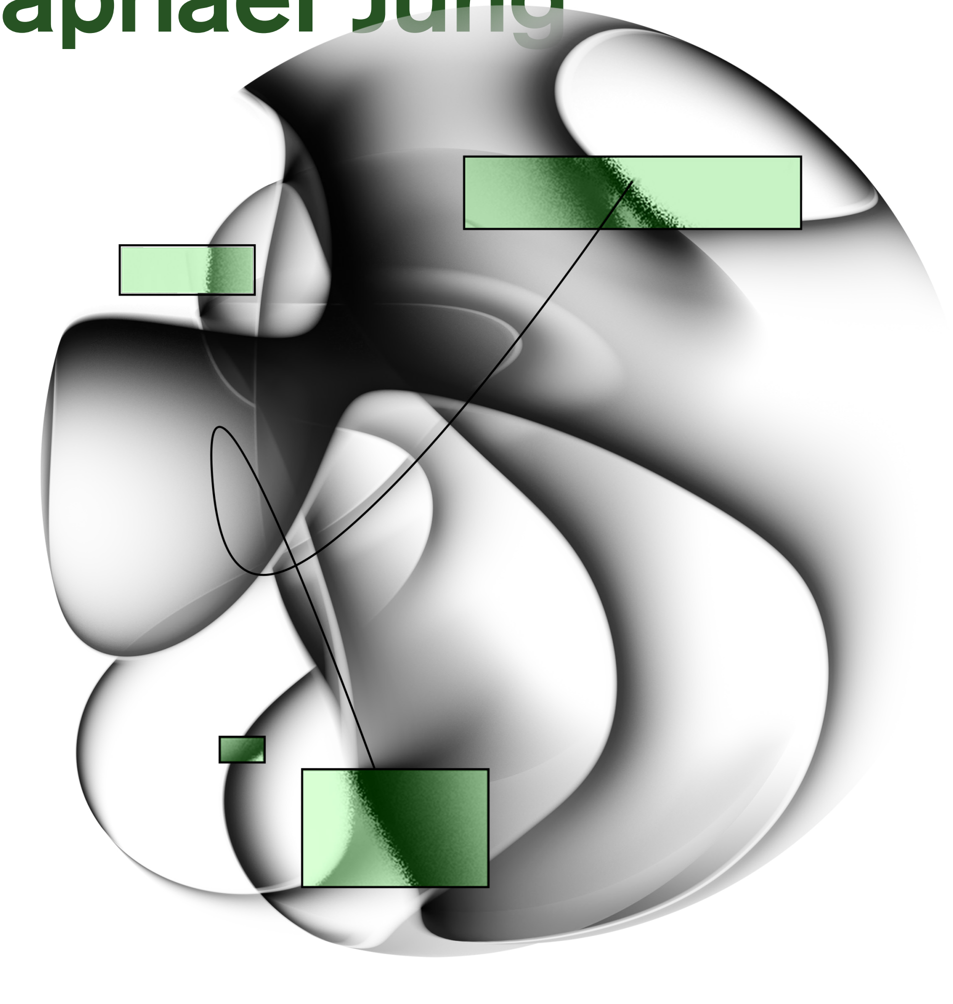
Towards Environmental Sustainability of GenAI

A strategy framework across the lifecycle

Raphael Jung



N° 5626056 MSc Thesis - Delft University of Technology 23.05.2025

This is a MSc thesis of Strategic Product Design from the Faculty of Industrial Design Engineering at the Delft University of Technology.

It is authored by Raphael Jung (No 5626056) and supervised by
Prof. Dr. Gerd Kortuem and Prof. Dr. Conny Bakker.

The thesis is based on the state of knowledge before the 23rd of May 2025.

