a qualitative manner based on the interpretation of the literature and interviews by the author. This matrix will be used to validate the qualitative relationships between the barriers and techniques based on semi-structured interviews with practitioners and managers.

Characteristics Barriers	Complexity	Effort	Disruption	-E%
Lack of perceived benefits				
Lack of environmental awareness				
Lack of information				
Lack of resources				

Table 4: Analysis matrix barriers and characteristics

4. Research Design

This chapter contains the research strategy for this thesis for the purpose of the creation of an artefact which enables businesses to assess the adoption viability of energy reducing techniques in AI development.

4.1 Design Science Research

The design science research (DSR) method is a research methodology aimed at more pragmatic research while still maintaining academic rigor and useful academic contributions. Academic management research in the mainstream leans towards descriptive research which can affect the usefulness of the research in practice (Aken & Joan, 2004).

Because DSR is solution oriented but still needs to contribute to the academic knowledge base, design science problems need to be evaluated from two sides. Hevner et al. (2004) lays out the three cycles related to design science namely, the relevance cycle, the rigor cycle, and the design cycle.

The relevance cycle defines the requirements that need to be fulfilled and how the fulfillment of these requirements can be measured. This cycle ensures that the research improves the environment it is focused on (Hevner, 2007).

The rigor cycle ensures that the created artefact is grounded in existing theories of the research field. This ensures that the artefact created constitutes an innovation over existing theories and artefacts. However, Hevner (2007) states that the assertion that all DSR must be directly grounded in existing descriptive theories is unrealistic.

Finally, the most important part of DSR is the design cycle. The research activities conducted in the design cycle are informed by the rigor and relevance cycle and are iteratively applied and evaluated. The challenge in the design cycle is to maintain the balance between the rigor and relevance cycles. To have a strong grounding is not enough if the relevance and evaluation methods are weak (Hevner, 2007).

This thesis embraces design science methodology by creating a utilizable artefact with academic grounding while providing additional perspectives to contribute to the existing knowledge base.

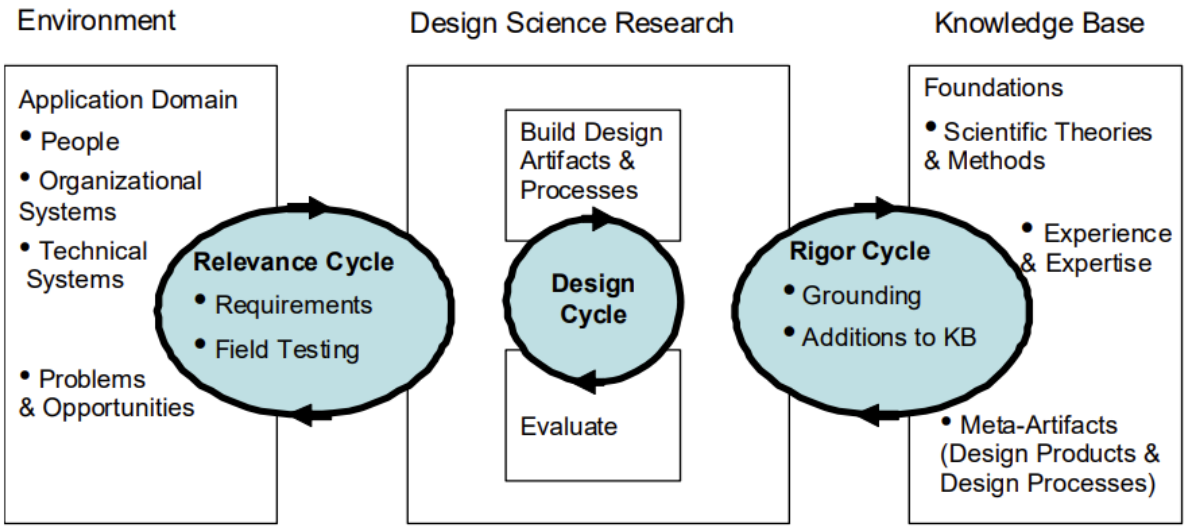

Figure 4: Design Science Research Cycles (Hevner et al., 2004)

4.1.1 Rigor Cycle

To ground this research in existing academic work and to ensure an academic contribution, a knowledge question based on the research is formulated which facilitates the design cycle. The knowledge question answered by this research is laid out in sub question two:

"What are the barriers to adoption of energy efficient AI development techniques?"

This question is answered through a literature review and semi-structured interviews. The literature review identifies barriers to the adoption of sustainable practices in many different sectors and disciplines while the interviews aim to verify the existence of these barriers in the context of energy efficient AI development techniques and adapt them to the relevant field. The literature review also identifies the basic characteristics of energy reduction techniques which interact with the barriers to adoption.

4.1.2 Relevance Cycle

To ensure that the created framework is relevant and fulfills the intended purpose, it is important to gather information on the environment that the artefact is intended to operate in. In this research this is achieved through semi-structured interviews and iterative exploration of artefact designs with relevant stakeholders. Semi-structured interviews will also be used to verify and identify the energy reduction technique characteristics that have the biggest influence on adoption.

4.1.3 Design Cycle

This research builds an initial artefact based on existing literature, this framework is then evaluated, improved, and then evaluated again. This research is limited in time and scope and thus will be limited to two improvement iterations of the framework, one for usability and one for validity.

4.2 Data Collection

In this section, the process and methods of data collection used in this research will be discussed and substantiated.

4.2.1 Problem exploration

In DSR the first step of valid research is the exploration of the problem and its context. In this research this exploration is achieved through informal conversations and semi-structured interviews with company employees and managers. For this exploration it is imperative that the formality of an interview is avoided at first since it can cause social and time pressure on the interviewee if the necessary precautions aren't taken (Myers & Newman, 2007). Once a connection has been established through informal communication, semi-structured interviews can be utilized to get a deeper understanding of the problem and its context. In DSR this is an iterative process, getting to the root of a problem requires time and revision (see Discussion).

In Behavioral Research, a gap in the literature will inform the research question. However, in DSR the research question might arise from a practical need rather than just a theoretical gap. This means that problem exploration and field research can occur before and during the process of literature review.

4.2.2 Case selection

As a result of the problem exploration, a case was selected which will maximally benefit from an artefact and which can facilitate the research activities. For this research, the case of AI development was chosen due to the current lack of adoption of sustainability practices and the availability of experts in their field that are open to interviews.

Interviewees were chosen based on their expertise and their ability to influence the adoption of sustainable practices in AI development. To ensure a complete picture of the barriers to adoption from multiple perspectives and the interaction with technique characteristics, interviewees were chosen from multiple layers of the organization. The roles chosen to interview were Computer Vision Engineer, Product Manager for the Vision department, commercial director, and Technology director.

Semi-structured interviews were chosen as the sole interview method, in semi-structured interviews, the interviewer will have prepared questions beforehand but can decide to not ask certain questions or improvise new questions based on the responses of the interviewee (Myers & Newman, 2007). The choice for semi-structured interviews was made due to the complexity of the subject matter of AI development which increases the chance that further clarification is needed. Furthermore, the current literature suggests that awareness of energy reducing practices in AI development is low which could lead to misinterpretation of questions if there is no room for additional explanation and improvisation. The interviews were recorded with consent of the interviewee and later transcribed to facilitate coding of the interviews.

4.2.3 Interview design

The design and execution of semi-structured interviews is crucial to the validity of the outcomes. For this research, the recommendations for qualitative interviews in Information System research by Myers and Newman (2007) are used (see figure 5).

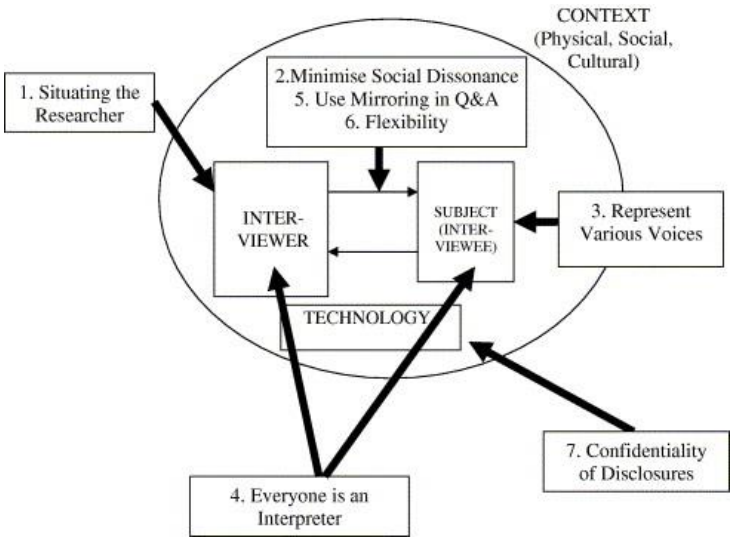

Figure 5: Recommendations for qualitative interviews in IS research (Meyers & Newman, 2007)

To ensure the validity of the interviews, an interview strategy was conceptualized based on the recommendations of Myers and Newman (2007). The first part of the strategy concerns the minimizing of social dissonance. To minimize the effects of social dissonance the interviewer will first approach the interviewee in an informal setting and introduce himself. Only after having established contact in an informal matter will the interviewer invite the interviewer to participate in the study. Next to represent various voices, the interviewees are chosen from different layers of the organization and triangulation of subjects is achieved through a combination of recommendations from previous interviewees and the “random” informal approach by the interviewer based on the relevant roles.

To facilitate the interviewee in expressing their own thoughts without enforcing the interviewer’s bias, mirroring of jargon and expressions will be used to rephrase subsequent questions.

Furthermore, the semi-structured nature of the interview will allow the interviewer to explore lines of questioning in a more flexible way. Finally, the interviewer will clearly describe the process of recording the interview and what will happen to the recording and the answers provided by the interviewee after the interview is over to ensure the confidentiality of disclosures.

4.2.4 Interview Protocol

To support the interviewer in conducting the interview and to facilitate comparison between the responses of different interviewees, an interview protocol is created (table 5). The first topic is an introduction of the research topic and the motivation for this research. It serves to set a clear subject of the interview and gives the interviewee the necessary context. Next the background information of the interviewee is collected for the purposes of validating the results of the interviews. What follows is the central part of the interview where the different barriers identified in the literature review are categorized into three topics, namely Awareness, Impact, and Willingness. The interviewer will ask questions which assess the existence of barriers identified from the literature. Next the possible remedies for these barriers will be explored, this will provide insight into the severity of the influence of the described barriers. The following topic is centered around the specific characteristics of energy reducing techniques which would make these techniques more or less attractive to implement, this will inform the severity of impact of technique characteristics on the identified barriers. Finally, the interviewer thanks the interviewee for the participation and asks for any final comments or questions for the interviewer.

Topic	*Contents*
Introduction	*Introduction of the author, the thesis research, and the motivation behind the research*
Background information interviewee	*Gathering necessary contextual information about the interviewee such as role, responsibilities, and tasks*
Awareness	*Assessing knowledge of interviewee on environmental sustainability, energy consumption, and energy saving techniques*
Impact	*Questions centered around the perceived environmental impact of products.*
Willingness	*Assessing the motivation of the interviewee to address energy consumption*
Remedies	*What could be done to make energy reduction more desirable?*
Technique Characteristics	*What would make energy reducing techniques attractive to implement?*
Closing	*Closing the interview, giving interviewee the opportunity to add any last points or ask questions to the interviewer.*

Table 5: Summarized interview protocol

4.3 Data Analysis and Validation

This research analyzes the data gathered by means of deductive coding based on the analysis matrix in table 4. Deductive coding was chosen as the main method of analysis due to the theoretical framework being already defined based on the literature. This framework will guide the coding process to attribute statements of the interviewees to the different characteristics, barriers, and their interactions.

The weight of the barriers, characteristics and their interactions will be based on the frequency of the occurrence in the transcripts, the interpretation of influence by the author, and framework evaluation interviews.

Based on the interviews conducted, an initial instance of the framework will be operationalized. This framework will consequently be presented for evaluation to intended users, these evaluations will then inform revisions to the framework and will shed light on its usefulness and the influence of the different barriers and characteristics.

4.4 Stakeholders Involved

The Chief Technology Officer (CTO) is the problem owner and oversees the effort of integrating sustainability in Prime Vision developed information systems. The CTO is in close contact with customers and is party to discussion on long term strategic goals of the customers which include sustainability goals. His goal with regards to sustainability is to measure, track and improve the environmental sustainability of solutions and product teams for the purpose of customer satisfaction and reporting. This goal is intertwined with measuring quality, productivity, and cost performance of the product team.

Developers. The developers should be a large part of the design process of the artefact since they can affect the energy consumption of the software through the source code. Furthermore, developers can provide crucial input on the factors influencing energy consumption and effort in the development process. The main driver of the developers is the ability to continue doing their work in such a way that does not compromise the quality of their work and does not require an exorbitant amount of time.

Product manager. For every product line at Prime Vision, a product manager is responsible for the overall development and evolution of the product. This stakeholder is an integral part in defining the product strategy and thus is an important player in implementing and adopting energy reducing initiatives.

Continuous Integration and Continuous Deployment (CI/CD) guild. The CI/CD guild is responsible for developing and integrating the future CI/CD software development infrastructure into new and existing projects. The artefact will ideally integrate, and at a minimum does not interfere, with the desired tooling laid out in the roadmap of the CI/CD Guild. The main driver of the CI/CD guild is the standardization of tooling used in future projects.

5 Results

In this chapter, the results of the research will be presented. The chapter consists of the results of the interviews conducted, the occurrence of the barriers and their interactions with the technique characteristics. Next the validation of the framework will be discussed based on the evaluation of practitioners.

5.1 Code Results

All codes were decided a-priori based on the literature and initial informal conversations held with stakeholders within the case study organization. The codes are taken from the framework created in chapter 3 and are split up into organizational barriers and technique specific characteristics. The codes were assigned to quotes within the interview text based on the coding manual set out in Appendix D.

In table 6 the occurrence count of the organizational and practical barriers in the interviews are laid out, this sheds light on what the interviewees see as the most important inhibitors of implementing energy reduction techniques in the products. An interesting insight that becomes clear when looking at the occurrences is that the operational layer of the organization is more focused on the technological barriers whereas the management layer is more focused on the organizational barriers.

The most occurring organizational barrier is Lack of information with eighteen occurrences, this shows that there is a difficulty in knowing where to start and how to find suitable energy optimizations in development. Both lack of awareness and lack of perceived benefits were also mentioned in almost every interview.

Among the technical barriers, disruption was seen as the most significant barrier with twelve occurrences. Potential energy savings weren't seen as much of a redeeming factor when compared to the barriers and complexity was not seen as a significant inhibitor to adoption.

		CV1	CV2	CV3	PM	TD	CD	Total
Barriers	Lack of Perceived Benefits	2	2	1	3	0	1	9
	Lack of Awareness	2	3	1	1	1	2	10
	Lack of Information	2	2	4	4	4	2	18
	Lack of Resources	1	2	0	1	1	0	5
Characteristics	Complexity	0	1	1	0	0	0	2
	Effort	1	1	1	0	0	0	3
	Disruption	1	5	3	2	1	0	12
	-E%	0	1	1	0	1	0	3

Table 6: Code occurrences

5.2 Interviews

In total, 6 interviews and 3 evaluations were conducted in this study. The interviewees are assigned an identifier code according to their role within the organization. CV stands for computer vision engineer, PM stands for product manager, CD for commercial director, and TD for technology director.

Identifier	Job title	Type	Date
CV1	Computer Vision Engineer	Interview	17-04-2024
CV2	Computer Vision Engineer	Interview	24-04-2024
PM	Product Manager	Interview	26-04-2024
CD	Commercial Director	Interview	01-05-2024
CV3	Computer Vision Engineer	Interview	07-05-2024
TD	Technology Director	Interview/Validation	13-05-2024
CV1	Computer Vision Engineer	Evaluation/Validation	14-05-2024
CV3	Computer Vision Engineer	Evaluation/Validation	15-05-2024
CV2	Computer Vision Engineer	Evaluation/Validation	16-05-2024

Table 7: Interview schema

From the interviews it is clear that the most important factor that is mentioned in every interview is the performance and quality of the product. Anything that could compromise this would not be worth implementing.

"The product should always adhere to the quality requirements and should be stable. And it (the energy optimization) should have a real impact."

"We should choose anything that saves energy as long as we're hitting performance, the amount of energy saved is secondary."

The possibility of integration of energy reducing techniques with existing processes and tools is also seen as an important factor which could inhibit adoption.

"It is less about the specific technique and more about being able to integrate it with the other parts of optimization and conversion."

Energy consumption is not discussed within the product team, it is not seen as an issue even though all engineers are aware of the resource intensity of training and running AI models. Interestingly, all interviewees are in favor of reducing energy consumption and all interviewees note willingness to do it but are not aware of exactly how to do this. This is also reflected in the code count as lack of information was mentioned the most out of all barriers.

"I think if you just start with some initial techniques that could start to reduce the energy consumption that is something that could help a lot, because that is something that we can do internally. If there could be something that could tell us like okay this function consumes more energy, so use this one"

"Within the team we really don't talk about energy consumption, it stays at the performance level. In contrast to some of the green initiatives flying around at PV."

Some engineers believe that adding energy consumption as another metric in model optimization would become too complex to work with.

"What we're not waiting on is another dimension to contend with when optimizing models, then you have to juggle so many different things and that becomes untannable."

Some engineers worry that improving energy consumption will take away time from other optimizations.

"If you want to optimize energy consumption, then the team will have to put work in, and that means that we have to put other work aside."

In general, there is willingness among engineers and managers alike to reduce energy consumption. This shows that for this case study the willingness aligns with the literature findings around willingness of programmers to reduce energy consumption (Pang et al., 2016).

"I honestly think that you would get no pushback or complaints from the engineers if you asked them to work on this, I think they would be very happy to do that." – Product Manager

"I think we give a lot of room to employees to engage with sustainability, but it also comes down to the motivation of the employees themselves if they want to use that room." -Commercial Director

*"**Interviewer:***
Do you think it would be worthwhile to start to reduce energy consumption where possible?

CV 3:
Of course, yeah." – Computer Vision Engineer

It is also apparent that even though the willingness is there, sustainability has not yet seeped into product development activities. Despite the rest of the organization pushing for sustainable initiatives.

"Sustainability is top-down, basically the company will first encourage the easy things and the development at Sight is not a low hanging fruit. It's also more IP and a value creating activity."

"The engineers are focused on what they do, they see that we have to be greener but when you look at the actions it's more like separating trash and taking the bike, not in product development."

Also noted by multiple interviewees is a switch from very customer driven development to productization and standardization. This is also noted as a reason for increased interest in sustainable product development within the organization.

"We are starting to go more towards productization and that requires looking further into the future and not just focusing on a single customer but on all potential customers and their interests as a whole"

5.3 Framework Operationalization and Validation

To support the practical application of the framework created in chapter 3, the framework should be operationalized in a way that facilitates assessments of energy reduction techniques in the case study organization. In this section this operationalization and the validation of the framework will be discussed.

To validate the framework, three additional interviews were held with vision engineers. The vision engineers were first asked to evaluate a small catalogue of energy reducing techniques in AI development according to the identified technique specific characteristics (see Appendix C). After they evaluated the techniques, an interview was scheduled to discuss their evaluations and the rest of the framework including the organizational barriers. The vision engineers were asked to evaluate the techniques on the characteristics on their own, and during the interview were asked to evaluate the organizational barriers in company of the interviewer. This approach was chosen due to the need for extra explanation of these organizational barriers to avoid overly subjective results.

5.3.1 Framework Usability Revisions

The analysis of interviews revealed that simply noting the presence of barriers in the coding of interviews does not adequately capture the nuances necessary to convey the extent of their impact. To address this, it is essential to revise the framework. Previously, the framework only provided a binary option to indicate the presence or absence of a barrier. The revised framework now includes three levels of severity for each barrier (Low/Medium/High), allowing for a more nuanced assessment. Furthermore, based on user feedback the framework was split in two to better reflect the fact that the organizational barriers are not technique dependent but organization wide. Other revisions were made based on feedback to improve usability of the framework. The barriers were no longer formulated as “A lack of” since this caused some confusion in interviews due to being counter intuitive. For example, if there are abundant resources then the lack of resources barrier would be low. Furthermore, the wording of the energy savings characteristic was changed to be clearer. The measure of energy reduction was also changed to Low, Medium, and High to ensure consistency in the measures.

	Characteristics			
Techniques	**Complexity (L/M/H)**	**Effort (L/M/H)**	**Disruption (L/M/H)**	**Energy Reduction (L/M/H)**

Organizational Barriers			
Perceived Benefits (L/M/H)	**Environmental Awareness (L/M/H)**	**Information (L/M/H)**	**Resources (L/M/H)**

Table 8: Revised framework

5.3.2 Framework Operationalization

The framework is operationalized using a modified weighted multi-criteria Analysis methodology. Each barrier and characteristic level is assigned a weight ranging from "- -" to "++", representing the factor's positive or negative influence on adoption. These weights are derived from the interviews and evaluations conducted within the case study organization. Each "+" counts as 1 and each "-" counts as -1.

The weights of the organizational barriers are then compared against the weights of the technique-specific characteristics. These calculations are used to sum up the positive and negative influences on the adoption of a specific technique, resulting in either positive, neutral, or negative advice for adoption based on user input. The interactions between these weights and the reference matrix can be found in Appendix E.

The weights are calibrated so that the thresholds for positive and negative advice are centered around zero. A positive adoption advice requires a sum greater than 2, while a negative advice requires a sum less than -2. Values between -2 and 2 result in neutral advice.

5.3.3 Weight Calibration

This section will discuss the initial calibration of the weights necessary for the operationalization of the assessment framework, both the interviews conducted, and the literature on the strength of barriers discussed in chapter 2.3 were considered. The weights calibrated from the interview phase were used to create an initial model which was then used in the evaluation interviews discussed at the start of this chapter. The complete calibrated weights can be found in Appendix E.

Perceived Benefits

In the literature the lack of perceived benefits was found to be the most influential barrier to adoption, this was not as clear in the interviews since the attitude towards improving environmental sustainability was generally positive. However, the interviews did show that even with perceived benefits in the form of strategic importance, the upside of this was modest. Therefore, this barrier is found to have a mostly negative effect with a small positive effect when perceived benefits are high.

Awareness

Lack of awareness was named in every interview as a barrier to adoption, it was however not seen as a very significant barrier to overcome and was sometimes attributed to a lack of priority within product teams. This is in line with the findings in the literature which note a lack of perceived benefits as a more serious threat to sustainable practice adoption. Therefore, the overall strength of this barrier was found to be moderate.

Information

In the interview phase, a lack of information was noted the most as an inhibiting factor for adoption of energy reduction techniques. Additionally, the strength of the barrier was seen as significant due to the impact it had on the ability to evaluate both the impact and the effort required of adoption. However, it was also noted that increased information availability would greatly improve decision making ability and could further incentivize initiatives. Therefore, the effect of low information availability was found to be strongly negative while high information availability was found to be strongly positive.

Resources

In the interview phase, a lack of resources was not often mentioned as a barrier to adoption. However, the need for available resources is vital to ensure the possibility for adoption of energy reduction techniques. That is why resources are seen as a facilitator for adoption, in line with this, the influence of the resource barrier is found to have an overall negative impact on adoption while high resource availability only boasts a moderate positive impact.

Complexity

Contrary to existing literature, the complexity characteristic was not seen by interviewees as a highly influential factor in deciding on adoption of a technique. Consensus among interviewees was that complexity should not be unreasonably high, but that complexity is not a big hurdle. This is why Complexity is found to have a negative effect only if it is deemed as high complexity.

Effort

The interviewees remarked on the effort of implementation as a potential barrier but were not explicit in the strength of this influence. This is why the influence of effort was initially calibrated to be moderately influential across all levels.

Disruption

Disruption was found to be the most influential characteristic by the interviewees. It was mentioned most frequently and was also seen as a major hurdle for the adoption of new techniques. Fear of quality losses dominated the responses from interviewees; however, process complications and disruptions were also among the concerns. Based on these responses, the influence of disruption was found to be negative even at medium levels and it was seen as a major positive if disruption could be avoided.

Energy Reduction

The amount of energy reduction achieved by a technique was seen by most interviewees as a nice to have and not a deal breaker or deciding factor. Also mentioned by interviewees was a sentiment of every little bit helps, this is reflected in the strength of the energy reduction characteristic. This characteristic is found to have a low overall influence and will only positively influence adoption if the energy reduction is sufficiently high.

5.3.4 Weight Revisions

Based on the evaluation interviews with vision engineers, the weights of technique characteristics were fine-tuned. The initial weights based on the interviews are displayed in table 9. The weights of the organizational barriers were not significantly impacted as a result of the evaluation interviews.

	Complexity	*Effort*	*Disruption*	*Energy Reduction*
L	0	+	++	0
M	0	0	-	0
H	-	-	--	+

Table 9: Technological characteristics weights before evaluation

Several changes were made to the weights based on the evaluation interviews (see table 10). Namely the influence of effort was increased, and the weight of medium disruption was decreased. Interviewees were more incentivized by effort than was found in the interview phase. Furthermore, the consensus among engineers was that some disruption was acceptable and most of the time unavoidable.

	Complexity	*Effort*	*Disruption*	*Energy Reduction*
L	0	++	++	0
M	0	0	0	0
H	-	--	--	+

Table 10: Revised technological characteristics weights

5.3.5 Validation Results

The outcome of the framework based on the input of the interviewees was discussed in the validation interviews. The interviewees were asked if they agreed with the assessment of the techniques and the advice that the framework provided. In all cases the interviewees agreed with the advice as presented by the framework with some comments, these comments were then used to revise the weights of the framework to better reflect their impact (see section 5.3.4). An important outcome from the validation interviews was that interviewees all mentioned having learned from the experience and expressed interest in following up and trying out the techniques with positive or neutral advice from the framework. This shows that the assessment framework is useful for facilitating knowledge transfer from academia to industry, promoting awareness, and aiding in assessment of techniques.

6 Discussion

This chapter will interpret the results and insights gained from this research and will link these insights to the literature to highlight missing pieces which still inhibit organizations from taking the next step in sustainability by incorporating it in the design and development of products. First the barriers from literature and their influence on the adoption of energy reducing techniques in AI development will be discussed. Next the influence and definitions of the technique characteristics defined in this research will be discussed.

Following this, a broader possible approach on software energy reduction and the shortcomings of this approach will be analyzed. And finally, the strategic value and organizational drive for sustainable products of the organization in this case study and organizations in general will be discussed.

6.1 Barriers And Their Influence

The most influential barrier found during interviews and evaluations was a lack of information. For many interviewees this barrier was at the core of why energy reduction is not yet integrated into the development process of AI models and solutions. Also seemingly influencing the severity of other barriers.

"For us it's a question of where do we start, and how can we measure the impact."

"Sustainability is a primary concern for us, but I can't yet see the benefits on the software side."

"Because we don't know the benefits and costs, we can't estimate how many resources we need."

This lack of information causes uncertainty in the value of sustainability in the software development process. According to management closer to development, there is a dire necessity for the creation of a business case for sustainability in software development. Management also admits that the integration of sustainability into product development might end up costing more than it will earn. Another consequence of a lack of information is the decreased ability to assess other barriers. This effect is not studied in this research, but future research should be conducted to incorporate this effect into an improved methodology for sustainable technique and barrier assessment.

In this case study an interesting deviation from the literature is found for the barrier "lack of resources". It was noted by the interviewees much less than expected. A lack of resources was only mentioned in passing but was not actually seen as an issue. Instead, the issue shifted to fit more with a perceived lack of benefit, the interviewees did not perceive a barrier due to a lack of resources but merely saw it as a lack of priority. One interviewee exemplified this by stating:

"I think incentive is a big part of it, if the customer comes tomorrow with a requirement for energy efficiency, then we would do it."

This could be explained by the fact that the case study organization is used to working on a project-by-project basis where the customer pays for the development, so if energy reduction is a requirement, then the customer should pay for it. However, the product manager explained it in a more nuanced way by stating:

"I think the driver is more on the operational side, if we have to invest 200 hours to make our product greener, it will either come from the profits or it should be invested into a roadmap for the product. We are currently transitioning to new AI systems, and perfecting those should be step 2."

This suggests that in the case study organization, resources are not perceived as limited. Instead, the prioritization of how resources are allocated appears to be the more significant inhibiting factor. This raises questions about the barrier 'lack of resources,' especially when there's an overlap with perceived benefits. For instance, if an organization allocates resources to all its priorities up to the fifth one, but places sustainability as the seventh priority, the issue might not be resource scarcity but rather a low priority assigned due to perceived insufficient benefits. While theoretically, infinite resources would allow every area to be addressed regardless of its priority, practical limitations necessitate choices. So, even with resource availability, prioritization and by extension 'lack of perceived benefit' can itself act as a barrier. This maintains the relevance of the 'lack of resources' and 'lack of perceived benefit' barriers, as it can indicate that high-priority areas might still be neglected due to actual resource limitations.

Another deviation from literature is seen in the influence of the characteristic 'Complexity'. In the literature on software energy reduction and sustainable practice adoption, the consensus is that programmers lack knowledge on reducing energy consumption and that SMEs often lack the technical ability to implement sustainability related initiatives (Pang et al., 2016; Caldera et al., 2019; Parker et al., 2009). However, 'Complexity' was rarely mentioned as a barrier or hindrance to adoption by the interviewees. Energy reduction techniques were simply seen as another problem to be solved, no more complex than optimizing for performance or latency. This is likely due to the inherent complexities of AI development which requires a high level of expertise and research skill.

The technique specific characteristic 'Effort' was not often named as an inhibiting factor by the interviewees. This factor was more often seen in conjunction with 'Disruption' and 'Perceived lack of benefit'. This could be traced back to the fact that resources are not perceived as scarce within the organization. This mindset can cause effort of implementation alone to be seen as less of an inhibiting factor. However, the interviewees did use general terms to describe the disruption factor which could be interpreted as the effort factor as well. One interviewee stated that in all cases the process should stay "workable", another noted that it shouldn't become "too much". These statements were nonetheless classified as the disruption factor since they did not allude to the effort of implementation, but the effect that adoption has on the development process which can cause more effort to be put in after implementation.

The biggest concern related to adoption of techniques was the disruption that this would cause in the process and the possibility of product quality loss as a result of the technique. Disruption was not only named the most frequent but also as the most impactful when evaluating a technique. This finding is likely because it impacts both the value of the product and the direct working activities of the developers. This is also consistent with the main driver identified in the stakeholder assessment of developers in chapter 4.4.

The amount of energy saved was not seen a much of a redeeming factor in the face of other technical barriers Employees and management alike would rather save a little energy as opposed to a lot, if it means that the integrity of the process and product quality can be upheld. This indicates a clear priority for performance and quality over sustainability.

6.2 Energy Reduction in Software Development: Trade-off Model Approach

Some academic research suggests that using a trade-off approach to managing sustainability is the preferred option. The trade-off approach states that an organization should make clear choices and trade-offs when dealing with sustainability. This can allow for the creation of win-wins where environmental practices are chosen based on economic value, thereby avoiding the conflicting tension of profitability versus sustainability (Bansal, 2005; Margolis & Walsh, 2001). The employment of the trade-off approach to create win-win scenarios is often viewed through an instrumental view, where the creation of a business case is a central component (Van Der Byl & Slawinski, 2015).