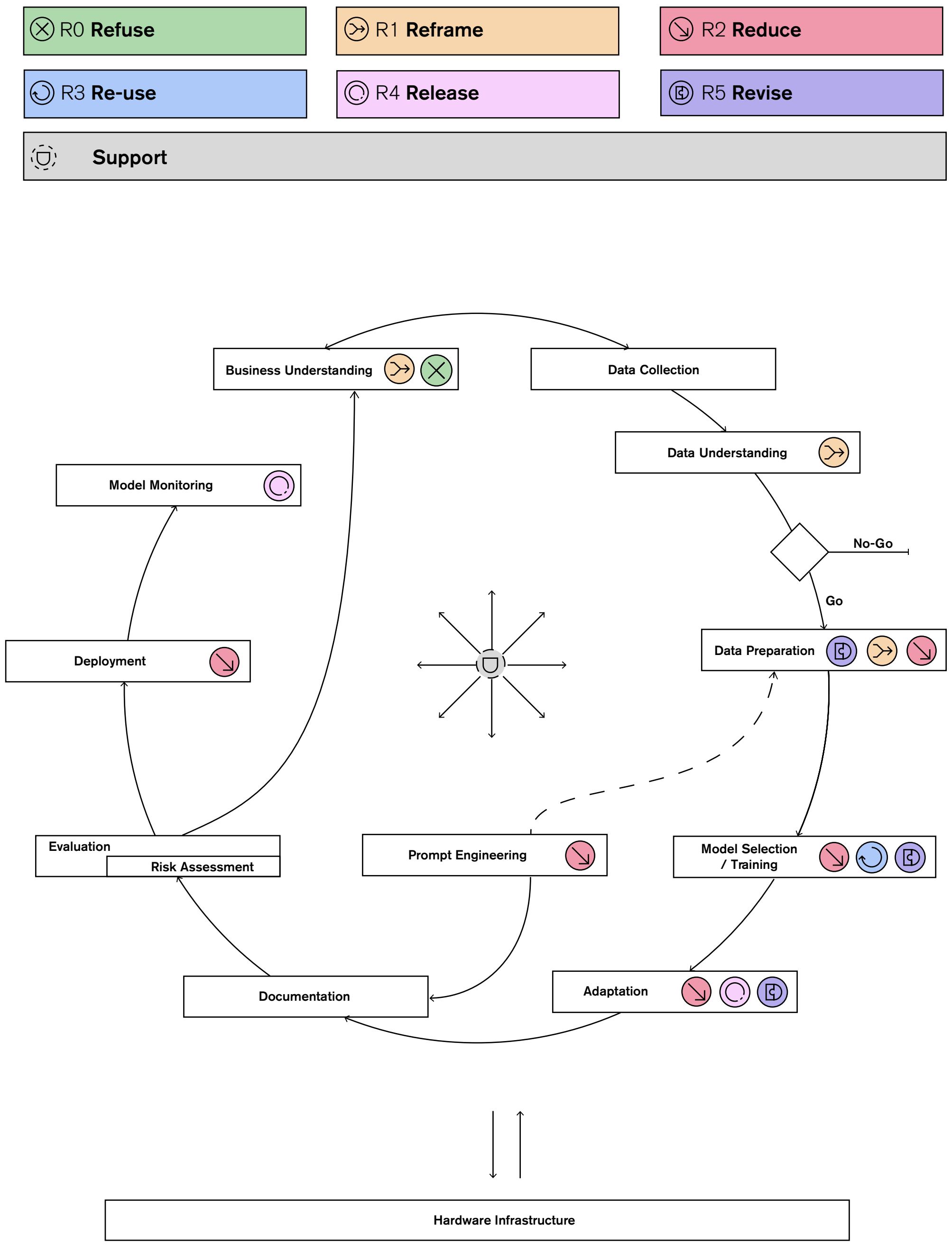
Abstract

The proliferation of Generative Artificial Intelligence (GenAI) poses substantial environmental challenges, including escalating energy consumption, water usage, and material extraction, which strain vulnerable planetary systems.

This work examines how the development and deployment of GenAI can be aligned with environmental sustainability objectives. It proposes a framework that categorizes sustainability strategies into seven distinct types: Refuse (abandoning the function GenAI was intended to fulfilll or employing alternative means), Reframe (modifying the context in which GenAI is embedded - such as governance or project framing - to reduce the number and scale of models), Reduce (efficiency enhancements to the technology itself to lower resource consumption), Reuse (reusing a model in a new context while preserving its functionality), Release (updating a model to restore its functionality, e.g. through data updates or bug fixes), Revise (using components of an existing model to develop a new one, such as via transfer learning), and Support (measures that increase the likelihood of adopting other strategies without directly reducing environmental impact, such as impact quantification). These strategies are systematically mapped to the lifecycle stages of GenAI, showing where each type can be applied. Field applications of the framework are already underway, underscoring its practical relevance and potential for real-world impact. Find the framework on the following page.

To assess the current state of research, a comprehensive scoping study was conducted across IEEE Xplore and Web of Science. The objective was to identify practical examples aligned with each strategy type and to expose research gaps. While approaches were found for all categories, their distribution was highly uneven: Reduce strategies dominated, followed by Reframe. The remaining types - Refuse, Reuse, Release, and Revise - were considerably underrepresented, highlighting the need for further investigation.

Building on these conceptual and empirical insights, the framework was operationalized through the development of a governance blueprint within a global professional services firm focused on IT. Designed through field research and participation in an active GenAI development project, the blueprint translates the framework into practice by identifying suitable sustainability strategies during development and embedding sustainability guardrails for roll-off. Its applicability was validated through stakeholder interviews across strategic, managerial, and technical domains. The resulting model offers a practical foundation for governing the environmental sustainability of GenAI, demonstrating how the conceptual framework can be used to inform industry practice.