Within the case study organization, a model where the costs and benefits of interventions for the purpose of energy reduction in software can be weighed against one another was highly requested. Therefore, this research explored the creation of such a trade-off model using the instrumental view where the economic and environmental value of environmental sustainability initiatives in software development could be analyzed. It was found that the creation of a trade-off model in the case study organization is not yet feasible. This section will explore the required knowledge and enablers to facilitate the creation of such a trade-off model.

For the design of a trade-off model the most fundamental factors to understand are the Energy Optimization Opportunity and the Energy Optimization Cost.

6.1.1 Energy Optimization Opportunity

Finding the Energy Optimization Opportunity is dependent on the identification and quantification of energy saving measures in the existing software.

Identification

For the identification of energy inefficiencies there are two ways to analyze an application, static analysis, and dynamic analysis. Static analysis is the analysis of source code without running it. This means that it will not interact with the hardware at all. Dynamic analysis is the analysis of a program or system while running the program or system on hardware.

Static analysis

This approach requires that energy inefficient code, also known as “energy smells”, are identified and catalogued. This is a popular research topic for mobile application development. Goaër and Hertout (2022) identified 40 energy smells specifically for the Android platform. This study creates a plugin for the popular static code analyzer SonarQube. The impact of these energy smells on energy consumption is not measured. Further research into energy aware programming and refactoring is done by Couto et al. (2020), Cruz & Abreu (2017), and Morales et al. (2018).

Attempts to catalogue energy smells in traditional server-based applications have been less frequent. Gottschalk et al. (2012) identifies eight generic energy smells that could contribute to unnecessary energy consumption. However, in this research, the effects of these energy smells on energy consumption are not measured and the study still uses the Android platform to identify and refactor these energy smells.

Another important finding was made by Verdecchia et al. (2018), in this study it was found that the measured static software metrics could not give an indication of the energy behavior of the tested open-source applications, this implies that simple static analysis is not enough to estimate or benchmark software energy consumption. This study also finds evidence that performance and

energy consumption are not always related, this dilutes the possibility of using performance metrics as a proxy for energy efficiency.

Within Prime Vision, the need for automatic identification of energy smells in the code is high, the manual identification of energy inefficient code would be time consuming and would require additional training since the awareness of energy optimization techniques among developers is low. The tool in use for static analysis at Prime Vision is SonarQube, this tool does not support the identification of specific energy smells for server-based applications. It does support some performance-based code smells but, as discussed, these cannot reliably be used as a proxy for energy smells.

Dynamic analysis

As opposed to static analysis, dynamic analysis is the analysis of the application during run-time or testing under load. From dynamic analysis, factors such as CPU, GPU and DRAM utilization can be gleamed. If these measures are combined with the specific source code locations that consume these resources, insights into optimization locations can be found. One way to do this is through a profiler, a profiler is a tool that reads the performance counters of the system and can couple these counters with the source code being executed at that moment. This allows for the identification of specific resource heavy methods and functions in the software. These methods do not necessarily indicate whether the use of resources is inefficient, however they can give insight into where possible optimizations would have the most effect.

Currently, profilers are not used consistently within Prime Vision, the use of profilers is seen as a last resort to solve particularly stubborn performance issues after they have been localized. Profilers are not used to profile complete applications to spot potential issues.

To gain insight into the behavior of an application it must be profiled under load to get a representative result. Prime Vision performs load testing of applications in a public cloud environment managed by the customer. This significantly hampers the ability to profile the application for the relevant resource usage and power consumption. For both use cases, access to the underlying hardware registry is needed to fetch the performance counters. Public cloud providers often offer their own performance counters, but these are not compatible with existing profilers, do not include power consumption, and cannot be reliably used to estimate power consumption.

Running an application or application component under load on a hardware test bench, isolated from the rest of the environment is also not an option since the dependencies in the cloud environment are needed to perform a representative test.

Quantification

Verdecchia et al. (2018) study the energy impact of refactoring common non energy specific code smells in software. This study finds that in the best-case scenario for large applications, the energy consumption was reduced by up to 49.9% and performance was increased by up to 47.8%. However, it was also found that refactoring all identified code smells resulted in a decrease in performance of 6.8% but an energy savings of 10.7%. This shows that there is a need for energy specific code smells, which currently do not exist for many programming languages and platforms. Furthermore, the impact of these individual code smells must be quantified to estimate the optimization opportunity.

6.1.2 Energy Optimization Cost

The energy optimization cost consists of two main sub factors, namely optimization energy cost and optimization effort. Where the energy cost consists of the required energy to optimize the software, and the energy optimization effort consists of the developer effort to implement the optimization in the software.

Energy

The energy cost of implementing an optimization to the energy efficiency of an application is comprised of the required energy of the workstations of the developers and the energy consumption of the CI pipeline. For the energy cost of a developer's workstation, an estimation of power consumed can be made using a windows PowerShell command 'powercfg.exe'. With this command, the energy usage of devices with a battery can be polled and divided into specific applications which use this power. The issue with this approach is that it records all power consumption of the developer and not just when the developer is working on a specific issue.

The energy cost of the pipeline is more difficult to estimate since the pipeline also runs in the cloud. This causes the same problems as with energy measurement of software in the cloud.

Effort

The second part of the equation is the effort it takes for a developer to implement the energy optimization in the software. There are numerous software project effort estimation techniques which can be used to determine effort; however, these methods are very coarse grained and serve to give an overall estimation of effort over the course of large software projects. These methods are rarely accurate in their estimations and are not suited to estimate the effort of small or iterative workloads. This leaves the quantification of effort to the individual developers' estimation. This is an issue when it comes to energy efficiency optimizations since the awareness of these optimizations is very low. Ultimately this means that estimations by developers are not a reliable measure of effort which inhibits the creation of a robust cost estimation framework necessary to construct a trade-off model.

6.1.3 Gap to a trade-off model approach to energy reduction in software

The state of existing research and practice do not facilitate the creation of a trade-off model for energy optimization of modern cloud-based applications. To enable the creation of such a model, there would have to be additions to the existing literature and practice. The first addition to the literature would have to be the creation of a catalog of generic energy smells to enable the identification of energy inefficient behavior in source code. Secondly, the relative impact of these energy smells would have to be catalogued so that the energy optimization opportunity can be quantified.

Where practice is concerned, the first necessary change would be the addition of energy metrics and the supported use of profilers by cloud providers. Currently, some cloud providers share carbon emission numbers on a subscription level, which is a very course level of granularity and does not allow for analysis of emissions or energy consumption of applications or systems (Vos et al., 2022). Other cloud providers do not provide these numbers at all. Furthermore, the carbon emissions of cloud data centers are variable and not just the result of the energy efficiency of the applications running in the datacenter. For example, the energy mix of a datacenter could become greener, and this would lower the carbon emissions. However, this would not indicate an increase in the energy efficiency of an application making the metric unsuited for tracking improvements.

Finally, the awareness of energy efficiency in cloud-based applications among programmers is low, this means that the ability to effectively estimate effort or impact is currently not present. The creation of a catalogue of energy smells as mentioned previously would already alleviate some of the consequences of low awareness through the creation of a knowledge base that programmers could draw from.

All in all, the creation of a trade-off model for software energy consumption for cloud-based applications is missing key aspects in both research and practice. Currently the creation of such a trade-off model would not provide businesses with the required information to be a useful tool for identifying and prioritizing energy optimization efforts.

6.3 Organizational Drive for Sustainable Products

The broader question underlining this research is what drives organizations to improve the environmental sustainability of their products. This section will explore this drive for sustainability and the tensions that arise because of it.

6.3.1 Corporate Legitimacy

Current literature examining the relationship between corporate sustainability and organizational performance has not come to a consensus on whether sustainability has a positive or negative effect on organizational performance or if it even influences organizational performance at all (Singh & Misra, 2021). This begs the question why are organizations investing in becoming more sustainable?

According to Scherer et al. (2013) the drive for sustainable development within organizations is not necessarily fueled by organizational performance but by the need for corporate legitimacy. There are three strategies that organizations can utilize to gain legitimacy: Isomorphic adaptation, strategic manipulation, and moral reasoning (Scherer et al, 2013). Within the case study organization, normative isomorphism is used as the strategic approach to sustainability. Here, customer action and expectation are used as the main argument for the propagation of sustainability within the organization. This approach could be explained by the focus on customer relationships and the switch to standardized products in the case study organization, necessitating the projection of shared values and goals towards customers.

However, what became clear by interviewing employees and management is that moral reasoning is also employed as an internal legitimacy strategy. This is achieved by management through deliberation with employees as the most important internal stakeholders. This can act as a proactive strategy in dealing with future sustainability issues.

6.3.2 Organizational Tensions

Uncertainty leads to tensions within organizational goals, there is no certain financial return on investing in sustainability. However, the capability to produce environmentally sustainable products can become a strategic capability in the future. Should organizations focus only on improving the profitability of the product now, or does the or should organizations invest in uncertain future capabilities?

This sentiment of uncertainty is also experienced within the case study organization. There is an expectation that the sustainability of products will become a necessity to compile competitive offerings, but the effects of this are uncertain and not yet quantifiable.

“There is a competitive side to it, we expect customers to prefer a green product because they are working on it themselves.”

"We cannot quantify the financial value of offering green products."

Profitability might be impacted by the adoption of initiatives which improve the environmental sustainability of products, however due to the paradoxical nature of the profitability versus sustainability tension, any choice between them will only cause short term gain before the tension rises again. This is why paradox theory advocates for the acceptance of tensions and managing both sides simultaneously to ensure long term success (Lewis, 2000).

Currently the case study organization is trying to manage both exploration and exploitation in the domain of sustainability simultaneously by integrating new ideas into existing products.

"What we see is that we cannot offer a "green option" to customers and ask them to pay extra. It has to be an integral part of our products."

Another tension arises in the case study organization as a result of sustainability around who takes responsibility for improving sustainability of products. It is clear from the interviews that the whole organization is willing to improve the sustainability of products. However, who this responsibility should fall on is still ambiguous. Both management and operations see themselves as facilitators. If it needs to be done, we will do it. However, management expects initiative from employees and employees expect initiative from management.

"It's not like employees don't get time to work on sustainability, I think that we spend a good amount of time and give quite some space to facilitate that. But in the end, it also comes down to the motivation of the employee themselves."

"I think management should attach enough value to it (sustainable product design). It could be more expensive to do that but then management should say: this is a feature that we want because it helps in achieving our green goals"

Currently these tensions are latent, they are not in the foreground of discourse on sustainability within the case study organization. However, these tensions could become salient when environmental factors shift (Smith & Lewis, 2011). Currently customers are not asking directly for more sustainable products, however this will likely change. And when it does change, these tensions become salient and can form obstacles for efficient adoption and management of initiatives.

Another latent tension within the case study organization was found through observations, informal conversations, and interviews. Some felt that sustainability was being pushed too forcefully and not being thought through.

"I think if you really want to do something for the environment, there are other things that you can do before trying to optimize your models. So, I think that the impact here is low. I am a vegetarian and I think I am doing the best for the environment in that way, and I think that the meat industry does way more damage to the environment than our GPU training."

Some also shared the sentiment that making products sustainable was simply done for the sake of marketing while they thought that the environmental impact was very small and that resources would be better spent improving the products. There is some awareness among management that these sentiments could exist:

"Maybe the operational side doesn't share our idea of the value of sustainability, if you're on the shop floor so to say you might not have the same information, you might not feel the need to change anything. Possibly also because you don't know what the alternative would be"

There is a seeming contradiction in the statements made by engineers. While the general consensus is that the amount of energy saved is not a crucial factor when evaluating techniques, with a sentiment akin to "every little bit helps", some employees feel that the environmental impact of their work is too minimal to justify significant intervention. These statements appear contradictory but together suggest that the perceived environmental impact of their work is too minor to warrant efforts to reduce it. If energy reduction, the only positive aspect, is not considered significant, it implies that other negative factors are more influential in the decision-making process. This indicates a perception among engineers that the benefits of these efforts are low, even though the attitude towards reducing energy consumption in the interviews was positive.

When examining what drives pro-environmental behavior, there are three important constructs found in the field of environmental psychology: Perceived Behavioral Control (PCB), Attitude, and Moral Norm (Bamberg & Möser, 2007). From the interviews it can be established that both the moral norm and attitude towards energy reduction within the case study organization are positive. This leaves PBC, it is possible that the interviewees do not feel that they can perform the behavior necessary to reduce energy consumption. There could be many reasons for this, but in the context of this research, the perceived lack of impact that some interviewees feel that they have on energy consumption could contribute to a lack of PBC and in turn a lowered intention towards pro-environmental behavior. It can be speculated that the lack of information barrier also influences employees' PBC since the effectiveness and impact of initiatives cannot be tracked when there is a lack of information.

7 Conclusions, Recommendations, and Future Research

This chapter will discuss the key findings of this research and will answer the research question set forth in the beginning of this thesis. The academic and practical implications will be discussed and finally the limitations of this research and suggestions for future research will be considered.

7.1 Key Findings

When looking at the results of this study a few key findings emerge which are described below.

Firstly, it has become clear through this research that organizational barriers identified in the literature do influence the adoption of energy reduction techniques in AI enabled applications. Practitioners and managers alike noted the existence of organizational barriers and their inhibiting effect on further adoption of sustainability initiatives. Furthermore, this study shows that the characteristics of techniques also influence the adoption of these techniques. Finally, this research provides evidence that the interaction between organizational barriers and the characteristics of techniques influences the adoption of energy reduction techniques in AI enabled applications. Here, the organizational barriers have a moderating effect on the influence of technique characteristics on adoption.

Secondly, it was found that the lack of information was identified as the biggest inhibitor to adoption. The lack of information even influenced the ability of practitioners and managers to assess the other barriers and characteristics. This shows the importance of the barrier and the need for increased education and measurement of sustainability goals. Furthermore, due to a lack of information on benefits and impacts, the value of sustainability improvement of products is ambiguous. This causes tensions in the organization that can hinder progress and cause a rift between management and employees.

Third, this research explored the creation of a trade-off model to support the business case for environmental sustainability. This study finds that it is currently not feasible for the case study organization to create such a trade-off model. The following points should be addressed in future research:

- The creation of a catalog of generic energy smells to enable the identification of energy inefficient behavior in source code.
- The relative impact of these energy smells has to be catalogued so that the energy optimization opportunity can be quantified.

In practice there are also inhibitors that should be addressed to allow for the creation of a trade-off model:

- Cloud providers should increase the transparency and granularity of energy related metrics for cloud users.
- Programmer's awareness on energy consumption should be prioritized to allow for more accurate effort estimation.

Finally, this research provides insights into organizational tensions arising due to the drive for sustainable products. The perceived benefits of management did not align with the perceived benefits of employees. Employees questioned the usefulness of integrating sustainability in products by stating that the impact is minimal and that there are many other things to improve before improving the AI models and training. Furthermore, there are tensions around who is responsible for integrating sustainability in products, especially software products. Management supports initiatives to improve the sustainability of products but is not yet actively pushing for them. Employees are

open to improving sustainability of products but are also not actively pursuing it. This leads to stagnation of implementation of initiatives and if left unaddressed can lead to frustration between management and employees.

7.2 Conclusion and Recommendations

This research endeavored to answer the question: *How can the adoption viability of energy reduction techniques in AI development be assessed?* In this thesis barriers to adoption of sustainable practices are used to create a framework which enables organizations to assess energy reduction techniques in AI development. This framework acts as an enabler for organizations to measure and assess their barriers and can break the stagnation that occurs in sustainable practice adoption due to a lack of explicit external drivers. From this research it can be concluded that organizational barriers and technique characteristics can indeed be utilized together to assess the adoption viability of energy reduction techniques in AI development.

Moreover, this research concludes that the most influential factors impacting adoption and implementation of environmental sustainability initiatives in AI development are the availability of information and the disruption caused by the initiative. These findings inform assessment and can allow for the remediation of these factors to enable implementation in a way most suited to the organization at hand.

7.2.1 Recommendations

From the findings and conclusions of this research, recommendations to the case study organization can be compiled. Accordingly, this study issues the following recommendations:

- The case study organization should invest in improving the information availability of sustainable development of AI systems within the engineering teams. This can be achieved by allocating even minimal time for research activities in this direction and facilitating sharing of finding among the teams.
- The case study organization should keep working to quantify the impact of AI systems and their development and should endeavor to create internal KPI's to enable tracking initiatives and improvements.
- The case study organization should be wary of attitudes towards the effectiveness and impact of sustainability initiatives within the product teams. Disparity between the perceived benefits of sustainability within product teams and management can cause serious resistance if left unaddressed.
- Management of the case study organization should explicitly state who is responsible for taking initiative in improving the sustainability of software products. Management should also be explicit about the nature of initiatives in sustainable software development and provide clarity on whether sustainability initiatives are in principle incidental or structural activities.

7.3 Academic Implications

This research contributes to academic knowledge by studying the challenges and determinants for sustainable practice adoption in a technology and innovation driven field. It shows that traditional barriers to sustainable practice adoption can be utilized in this field and adds knowledge by creating four constructs which also influence adoption. This lays the foundation for future research aiming to explore adoption of environmental sustainability initiatives in fields where quantitative data is scarce and quick decision making is crucial.

This research also represents a first step in promoting the transfer of knowledge between academia and industry by providing a framework which is easily accessible and can be used to initiate dialogue and discussion around new sustainable techniques found in research.

7.4 Practical Implications

The practical implications of this thesis are derived from the value that the framework provides to organizations using it. The framework allows for decision support where limited quantitative information is present, which is often the case in fields utilizing novel technologies and techniques. Using this framework, organizations can also identify problem areas based on the evaluation of barriers and implement remedies to mitigate these barriers and improve the organizational conditions for sustainable practice adoption.

7.5 Limitations

Despite efforts to maximize validity and quality, this research encountered some limitations that will be discussed in this section.

Firstly, this research takes place within a single case study organization which limits the amount and diversity of data that could be collected which can influence the validity of the data. However, due to the qualitative nature of this research and the academic grounding of the concepts, the principles discovered in this study hold merit and can serve as the basis for future research and add value to academic research and industry practice.

Secondly, during this study an interesting phenomenon was noted. Namely, the lack of information barrier was also stated by interviewees to limit their ability to assess other barriers. This means that a change in one barrier might influence the evaluation of other barriers and change the eventual assessment. This effect is not explicitly researched in this study. This relationship should be further explored in the future to increase the robustness and generalizability of the assessment framework. However, due to the validation interviews, the outcomes of assessments for the case study organization could still be verified.

Thirdly, the technique characteristics used in the assessment framework are supported by a smaller contingent of existing research than the organizational barriers are. The characteristics relied more heavily on implied relationships from the literature and the statements and interpretations of the interviewees within the case study organization. Therefore, the applicability of the identified technique characteristics for the purposes of adoption viability assessment should be researched further to confirm the selection.

7.6 Future Research

As briefly discussed in the limitations and key findings of this chapter, there are some interesting future research directions that could be beneficial to the fields of software energy reduction and sustainability management within organizations.

Firstly, future research could focus on improving the generalizability of the assessment framework. Currently the framework has only been validated based on AI enabled applications within a single organization. However, if further developed, the framework could be a valuable sustainability adoption assessment tool for a range of fields and applications. To achieve this, the diversity of organizations and fields should be expanded to enable the creation of a more refined version of the framework and its underlying principles which could be applicable to future technologies and fields as well. Furthermore, a standard methodology for determining the weight and extent of the barriers and characteristics should be developed to improve the reproducibility and comparison of results between different organizations.

Secondly, future research could identify energy smells as discussed in chapter 6.2 and create tools and plugins to be able to identify these energy smells within source code. In this study the need for the functionality of identification of energy inefficient behavior through static code analysis was reiterated multiple times by developers. Research which creates such tools would thus be a valuable addition to both academia and practice.

Thirdly, the creation of a business case for environmental sustainability in software development should be reexamined when the literature has caught up enough to facilitate this. The creation of a business case using a trade-off model would provide important insights into the perceived economic value of sustainability by organizations. This would also allow for the creation of a more quantitative assessment framework for sustainability initiatives. Identification of relevant KPI's for energy consumption and efficiency in AI enabled applications should also be further researched to improve tracking and evaluating energy reduction techniques.

Finally, the framework created in this study is practical in nature and integrates only a small part of the human component through the organizational barriers. Future research should dive further into the human component through the lens of organizational tensions arising from the adoption of sustainability within product development. Here, longitudinal case studies could be conducted in which researchers follow the process of sustainability integration and identify tensions and remedies. It would also be interesting to see which approach to dealing with organizational tensions in sustainability would be most beneficial for organizations.

Bibliography

1. Aken, V. J. J., & Joan, E. (2004). Management Research Based on the Paradigm of the Design Sciences: The Quest for Field-Tested and Grounded Technological Rules. Social Science Research Network. https://autopapers.ssrn.com/sol3/papers.cfm?abstract_id=513679
2. Alraja, M. N., Imran, R., Khashab, B. M., & Shah, M. (2022). Technological Innovation, Sustainable Green Practices and SMEs Sustainable Performance in Times of Crisis (COVID-19 pandemic). Information Systems Frontiers, 24(4), 1081–1105. https://doi.org/10.1007/s10796-022-10250-z
3. Arvanitis, S., & Ley, M. (2012). Factors Determining the Adoption of Energy-Saving Technologies in Swiss Firms: An Analysis Based on Micro Data. Environmental & Resource Economics, 54(3), 389–417. https://doi.org/10.1007/s10640-012-9599-6
4. Baker, J. (2011). The Technology–Organization–Environment Framework. In Integrated series on information systems/Integrated series in information systems (pp. 231–245). https://doi.org/10.1007/978-1-4419-6108-2_12
5. Balanza-Martinez, J., Lago, P., & Verdecchia, R. (2024). Tactics for Software Energy Efficiency: A Review. In Progress in IS (Print) (pp. 115–140). https://doi.org/10.1007/978-3-031-46902-2_7
6. Bamberg, S., & Möser, G. (2007). Twenty years after Hines, Hungerford, and Tomera: A new meta-analysis of psycho-social determinants of pro-environmental behaviour. Journal Of Environmental Psychology, 27(1), 14–25. https://doi.org/10.1016/j.jenvp.2006.12.002
7. Bansal, P. (2005). Evolving sustainably: a longitudinal study of corporate sustainable development. Strategic Management Journal, 26(3), 197–218. https://doi.org/10.1002/smj.441
8. Collins, E., Roper, J., & Lawrence, S. (2010). Sustainability practices: trends in New Zealand businesses. Business Strategy And The Environment, 19(8), 479–494. https://doi.org/10.1002/bse.653
9. Cooremans, C. (2007). Strategic fit of energy Efficiency (Strategic and cultural dimensions of energy-efficiency investments). Proceedings Of The European Council For An Energy-Efficient Economy (ECEEE) Summer Study (2007). https://archive-ouverte.unige.ch/unige:97652
10. Cooremans, C. (2012). Investment in energy efficiency: do the characteristics of investments matter? Energy Efficiency, 5(4), 497–518. https://doi.org/10.1007/s12053-012-9154-x
11. Couto, M., Saraiva, J., & Fernandes, J. P. (2020). Energy Refactorings for Android in the Large and in the Wild. 2020 IEEE 27th Int. Conference On Software Analysis, Evolution And Reengineering (SANER). https://doi.org/10.1109/saner48275.2020.9054858
12. De Chavannes, L. H. P., Kongsbak, M. G. K., Rantzau, T., & Derczynski, L. (2021). Hyperparameter Power Impact in Transformer Language Model Training. Proceedings Of The Second Workshop On Simple And Efficient Natural Language Processing. https://doi.org/10.18653/v1/2021.sustainlp-1.12
13. De Groot, H. L., Verhoef, E. T., & Nijkamp, P. (2001). Energy saving by firms: decision-making, barriers and policies. Energy Economics, 23(6), 717–740. https://doi.org/10.1016/s0140-9883(01)00083-4
14. De Oliveira, U. R., Menezes, R. P., & Fernandes, V. A. (2023). A Systematic Literature Review on Corporate Sustainability: Contributions, barriers, innovations, and Future Possibilities. Environment, Development and Sustainability. https://doi.org/10.1007/s10668-023-02933-7
15. Deely, J., Hynes, S., Barquín, J., Burgess, D., Finney, G., Silió, A., Álvarez-Martínez, J. M., Bailly, D., & Ballé-Béganton, J. (2020). Barrier identification framework for the implementation of blue and green infrastructures. Land Use Policy, 99, 105108. https://doi.org/10.1016/j.landusepol.2020.105108

16. Elkington, J. (1994). Towards the Sustainable Corporation: Win-Win-Win Business Strategies for Sustainable Development. California Management Review, 36(2), 90–100. https://doi.org/10.2307/41165746
17. Elkington, J. (1997). Cannibals with forks: The Triple Bottom Line of 21st Century Business.
18. Emmerloot, N. v. (2020). When Information Becomes Crucial: A Case Study about the Barriers of Information Sharing during the COVID-19 Crisis. Tilburg: Tilburg University.
19. Ervin, D. E., Wu, J., Khanna, M., Jones, C., & Wirkkala, T. M. (2012). Motivations and Barriers to Corporate Environmental Management. Business Strategy And The Environment, 22(6), 390–409. https://doi.org/10.1002/bse.1752
20. Fleiter, T., Schleich, J., & Ravivanpong, P. (2012). Adoption of energy-efficiency measures in SMEs—An empirical analysis based on energy audit data from Germany. Energy Policy, 51, 863–875. https://doi.org/10.1016/j.enpol.2012.09.041
21. Freitag, C., Berners-Lee, M., Widdicks, K., Knowles, B., Blair, G. S., & Friday, A. (2021). The Real Climate and Transformative Impact of ICT: A critique of estimates, Trends, and regulations. Patterns, 2(9), 100340. https://doi.org/10.1016/j.patter.2021.100340
22. Friedman, A. L., & Miles, S. (2002). SMEs and the environment: evaluating dissemination routes and handholding levels. Business Strategy And The Environment, 11(5), 324–341. https://doi.org/10.1002/bse.335
23. Gadenne, D. L., Kennedy, J., & McKeiver, C. (2008). An Empirical Study of Environmental Awareness and Practices in SMEs. Journal Of Business Ethics, 84(1), 45–63. https://doi.org/10.1007/s10551-008-9672-9
24. Goaër, O. L., & Hertout, J. (2022). ecoCode: a SonarQube Plugin to Remove Energy Smells from Android Projects. ASE '22: Proceedings Of The 37th IEEE/ACM International Conference On Automated Software Engineering. https://doi.org/10.1145/3551349.3559518
25. Gohoungodji, P., N'Dri, A. B., Latulippe, J., & Matos, A. L. B. (2020). What is stopping the automotive industry from going green? A systematic review of barriers to green innovation in the automotive industry. Journal Of Cleaner Production, 277, 123524. https://doi.org/10.1016/j.jclepro.2020.123524
26. Gómez-Bezares, F., Przychodzeń, W., & Przychodzeń, J. (2019). Corporate Sustainability and CEO–Employee Pay Gap—Buster or Booster? Sustainability, 11(21), 6023. https://doi.org/10.3390/su11216023
27. Gottschalk, M., Josefiok, M., Jelschen, J., & Winter, A. (2012). Removing energy code smells with reengineering services. GI-Jahrestagung, 441–455. http://www.se.uni-oldenburg.de/documents/gottschalk+2012.pdf
28. Guo, Y. (2018). A Survey on Methods and Theories of Quantized Neural Networks. arXiv (Cornell University). http://export.arxiv.org/pdf/1808.04752
29. Hevner, A. R. (2007). A Three Cycle View of Design Science Research. Scandinavian Journal Of Information Systems, 19(2), 4. http://community.mis.temple.edu/seminars/files/2009/10/Hevner-SJIS.pdf
30. Hevner, March, Park, H., & Ram. (2004). Design science in Information Systems Research. Management Information Systems Quarterly, 28(1), 75. https://doi.org/10.2307/25148625
31. Hoefler, T., Alistarh, D., Ben-Nun, T., Dryden, N., & Peşte, A. (2021). Sparsity in Deep Learning: Pruning and growth for efficient inference and training in neural networks. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2102.00554
32. Hubara, I., Courbariaux, M., Soudry, D., El-Yaniv, R., & Bengio, Y. (2016). Quantized Neural Networks: Training Neural Networks with Low Precision Weights and Activations. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1609.07061

33. Hwang, B. N., Huang, C. Y., & Wu, C. H. (2016). A TOE Approach to Establish a Green Supply Chain Adoption Decision Model in the Semiconductor Industry. Sustainability, 8(2), 168. https://doi.org/10.3390/su8020168
34. Journeault, M., Perron, A., & Vallières, L. (2021). The collaborative roles of stakeholders in supporting the adoption of sustainability in SMEs. Journal Of Environmental Management, 287, 112349. https://doi.org/10.1016/j.jenvman.2021.112349
35. Kang, M., Lee, Y., & Moonju, P. (2020). Energy Efficiency of Machine Learning in Embedded Systems Using Neuromorphic Hardware. Electronics (Basel), 9(7), 1069. https://doi.org/10.3390/electronics9071069
36. Karita, L., Mourão, B. C., Martins, L., Soares, L. R., & Machado, I. (2021). Software industry awareness on sustainable software engineering: a Brazilian perspective. Journal Of Software Engineering Research And Development, 9. https://doi.org/10.5753/jserd.2021.742
37. Krzywaniak, A., Czarnul, P., & Proficz, J. (2022). GPU Power Capping for Energy-Performance Trade-Offs in Training of Deep Convolutional Neural Networks for Image Recognition. In Lecture notes in computer science (pp. 667–681). https://doi.org/10.1007/978-3-031-08751-6_48
38. Lawrence, S., Collins, E., Pavlovich, K., & Arunachalam, M. (2006). Sustainability practices of SMEs: the case of NZ. Business Strategy And The Environment, 15(4), 242–257. https://doi.org/10.1002/bse.533
39. Lenox, M., & Ehrenfeld, J. R. (1997). Organizing for effective environmental design. Business Strategy And The Environment, 6(4), 187–196. https://doi.org/10.1002/(sici)1099-0836(199709)6:4
40. Lewis, M. W. (2000). Exploring Paradox: Toward a More Comprehensive Guide. ˜The œAcademy Of Management Review, 25(4), 760–776. https://doi.org/10.5465/amr.2000.3707712
41. Lueg, R., Pedersen, M. M., & Clemmensen, S. N. (2013). The Role of Corporate Sustainability in a Low-Cost Business Model – A Case Study in the Scandinavian Fashion Industry. Business Strategy And The Environment, 24(5), 344–359. https://doi.org/10.1002/bse.1825
42. Margolis, J. D., & Walsh, J. P. (2003). Misery Loves Companies: Rethinking Social Initiatives by Business. Administrative Science Quarterly, 48(2), 268–305. https://doi.org/10.2307/3556659
43. Morales, R. R., Saborido, R., Khomh, F., Chicano, F., & Antoniol, G. (2018). EARMO: An Energy-Aware Refactoring Approach for Mobile Apps. IEEE Transactions On Software Engineering, 44(12), 1176–1206. https://doi.org/10.1109/tse.2017.2757486
44. Murillo, D., & Lozano, J. M. (2006). SMEs and CSR: An Approach to CSR in their Own Words. Journal Of Business Ethics, 67(3), 227–240. https://doi.org/10.1007/s10551-006-9181-7
45. Myers, M., & Newman, M. (2007). The qualitative interview in IS research: Examining the craft. Information And Organization, 17(1), 2–26. https://doi.org/10.1016/j.infoandorg.2006.11.001
46. Ournani, Z., Rouvoy, R., Rust, P., & Penhoat, J. (2020). On Reducing the Energy Consumption of Software. Proceedings Of The 14th ACM/IEEE International Symposium On Empirical Software Engineering And Measurement (ESEM). https://doi.org/10.1145/3382494.3410678
47. Pang, C., Hindle, A., Adams, B., & Hassan, A. E. (2016). What Do Programmers Know about Software Energy Consumption? IEEE Software, 33(3), 83–89. https://doi.org/10.1109/ms.2015.83
48. Parker, C. M., Redmond, J., & Simpson, M. (2009). A Review of Interventions to Encourage SMEs to Make Environmental Improvements. Environment And Planning. C, Government & Policy/Environment And Planning. C, Government And Policy, 27(2), 279–301. https://doi.org/10.1068/c0859b
49. Pereira, R., Carção, T., Couto, M., Cunha, J., Fernandes, J. P., & Saraiva, J. (2020). SPELLing out energy leaks: Aiding developers locate energy inefficient code. ˜The œJournal Of Systems And Software, 161, 110463. https://doi.org/10.1016/j.jss.2019.110463

50. Pinto, G., & Castor, F. (2017). Energy efficiency. Communications Of The ACM, 60(12), 68–75. https://doi.org/10.1145/3154384
51. Saqib, Z. A., & Zhang, Q. (2021). Impact of sustainable practices on sustainable performance: the moderating role of supply chain visibility. Journal Of Manufacturing Technology Management, 32(7), 1421–1443. https://doi.org/10.1108/jmtm-10-2020-0403
52. Scherer, A. G., Palazzo, G., & Seidl, D. (2013). Managing Legitimacy in Complex and Heterogeneous Environments: Sustainable Development in a Globalized World. Journal Of Management Studies, 50(2), 259–284. https://doi.org/10.1111/joms.12014
53. Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. Communications Of The ACM, 63(12), 54–63. https://doi.org/10.1145/3381831b
54. Sen, S., & Cowley, J. C. P. (2012). The Relevance of Stakeholder Theory and Social Capital Theory in the Context of CSR in SMEs: An Australian Perspective. Journal Of Business Ethics, 118(2), 413–427. https://doi.org/10.1007/s10551-012-1598-6
55. Shahid, A., & Mushtaq, M. (2020). A Survey Comparing Specialized Hardware And Evolution In TPUs For Neural Networks. 2020 IEEE 23rd International Multitopic Conference (INMIC). https://doi.org/10.1109/inmic50486.2020.9318136
56. Singh, K., & Misra, M. (2021). Linking Corporate Social Responsibility (CSR) and Organizational Performance: the moderating effect of corporate reputation. European Research On Management And Business Economics, 27(1), 100139. https://doi.org/10.1016/j.iedeen.2020.100139
57. Singh, M., & Sahu, G. P. (2020). Towards adoption of Green IS: A literature review using classification methodology. International Journal Of Information Management, 54, 102147. https://doi.org/10.1016/j.ijinfomgt.2020.102147
58. Smith, W. K., & Lewis, M. W. (2011). TOWARD a THEORY OF PARADOX: a DYNAMIC EQUILIBRIUM MODEL OF ORGANIZING. ˜The œAcademy Of Management Review, 36(2), 381–403. https://doi.org/10.5465/amr.2011.59330958
59. Song, L., Deng, Y., Zhu, Z., Hua, H., & Tao, Z. (2021). A Comprehensive Review on Radiomics and Deep Learning for Nasopharyngeal Carcinoma Imaging. Diagnostics (Basel), 11(9), 1523. https://doi.org/10.3390/diagnostics11091523
60. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. The 57th Annual Meeting Of The Association For Computational Linguistics (ACL). Florence, Italy. July 2019. https://doi.org/10.18653/v1/p19-1355
61. Tornatzky, L. G., Fleischer, M., & Chakrabarti, A. K. (1990). processes of technological innovation. https://agris.fao.org/agris-search/search.do?recordID=US201300694725
62. Turner, R., Eriksson, D., McCourt, M., Kiili, J., Laaksonen, E., Xu, Z., & Guyon, I. (2021). Bayesian Optimization is Superior to Random Search for Machine Learning Hyperparameter Tuning: Analysis of the Black-Box Optimization Challenge 2020. arXiv (Cornell University). https://arxiv.org/pdf/2104.10201.pdf
63. Van Der Byl, C. A., & Slawinski, N. (2015). Embracing tensions in corporate sustainability. Organization & Environment, 28(1), 54–79. https://doi.org/10.1177/1086026615575047
64. Verdecchia, R., Saez, R. A., Procaccianti, G., & Lago, P. (2018). Empirical Evaluation of the Energy Impact of Refactoring Code Smells. EPiC Series in Computing. https://doi.org/10.29007/dz83
65. Verdecchia, R., Sallou, J., & Cruz, L. J. (2023). A systematic review of Green AI. WIREs Data Mining And Knowledge Discovery, 13(4). https://doi.org/10.1002/widm.1507
66. Vos, S., Lago, P., Verdecchia, R., & Heitlager, I. (2022). Architectural Tactics to Optimize Software for Energy Efficiency in the Public Cloud. 2022 International Conference On ICT For Sustainability (ICT4S). https://doi.org/10.1109/ict4s55073.2022.00019

67. Weng, M., & Lin, C. (2011). Determinants of green innovation adoption for small and medium-size enterprises (SMES). African Journal Of Business Management, 5(22), 9154–9163. https://doi.org/10.5897/ajbm.9000199
68. Wu, C., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., Chang, G., Behram, F. A., Huang, J., Bai, C., Gschwind, M., Gupta, A., Ott, M., Мельников, A. C., Candido, S., Brooks, D. J., Chauhan, G. S., Lee, B., Lee, H. S., . . . Hazelwood, K. (2021). Sustainable AI: Environmental Implications, Challenges and Opportunities. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2111.00364
69. Yang, T. J., Chen, Y. H., & Sze, V. (2017). Designing energy-efficient convolutional neural networks using energy-aware pruning. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5687-5695).
70. You, J., Chung, J., & Chowdhury, M. (2022). Zeus: Understanding and Optimizing GPU Energy Consumption of DNN Training. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2208.06102
71. Yu, T., & Zhu, H. (2020). Hyper-Parameter Optimization: A Review of Algorithms and Applications. arXiv (Cornell University). https://arxiv.org/pdf/2003.05689
72. Zhu, K., & Kraemer, K. L. (2005). Post-Adoption Variations in Usage and Value of E-Business by Organizations: Cross-Country Evidence from the Retail Industry. Information Systems Research, 16(1), 61–84. https://doi.org/10.1287/isre.1050.0045

Appendix A: Interview Protocol

Background information interviewee	What is your name?
	What is your role?
	How long have you been working in that role?
	Could you provide a brief overview of your primary tasks and responsibilities?
Familiarity with environmental sustainability	Are you familiar with the concept of environmental sustainability?
	What do you think environmental sustainability entails in your line of work?
	Do you think environmental sustainability is an important aspect of creating software?
Awareness	Have you personally thought about the energy consumption of the products and their development? What are your thoughts on it?
	How is the energy consumption of the products and development discussed within the team if at all?
	Do you have an idea of the amount of energy consumed by the products and development?
Impact	How big do you think the impact of the products and development you work on is on the environment?
	What do you think is there to gain from reducing the energy consumption of the products and their development?
	What feedback do you get from the customer about energy consumption of the products?
Willingness	Do you think it would be worthwhile to reduce energy consumption of the products and their development?
	Would you spend time on reducing energy consumption of the products and their development if you were allowed to?
	What do you think should be done to reduce energy consumption of the products if anything?
	What would be the best area to start improving if you would want to reduce energy consumption?
Remedies	What are the biggest hurdles for you to implement techniques that could reduce energy consumption of the products and their development?
	What would encourage you to reduce the energy consumption of the products and development?
Technique Characteristics	How do you assess possible (non-energy related) improvements now?
	What methods or techniques are you aware of that could reduce the energy consumption of the products you work on or their development? • Are you aware of any tools that support these methods or techniques? • Would you know where to look for energy reducing methods or techniques?
	What would make an energy reduction technique attractive for you to implement?
Closing	Is there anything that you would like to add, or do you have any questions for me?

Appendix B: Interviews

Interview CV1:

What is your name?

Sanjeet

What is your role?

Vision Engineer

How long have you been working in that role?

I have been working as a Vision Engineer for around a year, but my current tasks I have been doing for around 6-8 months now.

Could you provide a brief overview of your primary tasks and responsibilities?

In the team we have three sections: research, product, and solutions. I am working in the product team; the idea of the product team is to build tools and systems that are a platform for researchers and the solutions team to research or fine tune the internal products of the team. We try to provide standards to the research and solution teams to do their work. So, in MLops for example there are so many things to choose from, and we provide the standard for that. And we set it up and support them using it in the right way. We also create pipelines, for example to train deep learning models in a way that work with the company's storage systems and not everyone has to reinvent the wheel to train a new model or make optimizations. That is the biggest part of what I do.

Are you familiar with the concept of environmental sustainability?

I mean a bit. I don't think in a detailed way, I know there are some expectations from the EU on carbon emissions of companies and stuff like that but that is about all I know.

What do you think environmental sustainability entails in your line of work?

I think it comes down to making more efficient systems. You don't want to make systems which do the same things multiple times. The relevant part is how you can get things done with the least amount of resources and I think that correlates with sustainability.

Have you personally thought about the energy consumption of the products and their development? What are your thoughts on it?

Not particularly no. in general the solutions we provide are based on customer requirements so if a customer has a certain requirement, we adhere to them.

Have you had any requirements regarding energy consumption from customers before?

Not that I know of, but at the same time I haven't been dealing with customers myself so I might not be the best source for this. But it's for us always about performance measures.

How is the energy consumption of the products and development discussed within the team if at all?

If there is an interesting paper around it and somebody finds it, they will present it to the team. So we sometimes know what is going on in the research but it's not structural.

How big do you think the impact of the products and development you work on is on the environment?

I think it is tricky to answer because the things we make consume energy, but they also make customer processes more efficient. Also, these systems are expected to run for quite some time and so the development is not that large of a part of it.

So, do you think it's worthwhile then to reduce the energy consumption of inference?

I think for us the important thing is efficiency, if we can make it faster it is the best for us. And if a model is smaller, it is also faster and uses less GPU. So indirectly we do reduce energy consumption, but what is missing is how these optimizations are translating to reduction energy consumption.

So would you say that there is a lack of information on how to measure energy consumption?

I would say a lack of motive. If you are asked to do it, of course you are going to do it but otherwise not. How I see it, sustainability is top-down, basically is that the company will first encourage the easy things and the development at Sight is not a low hanging fruit. Its also more IP and a value creating activity. If you want to optimize energy consumption than the team will have to put work in, and that means that we have to put other work aside.

Would you spend time on reducing energy consumption of the products and their development if you were allowed to?

Yeah for sure, I think that is also a big reason why I am actively helping you. For me this is also an opportunity to get a better understanding of these methods. And if something good comes out of it the team can implement it. I think it's also an initiation factor right.

What would make an energy reduction technique attractive to implement?

I think incentive is a big part of it, if the customer comes tomorrow with a requirement for energy efficiency than we would do it. I think its not different perse than any other task, we would just do it.

What methods or techniques are you aware of that could reduce the energy consumption of the products you work on or their development?

No, not perse. For me, I think the more you make things optimal the more you reduce energy consumption, but we don't measure for energy consumption, so I don't know.

Where would you start with energy reducing methods or techniques?

I would start with the training side, I think if you just start with some initial techniques that could start to reduce the energy consumption that is something that could help a lot, because that is something that we can do internally. If there could be something that could tell us like okay this function consumes more energy, so use this one. That is something that you could look into as well.

Interview CV2

What is your name and what is your role?

Wouter, en ik ben een Computer vision engineer

How is the energy consumption of the products and development discussed within the team if at all?

Er wordt binnen het team niet echt gesproken over energieverbruik, dat blijft echt bij de performance in tegenstelling tot de groene initiatieven die hier bij PV rondvliegen.

We zijn wel bewust van de hardware claim die we nu maken dat we gewoon 8 A30 GPUs nodig hebben, en die gaan naar {klant} en die willen juist vergroenen, dus dan heb je wel zoiets van oh misschien is dit niet handig.

We doen in principe niet direct heel veel met groen moet ik zeggen. We zijn er niet erg bewust mee bezig, maar er zitten wel wat impliciete dingen die meeliften die hand in hand gaan. Hoe sneller je klaar bent met trainen hoe beter, hoe efficiënter je algoritme hoe minder je verbruikt en hoe minder dure GPUs je nodig hebt. Dus er zit wel een flink lineair verband tussen de rekenkracht en hoe veel je verbruikt. En dat zit dus niet alleen in de inference, dus als je aan het runnen bent. Ook al is dat wel een van de belangrijkere, het moet gewoon snel draaien bij de klant. Het heeft ook te maken met klanten die niet al te dure dingen willen kopen. En ook voor training, hoe eerder je klaar bent en hoe sneller je kan experimenteren hoe beter je daarmee af bent.

Ik denk dat binnen het team er niet echt kennis over is hoe je een model groen maakt, misschien indirect zoals ik al zei. Het is ook minder tastbaar, het is net als een lamp uitzetten, realiseer je je wel dat het zonde is dat er onnodig energie wordt verbruikt. Het is heel makkelijk om achter je bureau op de train knop te drukken en dan ga je koffie halen, maar er staat dan wel een half oerbos af te branden om die server aan te doen bij wijze van spreken.

Do you have an idea of the amount of energy consumed by the products and development?

We weten het sowieso niet, het moet wel goed kwantificeerbaar zijn en goed reproduceerbaar zijn. Je moet het wel kunnen zien en een educated guess kunnen maken om te zien of je de goede kant op gaat. En als de klant dat wil dan moeten we dat ook doen. Of als we dat zelf willen als dat nou eenmaal de propositie is geweest

What would be the best area to start improving if you would want to reduce energy consumption?

Ik denk bij inference, omdat uiteindelijk wordt dat heel vaak wordt uitgevoerd. Een train proces duurt een week misschien en inference gaat natuurlijk continu door.

How do you assess possible (non-energy related) improvements now?

Wij bekijken het vooral vanuit het snelheids oogpunt. We optimaliseren eigenlijk altijd op accuracy en dan de volgende belangrijke is latency. Je kent misschien wel dat speelgoed hamertje tik dat je drie paaltjes hebt waar je op kan slaan en als je er eentje naar beneden slaat dan komt er een andere weer omhoog. En zo bekijken we optimalisaties ook. Dus de valkuil van machine learning is om te zeggen van oh doe er maar een GPU bij. Zo van, we hebben een model gemaakt met een goede read rate en een goede error rate en dan kijken we wel gewoon hoeveel hardware we nodig hebben om die te kunnen runnen. Zonder dat je per se het model optimaliseert.

Voor klanten is het ook zo dat ze een bepaalde threshold hebben voor de latency, en als het dan iets sneller is of iets minder snel maakt niet zo veel uit, als het maar die threshold haalt.

Qua process zijn we best flexibel om het onderste uit de kan te halen, dus proces technisch maakt het niet zo veel uit want het gaat erom dat het project zelf goeie performance haalt.

You are working on integrating CI/CD principles and standardizing tooling. Are there any limitations there regarding tooling?

Daar gaan we inderdaad veel meer naartoe, ook met de modulariteit, de traditionele aanpak was we bouwen alles zelf en dan kun je alles optimaliseren. Maar je moet het dan ook zelf onderhouden en zelf fixen. En je moet ook zorgen dat het herbruikbaar is en als je iets gaat optimaliseren moet je ook zorgen dat alle andere producten ook blijven werken. Als je het helemaal zelf maakt dan kun je optimaliseren tot op het bot, maar dat is toch een hele andere tak van sport. Je wordt dan toch een soort schaap met 5 poten want je moet hardware optimalisatie aankunnen en ML en systeem optimalisaties en op een gegeven moment houdt dat op. En dan kun je met de standaard optimalisatie uit pakketten zoals pytorch en openvino al een heleboel die we zelf misschien niet hadden gekund. En dan hadden we het model nog wel sneller kunnen maken maar dan wel ten koste van de modulariteit.

Maar dat zeker wel een obstakel, als we zelf custom optimalisaties moeten maken die niet in de standaard tooling zitten dan gaat dat ook ten koste van de modulariteit. En ik weet zeker dat als er wordt gevraagd van maak dit model groen, dat er dan naar wordt gekeken maar als het dan read rate gaat kosten dan wordt het lastig. In principe meten we alles aan de performance af, ook niet eens aan de kosten. Misschien wordt het vanuit de business kant wel zo ingeschoten maar bij ons komt het altijd terug op de performance.

What methods or techniques are you aware of that could reduce the energy consumption of the products you work on or their development?

Wat eventueel een interessante gedachte zou kunnen zijn is om je code modulairder op te zetten dat je alleen de dingen de dingen test die je daadwerkelijk veranderd. Want je hebt een aantal ML componenten maar ook het opzoeken van dingen in de database en de reasoning zoals we dat noemen. Zoals het opzoeken van postcode en of die wel matched met de straat bijvoorbeeld. Dat is allemaal rechttoe rechtaan programmeer werk. Dus als je daar iets aan veranderd en dan alleen het moduletje test en niet alles wat ervoor zit, want het is toch deterministisch genoeg om hetzelfde antwoord te krijgen. Dat zou ook al wel kunnen bijdragen aan het verminderen van energieverbruik. Daar doen we nu nog niet zo veel mee, maar daar zijn we nu wel meer die kant op aan het gaan.

What would make an energy reduction technique attractive for you to implement?

Waar je natuurlijk niet op zit te wachten is nog een dimensie om op te letten tijdens je totaal optimalisatie. Dan moet je zo veel balletjes in de lucht houden dat is op een gegeven moment niet meer te doen. Dus trivialiteit en zo min mogelijk impact zou wel ideaal zijn. En dan trivialiteit bedoel ik in de zin dat het je werk niet opeens stukken meer complex maakt. Ik denk dat het over het algemeen goed is als je weet dat je er iets goed mee doet, zo van dit kost me wel wat, misschien niet eens zo veel maar ik doe er wel wat goeds mee. Het moet natuurlijk wel gewoon aan kwaliteitseisen voldoen, dat het stabiel is en dat je weet wat je kan verwachten. En het moet wel echt zin hebben.

Would you say that there is a scarcity of resources like time and computing power right now?

Ik weet het niet zeker, maar ik denk dat we de GPU tijd bijvoorbeeld wel efficiënter zouden kunnen inrichten, we hebben de resources maar ik denk niet dat we ze optimaal benutten.

Interview PM

We doen al veel met AI maar we doen nog niet alles met AI, we missen daar een stukje infrastructuur en een stukje kennis en misschien ook wanneer dat verantwoord is. Maar ik denk als de functionaliteit er is en het doet wat het hoort te doen, dan denk ik niet dat daar belemmeringen zijn.

Imagine that training of AI would take a little longer, but would reduce energy consumption. Do you think there would be resistance in the teams?

Nee dat denk ik niet, kijk als je training 20 keer langer duurt dan is het natuurlijk een ander verhaal. Maar als je training een nacht duurt dan is een uurtje extra echt geen ramp. En zeker omdat het ook een goed doel dient. Het belangrijkste is dat ze in controle zijn, dat ze kunnen plannen en dat ze weten wanneer het klaar is. Dus het gaat meer om voorspelbaarheid en je moet natuurlijk wel door kunnen met je proces.

Are people enthousiastic about this? Do you think they are prepared to work on reducing energy consumption?

Het feit dat je heel veel computerkracht gebruikt om iets te maken waar wij heel enthousiast van worden en het besef van hey wacht eens even, dit kost heel veel energie, dit kost heel veel geld, dit kost heel veel co2. Als er mogelijkheden zijn om dat minder ten koste te laten gaan van onze omgeving dan denk ik serieus dat niemand daar problemen mee heeft in het team. Zolang het werkbaar blijft natuurlijk, maar daar maak ik me niet zo zorgen over dat dat kan.

Is sustainability and energy consumption discussed within the team?

Nee, er wordt nu niet aangestippeld van hey kunnen we dit anders kunnen doen om het groener te doen maar dat is ook een deel kennis delen denk ik. Er is ook gewoon weinig kennis van wat mensen groener kunnen doen op dit moment. Dus het is ook een kwestie van die informatie en die mogelijkheden beschikbaar stellen. En ik denk dat de mensen op de werkvloer zeggen, ja geef me maar een groenere oplossing dan gebruik ik hem graag. Maar zij kunnen dat niet regelen, ik denk dat dat ook een beetje bij de managementlaag ligt om dat te faciliteren.

What if the underlying source code were to change to facilitate sustainability, what barriers do you see there?

Nou het belangrijkste is dat het niet ten koste mag gaan van de performance van je systeem. En als je de kwaliteit niet meer kan leveren dan heb je een probleem met je klant. En we kijken ook naar manieren om dingen leaner en goedkoper te maken. Want klanten die hebben ook liever een kleiner systeem. Op dit moment krijgen we nog weinig druk van klanten op het groene vlak, maar dat gaat ook komen en we weten donders goed dat veel klanten dat hoog in het vaandel hebben staan. Maar op dit moment gaat het vaak over extra kosten die eraan vastzitten.

Ja, en klanten hebben nu gewoon servers staan zonder GPUs en wij zeggen dan, we kunnen je iets beters aanbieden met een GPU. Dan vallen ze niet over het energieverbruik ervan, maar over de kosten die dat meebrengt. Mensen staan absoluut open om te kijken of er andere mogelijkheden zijn om modellen kleiner en groener te maken, als ze daarmee maar ook de oplossing kunnen bieden die nodig is voor de klant. Dus daar ligt wel een vervelende afweging, want een kleiner model is vaak ook wat minder krachtig.

Imagine that you hit all required performance levels and quality concerns, but there is extra time needed to invest in optimizing the model for energy efficiency. Do you think that would be worth it?

Vanuit de engineers denk ik dat het niet uitmaakt, omdat het allebei voordelen heeft. Vanuit verkoop is het hoe het gebracht wordt. Ik denk dat de driver daar veel meer op het operationele vlak zit, want die 200 uur gaan ten koste ofwel van de bedrijfswinst of die moeten geïnvesteerd worden in een soort roadmap. We zijn op dit moment bezig met de transitie maken naar allerlei AI-systemen, en het perfectioneren van die modellen dat is stap 2. Dus ik denk dat het management daar ook naar moet kijken en genoeg waarde aan moet hechten. Want het kan natuurlijk dat het puur bedrijfstechnisch gezien duurder is om die vergroening te maken, maar dan moet het management zeggen van, dit is een feature die we willen hebben want dit helpt ons bij groene doelstellingen of hier kunnen we ook mee marketen.

Do you think you have the required metrics to evaluate energy efficiency of AI models?

Ik denk dat we dat beter moeten meten, ik denk dat we dat meer in kaart moeten brengen. Dus daar zouden we wel degelijk als bedrijf naar kunnen kijken, van hoe maken we dit meetbaar. Op dit moment is het een beetje ontastbaar, het is er wel maar niet heel prominent. En we zien in de tests natuurlijk wel de snelheid maar hoe dat dan vertaald naar energieverbruik is niet duidelijk. Dus daar zou een KPI wel erg helpen om het in ieder geval tastbaar te maken. Want wat we eigenlijk doen in de ontwikkeling, is we zetten een aantal KPI's voor ons neer en dan sturen we de ontwikkeling op basis van die KPI's. dat zou je ook op het vlak van energieverbruik kunnen doen, en dat is dan misschien een interne KPI in plaats van een klant KPI.

Are you aware of any techniques that could reduce the energy consumption of AI?

Ja efficiëntere hardware, kleinere modellen, student teacher achtige constructie. Het outsourcen van trainingscapaciteit naar groenere servers, dat soort dingen. Maar ik claim niet dat ik daar een volledig overzicht op heb.

Zoiets als de gpus die we gebruiken en hoe verhouden die zich nou op gebied van footprint. Misschien kunnen we er ook anders naar gaan kijken, dat we zeggen van oke laten we vaststellen waar we op willen draaien en wat het beste daarin zou zijn en daar blijven we bij. In plaats van dat we zeggen, hey kijk er is een nieuwe GPU op de markt en die is weer krachtiger dus laten we die maar pakken.

Interview CD

Er staat natuurlijk in het strategisch plan, “wij minimaliseren altijd het energieverbruik van onze producten.” Waar komt die statement vandaan en waarom is dat strategisch belangrijk voor PV?
Eigenlijk is het een manier om op tijd te zijn bij de sustainability trend, we zien dat onze klanten steeds meer ermee gaan doen. Op het moment wordt er nog niet direct gevraagd om hele groene producten of offertes maar we zien wel dat dat eraan zit te komen. Wij hebben de keuze gemaakt om hier al mee te beginnen om de controle erover te houden, we willen niet reactief zijn met dit onderwerp en dan overrompeld worden.

Verder is het voor ons ook heel belangrijk om de relatie met de klant te versterken en de klant verder aan ons te binden.

Er zit natuurlijk ook een competitieve kant aan, we verwachten ook dat de klant eerder zal kiezen voor een groene oplossing, omdat ze zelf zeggen dat ze ermee bezig zijn.

Heb je een idee van de kwantitatieve waarde van de sustainability van de producten? Is dat iets wat wordt meegenomen in de strategische planning?
Nee, wij kunnen nog geen financiële waarde of cijfers verbinden aan de waarde van groenere oplossingen. Er is niet een bepaald percentage of indicatie die we hebben waar we op dit moment naar kunnen kijken waarmee we kunnen zeggen, oh we hebben dan een zoveel procent grotere kans om een opdracht te winnen bijvoorbeeld.
We hebben er wel vrede mee dat het misschien uiteindelijk ons meer gaat kosten dan dat het oplevert.

Waar zit de lijn voor PV, wanneer wordt er te veel opgeofferd voor sustainability? Want we hebben het natuurlijk nu over de producten van PV niet de omliggende initiatieven zoals zonnepanelen op het dak of de verwarming graadje lager.
Op dit moment is dat nog niet helemaal duidelijk, we hebben nog geen ervaring met het vergroenen van producten. Wat we wel zien, is dat we het niet als een optie kunnen gaan aanbieden waar de klant meer voor moet gaan betalen, dit moet een integraal onderdeel zijn van onze producten en niet een optionele offering. En we denken dat de klant dat ook meer gaat waarderen, zij zitten natuurlijk zelf ook in die situatie waarin ze groenere keuzes moeten gaan maken.

In de literatuur worden vier grote obstakels voor het implementeren van sustainability initiatieven genoemd, dus ik ben wel benieuwd of je PV herkent in deze obstakels:

Lack of awareness: M
Ik denk dat we heel erg ons best doen om onze impact in kaart te brengen en dit ook te communiceren binnen de organisatie, maar ik denk dat er nog wel een weg te gaan is voordat ik de claim zou durven te maken dat we helemaal aware zijn binnen Prime Vision.

Lack of resources: L
Faciliteren we het op dit vlak heel goed doen, het is niet alsof mensen geen tijd krijgen om aan sustainability te werken, ik denk dat we ook best wat tijd besteden en ruimte geven om dat te faciliteren. Maar uiteindelijk komt het ook neer op de motivatie van de medewerker zelf.

Lack of information: M/H
We meten natuurlijk al best wel wat qua onze reis bewegingen en dat soort dingen, maar op het gebeid van verbruik van onze producten missen we nog wel wat inzichten, ook omdat we niet alles weten van onze suppliers.
Als het gaat om informatie om op de lange termijn beslissingen te maken op het gebied van sustainability is het nog een beetje tasten in het duister. Een heleboel is gewoon nog niet duidelijk waar het naartoe gaat en daar zitten klanten denk ik ook nog een beetje naar te zoeken. Ze eisen nog geen groene producten maar dat zou ook kunnen veranderen.

Lack of perceived benefits: L
Nee ik denk dat we die hobbel wel over zijn nu, ik denk dat we wel de keuze hebben gemaakt om hiermee door te gaan. Ik denk dat we wel inzien dat sustainability van grote waarde is. Ik denk dat je daar tegenwoordig niet meer in achter kan blijven. Misschien dat er aan de operationele kant niet hetzelfde inzicht is, omdat je als je op de werkvloer staat bij wijze van spreken dan beschik je misschien niet over dezelfde informatie en misschien heb je dan niet het idee alsof het anders hoeft, ook omdat je misschien niet weet wat het alternatief is.

Interview CV 3

First off explain a little bit about what your role is in the vision team and what your day-to-day looks like.
Yeah so, my role is computer Vision engineer in the Research team of Sight.
Researching on ways to improve on our algorithms and these are generally deep learning-based methods.
So, my day-to-day work includes reading papers and running experiments, trying to improve performance on our current projects.

And would you say you are familiar with environmental sustainability and the energy consumption within your line of work?
Uh, not much, but I have an idea.
I mean, uh, deep learning training consumes a lot of energy, unfortunately, because it requires a lot of heavy matrix multiplication calculations during this process and it's kind of a rigorous process in the sense that you need to play around a lot. I experiment a lot and I realized that we're running a lot of GPUS constantly. They run over days and nights to get the most out of it, so I wouldn't say it's a waste of resources, but I think it consumes a lot of energy, let's say it's not a cheap process.
Even for bigger language models such as ChatGPT and like the mainstream models not even training them but inferencing on them on uses quite a lot of compute. So yeah, it's a bit resource intensive field let's say.

And have you personally thought about the energy consumption of the products during development before or is it discussed within the team?
Not really. I mean because for us the main goal is we have some devices, GPU storage, we should get the most out of them. So, it's not like not using them but using them. No, as efficient as possible so that we can use them more. Basically. So yeah, not very environmentally focused.

So, is there also then no feedback from customers or other stakeholders about energy consumption? Do you ever get questions about that?
Uh, not, not related to our training procedures.
Not this specifically. I mean, they probably they get some related to how do we store our data for example we use the what the interconnect for data storage and I think they are green energy storage center in the sense but not regarding to our own in facility servers and the energy consumption of those.
I don't know at least.

OK, so also not, not necessarily, let's say about the product itself?
No.

Do you think it would be worthwhile to start to reduce energy consumption where possible?
Of course, yeah.

And what do you think would be the best area to start improving that?
I mean, training is the most where the energy is mostly consumed.
So definitely that part. But yeah, to be honest, I'm not sure how you can really do that because as I said for us, we have some resources and we're trying to get the most out of them.
You can of course in your coding making it more efficient, but it's generally not to be greener but per say to get faster results so that we can run more experiments in the same amount of time. But since you're using the machine for a period of time, it doesn't matter if you do one experiment or 10 experiments in the period of time it's going to consume the same energy since it's up and running.
And yeah, for us, the main objective is like, let's do more with what we have, yeah.

Are you aware of any techniques from literature that's specifically focused on reducing energy consumption of training or inference of deep learning models.
So, the one research area is like, let's find models which are less resource intensive and which can give you the same performance. For example, smaller versions of the same models.
That's a smaller chat, GPT smaller and a bit less parameters which will be faster to go on. Another research area is let's work on the optimization procedure or the training procedure of the models in any model and then try to make it as efficient as possible.
This area I'm not very knowledgeable on we're not pushing research on this area.
The secondary is like the bigger is better, so doesn't matter how big my model is.
I want to go bigger if I can. If I have the resources and reach the best performance, so and we cannot also go to that area because we don't have as much resources as Google, Facebook and stuff.
So, for us it's like as an auxiliary task you said before, it makes more sense to go to smaller models and get this can get the same performance which will reduce the energy consumption as well.
And for us, it's easier to deal with basically.

And would you say that if you look at a paper that's that produces energy consumption, but maybe the performance is a little less. How would you go about evaluating that?
It depends on what we promised the customer because in general we promise numbers to customers like we will be this much accurate on these cases and stuff.
So that's, that's kind of what determines the threshold So if you're, if you're already satisfying the numbers. The required numbers then of course smaller is better. Why would we put a bigger model? We would, we wouldn't do that. But for us, the main concern is meeting the customer expectations.

And let's say that every everything can be met, right? The customer's expectations can be met. Everything is within the required thresholds. What other factors are there?
Yeah, I mean apart from the performance also one of the customer requirements like the how fast the algorithm runs rather than only performance itself.

So, you mean the latency?
The latency. Yeah.
Uh, so we would like to, I'd like to keep it minimum at it.
With minimum latency, we want to reach the maximum performance.
So according to that, whatever model is doing the job we would try to fix that.

And let's say outside of model performance. Are there any process related factors that would make you say: OK, this this isn't worth it?
So, we do research, but we're not like this, like a university research group, right.
So, the methods we try should have some, let's say proven performance and there should be relatively not super hard or we should have the feeling that OK this this will work for us if that's the case we will do it.
So, moving to ONNX or using Triton was one of these decisions that we made.
But let's say if it's like a research in very early stages and if we think that it will cause more problems to us, then the benefit then we wouldn't do that.

OK. So, it's also like the really proven ability to that, it's stable and works already.
Exactly. Yeah.

How much influence do you think it will have if I technique umm it saves a lot of energy for just a little energy, do you say OK we always pick the technique that saves a lot of energy even if it costs maybe a little bit of performance or would you say well we just choose anything that saves energy as long as we're hitting performance the amount of energy is then let's say secondary amount of energy saved.
Yes, I would say that energy saved is secondary.

Is there anything else that you can think that would really make an optimization technique for energy efficiency attractive?
That's uh on top of my head.
Not really, to be honest.

And how would you say like in general within the team, how would you say the awarenesses of energy consumption of the deep learning models?
I mean, I'm sure people are aware that it's kind of expensive or resource intensive let's say, but like in terms of quantity, I don't think people are really thinking of it like literally how much it's suspense. I'm not sure people are aware of it, and me, myself. I'm also not really aware.

So, there's not really any metrics right now where you're saying, oh, we know what the energy consumption is, or we know how to measure that.
I mean we can measure it, of course, because we know for how long it runs.
So I mean we could have tools to measure it, but we don't really measure it or analyze it or reflect on it basically. So, there's not much effort to say oh we should reduce this or this consumes too much energy. It's more like all we have this consumption and let’s fit as much as possible to this amount of consumption.

Interview TD

Wat is je rol binnen Prime Vision en wat zijn jouw verantwoordelijkheden?

Ik ben Technology Director bij prime vision, dag tot dag ben ik voornamelijk bezig met de mid tot lange termijn op het gebied van technologie om te kijken welke vernieuwingen er allemaal aankomen. Mijn hoofdthemas in de afgelopen jaren zijn met namen geweest: edge computing, A.I. en dan niet alleen vision maar ok copilot etc., Hybrid cloud en Security. Dat zijn de grote themas.

Zie jij dat er binnen prime vision veel aan wordt gewerkt om producten groener te maken?

We hebben natuurlijk een heel team daarop gezet (sustainability gilde), die zijn in de afgelopen jaren voornamelijk bezig geweest met de social capabilities, maar die zijn nu ook de slag naar software aan het maken. Zijn we genoeg bezig? Tsja we zijn natuurlijk nooit genoeg bezig.

Maar je ziet ook een vraag in de markt ontstaan, en wij zijn zelf in onze core altijd al mileu bewust bezig geweest. Alleen nu mogen we het wat meer gaan uitdragen en hete mag wat gaan kosten.

Is dit al een beetje doorgesijpeld naar product development?

Voor de hardware teams is dit wel een thema, die zijn echt bezig met waar haal ik mijn spullen vandaan en hoe ship ik ze etc. maar bij de softwareteams zit dat er minder in. En zeker bij de sight afdeling, dat zijn echt engineers die gefocused zijn op wat ze doen, dus ja die hebben echt wel door dat we groener moeten zijn. Maar als je bij de mensen zelf kijkt dan is het veel meer met het vuil opruimen en met de fiets gaan enzo, niet in de producten. Dus ja ze weten wat het is maar het is nog een beetje zoeken hoe ze dat nou integreren in hun basisproces.

En waarom is dat daar minder prominent?

Wij automatiseren en we willen steeds de nieuwste technologie bieden aan onze klanten. En dat betekent dat wij vrij klant driven bezig zijn geweest altijd. Dus dat betekent ook dat je naar je klantvraag moet gaan kijken. Als je dat doet dan moet de klant dus direct een vraag gaan stellen aan jou en dan speer je je daarnaar. En onze klanten doen zoveel aan sustainability, maar voornamelijk op het gebied van elektrische autos en transport, wat opzich logisch is in de logistieke sector.

Nu zie je dat wel een beetje omslaan dat onze klanten zich ook steeds meer aan het uitbreiden zijn op dat onderwerp en dat wij er ook aan onze kant meer belangstelling voor is. Als je nu met een CIO praat dan is de eerste priority security. En ook wel sustainability maar dan meer aan de kant van elektrische autos zegmaar.

Zou je zeggen dat PV dus redelijk reactief was op de klantvraag ook met betrekking tot sustainability?

Wij waren puur klant focus en we zijn nu steeds meer aan het focussen om te gaan productizen. Dan gaan we veel meer in producten en saas diensten leveren. En dan kun je en moet je veel meer proactief bezig zijn met toekomstige waarde van producten. En dan bekijk je klanten meer als een geheel in plaats van 1 specifieke klant.

framework

Lack of perceived benefits: L

Sustainability is voor ons wel een primaire zaak, maar wat ik nog niet zie is de benefits van de software kant. We zijn nog op zoek hoe kunnen we sustainability gaan embedden in de producten en wat voor impact gaat dat dan hebben.

Lack of information: H

Onze lack zit echt in wat levert het op, wat is de business case. Voor ons is het ook een beetje, waar moeten we in godsnaam beginnen. En hoe kunnen we de impact daarvan bepalen.

Lack of awareness: L

Iedereen weet wel dat er wat moet gebeuren, voor ons is het echt gewoon van hoe kunnen we dit dan gaan doen en hoe kunnen we onze impact meten.

Lack of resources:

Omdat ik die benefits niet weet ik niet of ik genoeg mensen heb. Ik zou deze op M zetten. We doen nu ook al een tijdje security en dat is eenzelfde verhaal. Als je dit echt wil oen dan moet je het gelijk goed doen en dan moet je het ok direct integreren en automatiseren. Als je dit met de hand gaat doen dan wordt het niks en dan heb ik inderdaad heel veel mensen nodig. Maar als ik dit in mijn proces integreer en automatiseer dan kan dat veel efficiënter. Als ik gewoon voor elke keer dat ik een load en performance test doe ook een check doe of ik groen bezig ben dan is dat veel sneller. Want die load en performance test moet toch wel.

Appendix C: Technique evaluation

Introduction

As AI models get larger and larger, and the number of use cases for AI continues to expand, the energy consumption of AI is expanding with it.

In existing research, techniques that tackle the energy consumption of using and developing AI have been identified, however the practical adoption of these techniques is low. I have created a framework which aims to facilitate the transfer of knowledge between academia and industry by evaluating the feasibility of practical implementation of these energy saving techniques.

To evaluate the energy saving techniques, three characteristics which can influence the implementation feasibility of these techniques were extracted from existing research. Since these techniques can be complex, it is important that the assessment of these characteristics is done by experts in their field (you).

So, I want to ask you to evaluate a few techniques on these three defined metrics.

In chapter 2 you will find the definition of the different characteristics that were extracted from the literature, in chapter 3 you will find the catalog of four energy saving techniques in AI development, and finally in chapter 4 you will find the assessment matrix and the instructions on how to fill it in.

Thank you very much for helping me with my research!

Glossary

Characteristics	*Definition*
Complexity	*Complexity* is defined as the technical complexity of the technique; this refers to the rarity of the knowledge required to implement the technique. If highly specialized knowledge is needed, then the *complexity* of the technique will be high. If specialized knowledge is required but can be acquired easily the complexity will be medium. If the skills and knowledge required is part of the already existing skillset the complexity will be Low.
Effort	*Effort* is defined as the estimated amount of work required to implement the technique in question.
Disruption	*Disruption* is defined as the impact that the implementation of the technique would have on the existing processes and tools. If implementation requires significant customization and/or interferes with the development process in a meaningful way, then the *disruption* is high. If implementation requires some changes but the overall process can remain the same, the disruption is medium. If there are no or only very minor changes to the process required to implement a technique, the disruption will be low.

Techniques catalog

Technique	Quote	Source	Theoretical Energy savings
Batch sampling	"We show that CNN training could be accelerated by a "frustratingly easy" strategy: randomly skipping mini batches with 0.5 probability throughout training. Stochastic mini-batch dropping (SMD)"	Wang et al, 2019 PDF	SMD = 33%
Energy consumption as a model constraint	"Leveraging the energy model, we augment the conventional DNN training with an energy-constrained optimization process, which minimizes the accuracy loss under the constraint of a given energy budget. Using an efficient algorithm, our training framework generates DNNs with higher accuracies under the same or lower energy budgets compared to prior art."	Yang et al, 2018 (https://github.com/hyang1990/model_based_energy_constrained_compression) PDF	32-74%
GPU-power limiting	"Power limiting can imply much lower energy consumption (up to 33% for V100 and Quadro 6000), along with a low to medium performance penalty."	Krzywaniak et al. 2022 PDF	Up to 33%
Energy Aware Pruning	"The experiments show that the proposed pruning method reduces the energy consumption of AlexNet and GoogLeNet, by 3.7× and 1.6×, respectively, compared to their original dense models."	Yang et al. 2017 PDF	24 – 73%

References

Wang, Y., Jiang, Z., Chen, X., Xu, P., Zhao, Y., Lin, Y., & Wang, Z. (2019). E2-train: Training state-of-the-art cnns with over 80% energy savings. *Advances in Neural Information Processing Systems*, *32*.

Yang, H., Zhu, Y., & Liu, J. (2018). Energy-constrained compression for deep neural networks via weighted sparse projection and layer input masking. *arXiv preprint arXiv:1806.04321*.

Krzywaniak, A., Czarnul, P., & Proficz, J. (2022, June). GPU power capping for energy-performance trade-offs in training of deep convolutional neural networks for image recognition. In *International conference on computational science* (pp. 667-681). Cham: Springer International Publishing.

Yang, T. J., Chen, Y. H., & Sze, V. (2017). Designing energy-efficient convolutional neural networks using energy-aware pruning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5687-5695).

Assessment Matrix

Please look at the techniques and the associated sources in the techniques catalog and evaluate each of the techniques on their complexity, the effort required for implementation, and the disruption to the process and tools you think these techniques would cause if implemented at PV. Please also provide a short explanation in the cell for your choice of High, Medium, or Low for each of these characteristics.

Technique	Complexity (L/M/H)	Effort (L/M/H)	Disruption (L/M/H)
Batch sampling			
Energy consumption as a model constraint			
GPU-power limiting			
Energy Aware Pruning			

Appendix D: Coding Manual

Code	Rules
Lack of perceived benefits	This code will be assigned when: • The interviewee states that they do not see the benefit of reducing energy consumption. • The interviewee does not see reducing energy consumption as a priority. • The interviewee shifts responsibility to a third party.
Lack of awareness	This code will be assigned when: • The interviewee is not aware of environmental sustainability issues in their line of work. • The interviewee does not consider energy consumption in their work-related activities.
Lack of information	This code will be assigned when: • The interviewee states that energy consumption is not measured. • The interviewee states that energy metrics could assist in making improvements. • The interviewee states that they are unaware of the direct causes of energy consumption. • The interview states that they do not know how to reduce energy consumption.
Lack of resources	This code will be assigned when: • The interviewee mentions being in a time crunch, or generally very busy/ no time for anything else. • The interviewee states that physical resources (like computing power) are constraining his/her work. • The interviewee mentions lack of tool support to reduce energy consumption.
Complexity	This code will be assigned when: • The interviewee states that they do not possess the required expertise to implement a certain technique. • The interviewee states that implementing a technique would make their work more complex.
Effort	This code will be assigned when: • The interviewee states that the amount of time required to implement a technique would be an inhibiting factor
Disruption	This code will be assigned when: • The interviewee emphasizes the importance of maintaining the development process. • The interviewee states a fear of quality downsides in the product due to technique implementation. • The interviewee notes the compatibility of techniques with existing tooling as an issue.
-E%	This code will be assigned when: • The amount of energy saved is used as an argument for or against implementation of a technique. • The interviewee brings up the positive environmental impact (or lack thereof) of a technique.

Appendix E: Framework Reference Matrices

	Perceived Benefits	*Awareness*	*Information*	*Resources*
L	--	-	--	--
M	-	0	-	-
H	+	+	++	+

	Complexity	*Effort*	*Disruption*	*Energy Reduction*
L	0	++	++	0
M	0	0	0	0
H	-	--	--	+

Characteristics / Barriers		Complexity		
		L	M	H
perceived benefits	L	--	--	---
	M	-	-	--
	H	+	+	0

Characteristics / Barriers		Complexity		
		L	M	H
environmental awareness	L	-	-	--
	M	0	0	-
	H	+	0	--

Characteristics / Barriers		Effort		
		L	M	H
perceived benefits	L	0	--	----
	M	+	-	--
	H	+++	+	-

Characteristics / Barriers		Effort		
		L	M	H
environmental awareness	L	+	-	---
	M	+	0	-
	H	+++	+	-

Characteristics / Barriers		Disruption		
		L	M	H
perceived benefits	L	0	--	---
	M	+	-	---
	H	+++	+	--

Characteristics / Barriers		Disruption		
		L	M	H
environmental awareness	L	+	-	---
	M	++	0	--
	H	+++	+	-

Characteristics / Barriers		Energy Reduction		
		L	M	H
perceived benefits	L	--	--	0
	M	-	-	+
	H	+	+	++

Characteristics / Barriers		Energy Reduction		
		L	M	H
environmental awareness	L	-	-	0
	M	0	0	+
	H	+	+	++

Characteristics / Barriers		Complexity		
		L	M	H
information	L	--	--	---
	M	-	-	--
	H	++	++	+

Characteristics / Barriers		Complexity		
		L	M	H
resources	L	--	--	---
	M	-	-	--
	H	+	+	0

Characteristics / Barriers		Effort		
		L	M	H
information	L	0	--	----
	M	0	-	--
	H	++++	++	0

Characteristics / Barriers		Effort		
		L	M	H
resources	L	0	--	----
	M	0	-	--
	H	+++	+	+

Characteristics / Barriers		Disruption		
		L	M	H
information	L	0	--	----
	M	+	-	--
	H	++++	++	0

Characteristics / Barriers		Disruption		
		L	M	H
resources	L	----	--	----
	M	---	-	---
	H	+++	+	-

Characteristics / Barriers		Energy Reduction		
		L	M	H
information	L	--	--	-
	M	-	-	0
	H	++	++	+++

Characteristics / Barriers		Energy Reduction		
		L	M	H
resources	L	--	--	-
	M	-	-	0
	H	+	+	++

Appendix F: Evaluation interviews

Framework Evaluation CV1

Technique	Complexity (L/M/H)	Effort (L/M/H)	Disruption (L/M/H)
Batch sampling	L The Batch Sampling technique was quite simple, it didn't affect the complexity of the architecture or anything at all.	L All you have to do is add a condition to drop 50% of samples, it is not difficult to implement and it is not complex.	M I saw in the paper that there was around a 3% performance hit possibly. You can maybe optimize that away but that means you have to put in effort somewhere else.
Energy consumption as a model constraint	H This is quite complex to implement, but I don't really mind that I like the challenge.	H Because it is quite complex it might take some effort to integrate and work out.	L I think disruption is quite low, we can do all of this internally in our training process and it didn't seem to affect the accuracy that much.
GPU-power limiting	M I found this to be medium complexity because you have to set it up for every GPU you use, and at the same time do it for the right drivers and you need to know the right limits. And that knowledge needs to be learned and presented to the person that is doing the training.	M I think it is the same for effort, because you have to do it for everyone and then the whole team has to get on board and get that done. I also think that the disruption and effort are linked, if the disruption is higher, the effort we need to	H (changed to L) I said disruption was high because I read that the performance hit was like average 5% and that is on a classification network so that is quite high. We usually only fluctuate by 0.5 percentage points on performance. (Interviewee changed answer to L since the paper was vague in using the term "Performance". Instead of accuracy,

		put in to fix it will also become higher.	performance was used as a term to refer to duration of training, 20-30% extra duration was not seen by the interviewee as a large hindrance.)
Energy Aware Pruning	H	H	M

Framework Validation

Framework was filled in with the evaluations given by CV1

Interviewer

Now that we have filled in the framework, would you say you agree with the outcomes?

CV1

I think that the batch sampling is quite simple, I think that could really be green. Because with the three percent performance loss it is bad but because the complexity and effort are low, it wouldn't hurt to try out and maybe you can deal with performance somewhere else. So for me complexity and effort has higher weightage here, Complexity does not really matter that much, but as long as the effort is low we can try it out anyway. I also think that if it saves enough energy it could be worth it.

Interviewer

So would you say that if it is balances with high energy savings the disruption doesn't have to be a deal breaker because you can fix it in another way?

CV1

Yeah exactly, because the thing is, that disruption then translates into efforts increasing somewhere else right. So that requires some separate incentive but yeah if the benefits are high enough that should be doable.

Interviewer

So complexity is not that much of a limiting factor, except if it was really high. But effort is more of a factor in deciding whether to try a technique or not?

CV1

Yeah I would say so.

Interviewer

Alright. Right now, disruption is the most influential factor in this framework. Would you say you agree with that?

CV1

I mean yes, I understand that. But I think that the expectation is that there is always going to be some disruption from these techniques and that has to be tackled in a separate way. I think it's always going to be there so I think its natural for disruption to be elevated, but if it brings enough energy savings then I think it could be done. It just means more effort somewhere else.

Interviewer

I also wanted to ask you for your opinion on the organizational barriers. These were determined by speaking to the vision engineers, product manager and technology and commercial directors. Would you say you agree with the values for these barriers?

CV1

Let's see, lack of perceived benefits are Low, Lack of environmental awareness low, lack of information High and Lack of resources high. Yeah I would say I agree with that, that sounds reasonable to me.

Interviewer

Do you have any final thoughts or remarks?

CV1

The interesting thing that I found just going through the papers that it seems there is quite some good studies on using the GPUs differently and it is good to see that there is room for improvement. I also thought the papers you found were very nice, they were very descriptive and answered good questions.

Framework results based on technique evaluation

	Characteristics			
Techniques	**Complexity (L/M/H)**	**Effort (L/M/H)**	**Disruption (L/M/H)**	**Energy Reduction**
GPU Power Limiting	M	M	L	M
Batch Sampling	L	L	M	M
Energy Aware Pruning	H	M	H	H
Energy consumption as a model constraint	H	H	L	M

Organizational Barriers			
Lack of Perceived Benefits (L/M/H)	**Lack of Environmental Awareness (L/M/H)**	**Lack of Information (L/M/H)**	**Lack of Resources (L/M/H)**
L	**L**	**H**	**M**

Influence on weights

Complexity = no change

Effort = more influence

Disruption = less influence of medium disruption

Energy Reduction = no change

Framework Evaluation CV3

Technique	Complexity (L/M/H)	Effort (L/M/H)	Disruption (L/M/H)
Batch sampling (active learning) **Suggestion from CV3:** You should also look at active learning, active learning aims to select the best subset from your whole dataset to train on. And this improves both your performance, and you need less data and thus less compute.	M It is easy to apply for us, you can just drop half of the batches. But we would rather not drop batches randomly because then maybe the data would get skewed. We are very good at reading machine writing and less at handwriting. We also have way more samples for machine writing so we would rather drop more machine writing samples than handwriting. So, we would have to tweak it a little bit. That is why active learning would be better.	M For the random sampling I would say it is very easy, but if we want to do smarter sampling then maybe it would be more effort.	L If it was again random sampling then disruption would be M, but if we do it smarter then it would be L.
Energy consumption as a model constraint	M This is more complex and it requires modifying the training procedure.	H It doesn't quite fit with the techniques that we want to be using.	H It requires quite some manual modification in our training system, so that's why I didn't really like it. It also focuses a lot on CNN and we are moving more and more to transformer based model. Something we use for energy reduction shouldn't cause more energy consumption so it should be very generic and very easy to implement.

GPU-power limiting	L It is low complexity because it is just capping the power, its something you can do with nvidia itself. So you don't change your model, you don't touch the architecture etc.	M You have to find the sweet spot for the power limit and that might not be the same for all the models. So you might have to do some experiments.	L There is no performance loss and you don't have to change anything in your process. This is one I really want to play with and see if its worth it.
Energy Aware Pruning	M You have to change the model architecture and that always adds complexity.	H	H

Framework Validation

Interviewer

Which of the three characteristics do you think has the most influence on whether you would implement one of these techniques?

CV3

I'll say disruption is the most influential. Because complexity, it matters of course but if its going to bring value we already go for complex things so its not really much of a problem. Effort is also maybe correlated with complexity but effort is a little more influential than complexity. But if something is not disruptive but is very complex and takes high effort, I will do that.

Also if something is complex and takes effort, you learn from it and get better at it, but if you really have to change how you do things its really a pain.

Interviewer

Do you also say that because the effort might be a one-time thing whereas the disruption could be ongoing?

CV3

Yeah for sure, because you have to do it again and again and it probably wont go away in the future either.

Framework was filled in with the evaluations given by CV3

Interviewer

Is there anything that stands out to you that you maybe don't agree with in the results? You said that gpu power limiting would be best but it is orange here and batch sampling is green.

CV3

Right so I think I agree with that, its just that the batch sampling would need some customization right. So we can make it work but not out of the box, we would need to use more smart sampling. As it is I would not do it.

GPU power limiting is new to me and it sounds exciting so it is more a personal thing that I would like to try. But yeah the list looks correct if I look objectively.

Interviewer

Would you say that maybe if you apply the batch sampling as it is, the disruption would be higher?

CV3

Yeah, I would say so.

Interviewer

So, if we look at your idea of smart sampling with active learning, how would you rate the technique then?

CV3

I would say complexity is medium, and effort is also medium, and then probably disruption is low.

Interviewer

So, it turns to orange then. So, would you say that complexity here doesn't matter at all?

CV3

It really depends also on how much energy you're saving right. If it's highly complex and it only saves maybe a few percent then maybe I wouldn't do it.

Interviewer

So, would you say that energy is more of a positive factor than that complexity is a negative factor?

CV3

I would say so yes

Interviewer

Earlier we shortly discussed the organizational barriers but I want to ask you if you agree with the assessment of these barriers that is made here.

CV3

I would say we are not really aware of anything. I think if you really want to do something for the environment, there is other things that you can do before trying to optimize your models. So I think that the impact here is low. I am a vegetarian and I think I am doing the best for the environment in that way, and I think that the meat industry does way more damage to the environment than our GPU training.

Its nice of course, and I don't want to diminish the effort. I do think that there is a lack of awareness, we do some actions like only going with one person to clients and stuff if we don't need more. And I think that's good and I don't want to diminish the effort, but I think that there should be more thought into what is the actual impact of this action that were taking.

Interviewer

So would you say that that is a lack of awareness, or a lack of information because the impact is not really measured or known? We can't make people aware of the impact because we don't know it exactly.

CV3

Yeah exactly.

Framework results based on technique evaluation

	Characteristics			
Techniques	**Complexity (L/M/H)**	**Effort (L/M/H)**	**Disruption (L/M/H)**	**Energy Reduction (L/M/H)**
GPU Power Limiting	L	M	L	M
Batch Sampling	M	M	L	M
Energy Aware Pruning	M	H	H	H
Energy consumption as a model constraint	M	H	H	M

Organizational Barriers			
Lack of Perceived Benefits (L/M/H)	**Lack of Environmental Awareness (L/M/H)**	**Lack of Information (L/M/H)**	**Lack of Resources (L/M/H)**
L	**L**	**H**	**M**

Influence on weights

Complexity = no change

Effort = no change

Disruption = less influence of medium disruption

Energy Reduction = more positive effect than negative effect of complexity

Framework Evaluation CV2

Technique	Complexity (L/M/H)	Effort (L/M/H)	Disruption (L/M/H)
Batch Sampling	WH: M. More time spent on performance optimization.	WH: L. Implementation seems straightforward. Omdat het redelijk makkelijk te implementeren is, kan je het altijd een keer uitproberen. En als je er dan achter komt dat het voor een bepaalde architectuur dit ook redelijk strafloos kan doen dan kan je voor dat soort modellen dat al meenemen.	WH: L. Implementation fits current training process; In de paper zeggen ze dat het erg simpel is, maar je hebt wel toch nog een soort performance impact. Het implementeren is triviaal, maar als je daardoor meer moet trainen om de performance te optimaliseren dan krijg je misschien een win/loss.
Energy consumption as a model constraint	WH: M. More time spent on performance optimization;	WH: M. Implementation seems straightforward, but influences overall training process	WH: M. Additional processing in current pipeline Ik denk dat het trainingsproces hierdoor aardig wordt aangepast. Dat zal ook wel afhangen van de implementatie verwacht ik.
GPU-power limiting	WH: L. No development required	WH: M. Tailor made optimization per GPU/Model Omdat je hier per GPU-model een andere afstelling moet	WH: M. Training takes longer De training kan wat langer worden en dan moet het wel passen met planning.

		maken en we hebben nogal wat verschillende modellen GPU's is de effort hier wel wat hoger.	Voor het Solutions team zal het echt in de planning moeten passen, want die hebben gewoon een bepaald tijdslot om een model te trainen. Maar voor de research activiteiten zal het meer te maken hebben met convenience denk ik.
Energy Aware Pruning	WH: M. More time spent on performance optimization	WH: H. Implementation from paper	WH: M. Additional processing in current pipeline Performance drop is redelijk bescheiden maar zelfs 1-2% is al redelijk hoog voor onze processen.

Framework Validation

Framework filled in according to evaluations CV2

Interviewer

Ben je het eens met deze assessment van de technieken op basis van jouw evaluaties?

CV2

Ja misschien dat bij de power limiting de effort en disruption wat meer mee gaan wegen. Want dat leek me er juist wel eentje van die kunnen we wel uitproberen. Want dat is een switch die je aan kan zetten mits je die energy reduction ook kan meten. En dat is ook een beetje, baat het niet dan schaadt het niet, als het niet zo veel werk is om het uit te proberen. Het is een ander geval natuurlijk om het daadwerkelijk te adopteren.

Ja dus die eerste twee zijn wel degene die het dichtste bij liggen en dat die batch sampling groen is dat herken ik wel ja.

Interview

Uit de drie karakteristieken die je hebt beoordeeld, welke heeft volgens jou de meeste impact?

CV2

Vanuit mijn eigen rol is de complexiteit wel een belangrijke, als de processen complex worden of lastig te onderhouden en implementeren dan wordt de code waarmee je moet werken ook lastiger. En ik kan me voorstellen dat voor het solutions en researchteam het trainingswerk vervelend kan worden want dat gaan ze in hun eigen werk merken. En effort ja ook wel een beetje.

Ik denk dat als bedrijf uiteindelijk effort vaak de doorslaggevende is. Ook al heb je dan wel een betere footprint, dus ja.

Interviewer

Oke dus voor jou is de complexity wat belangrijker. Je had het net over het complex worden van het proces, hoe bedoel je dat precies?

CV2

Ja wat je graag wil is dat je de processen die je hebt dat die aan een aantal criteria voldoen en dan ze goed te reproduceren zijn en dat mensen ervoor kiezen om ze te gebruiken. En het lastige van bijvoorbeeld een ander train proces is dat mensen ook een alternatief hebben. Dus dan kan je krijgen dat mensen zeggen, ja is allemaal leuk en aardig die groene dingen, maar als ik gewoon een git repo pull en ik druk op train, dan is dat

veel sneller dan als ik een framework moet gaan aanpassen en ik moet opgeven welke gpu ik gebruik en dan moeten we de scheduler weer checken, en dan moeten we nog even checken of de goede driver wel is geïnstalleerd.

Dus als het veel handwerk is wordt het lastig. Of je kan het natuurlijk automatiseren, en daar zijn we steeds meer mee bezig. Maar als die automatisering dan steeds complexer wordt dan wordt het ook minder stabiel, dan hoeft er maar iemand een driver te updaten en dan opeens werken je train processen niet meer.

Interviewer

Dus eigenlijk zorgt de complexiteit voor een complexer proces, en dat kan uiteindelijk weer zorgen voor disruption?

CV2

Ja inderdaad, of disruption in de vorm dat mensen het niet gaan gebruiken. Complexiteit hoeft niet erg te zijn want het kan ook een voordeel hebben als ik het wel doe.

Interviewer

Ja dus dat is dan weer een trade-off die gemaakt moet worden.

Hoe belangrijk is de hoeveelheid energie die wordt bespaard met een techniek voor jou?

CV2

Dat is een lastige vraag, alle beetjes zouden moeten helpen. Wat bij mij nog een beetje ontbreekt is wat het energieverbruik daadwerkelijk is. Bijvoorbeeld hoeveel verbruikt het trainingsproces ten opzichte van het normale kantoor energieverbruik. Je wil natuurlijk niet een heleboel energie en tijd steken in iets wat aan het einde maar heel weinig bijdraagt. Dus dan ben ik wel echt benieuwd naar de absolute waardes, dus echt hoeveel KWh. Want met percentages weet je nog niks als je niet waarvan je een percentage neemt.

Framework results based on technique evaluation

	Characteristics			
Techniques	**Complexity (L/M/H)**	**Effort (L/M/H)**	**Disruption (L/M/H)**	**Energy Reduction (L/M/H)**
GPU Power Limiting	L	M	M	M
Batch Sampling	M	L	L	M
Energy Aware Pruning	M	H	M	H
Energy consumption as a model constraint	M	M	M	M

Organizational Barriers			
Perceived Benefits (L/M/H)	**Environmental Awareness (L/M/H)**	**Available Information (L/M/H)**	**Available Resources (L/M/H)**
H	**H**	**L**	**M**

Influence on weights

Complexity = no change

Effort = more influence

Disruption = no change

Energy Reduction = no change