Content

Abbreviatons

AI	Artificial Intelligence
ANN	Artificial Neural Network
CISP-DM	Cross-industry standard process for data mining
CoE	Center of Excellence
CSDDD	Corporate Sustainability Due Dilligence Directive
CSRD	Corporate Sustainability Reporting Directive
DL	Deep Learning
FFNN	Feed-forward Neural Network
GAN	Generative Adviserial network
GenAI	Generative Artificial Intelligence
GPT	Generative Pretrained Transformer Model
IT	Information Technology
KPI	Key Performance Indicator
LLMs	Large Language Models
PoC	Proof of Concept
SLMs	Small Language Models
ML	Machine Learning
MVP	Minimum Viable Product
NLP	Natural Language Processing
PB	Planetary Boundary Framework
ReLU	Rectified Linear Unit
RQ	Research Question
SCI	Software Carbon Intensity
SNN	Spiking Neural Network
UI	User Interface
VAE	Variational Autoencoders

Glossary

As abstract concepts are discussed within this work, it is important to clearly differentiate between concepts and their meanings. Within this thesis the listed concepts are defined as the following:

Approach — *A concrete and defined action. In this context, it refers specifically to an action aimed at improving the environmental sustainability of an application or model - for example, rigorous data cleaning.*

Strategy Type — *A cluster of approaches that all aim to achieve the same specific effect by using a shared mechanism. In this context, a strategy type might include a sequence of actions that all contribute to improving model efficiency.*

Strategy — *A planned formation or sequence of approaches that collectively aim to achieve an overarching goal - in this case, the reduction of environmental impacts. A strategy can consist of multiple strategy types; for example, some elements might focus on improving model efficiency, while others target a reduction in overall usage.*

Framework — *A basic structure underlying a system or concept. It serves as an abstract representation of the system.*

Blueprint — *A plan or programm of action.*

Note that chapter specific wordings are defined within the corresponding chapters.

Introduction

In the fast-paced world of technological advancement, progress seems driven by competition, curiosity, or the pursuit of bigger, better, faster solutions - often at the expense of environmental sustainability. As our advancements push the limits of our planet's capacity, we risk undermining the ecosystem that sustains us. This raises an urgent challenge: How can we guide technological development to serve humanity as a whole while respecting environmental boundaries?

This thesis investigates how the development and application of generative artificial intelligence (GenAI) can align more closely with environmental sustainability. Making sense of the context, technology and processes present in the field. The thesis is divided into three parts: The background, a conceptual framework and an exploration of the practical applicability of the framework in a sustainability blueprint for a consultancy and IT firm. In the background, literature is explored to provide an understanding of the context and challenges that come with GenAI development and implementation in regard to their sustainability. The goal is both to provide sufficient knowledge to make sense of the research landscape and following chapters and to lay out the purpose and impact of this project. The framework section proposes a systematic approach to connecting the GenAI lifecycle to concrete sustainability approaches by proposing a novel taxonomy of sustainability strategies in the context of GenAI and their relation to the lifecycle phases. Lastly, the gained knowledge and designed framework are integrated in a sustainability blueprint to inform development projects for GenAI based applications.

The intended audience includes researchers and industry professionals such as project managers, product owners, and executives. To ensure accessibility to non-AI-experts, foundational concepts of GenAI are explained in the appendix.

The ultimate goal is to empower diverse industry actors with a means to actively influence the development trajectory of GenAI toward a future that is both technologically advanced and environmentally sustainable.

Research Objective

Context

The rapid expansion of generative artificial intelligence (GenAI) has made it a crucial tool for business organizations, with the market projected to grow by 212 %, reaching $206.95 billion by 2030 (Statista, 2025). This growth comes at a time in which the world faces significant environmental challenges, such as climate change and resource depletion (Falk et al., 2024).

Proponents of AI-driven innovation highlight GenAI's potential to solve complex business problems and improve operations, including contributions to sustainability efforts. Yet, these claims are tempered by increasing concerns from researchers and environmental advocates, who point to the high environmental costs of GenAI technologies (Falk et al., 2024; Robbins & van Wynsberghe, 2022; van Wynsberghe, 2021; Wu et al., 2021). Training and deploying GenAI models consumes vast amounts of energy, and operating AI infrastructure significantly increases the carbon footprint and resource consumption (Falk et al., 2024; Wu et al., 2021).

Despite GenAI's potential to drive sustainable innovation in business, current policies and practices do not adequately address its environmental impact (Bashir et al., 2024; van Wynsberghe, 2021).

This lack of a structured and comprehensive perspective presents a critical research gap. Without a clear understanding of the sustainability strategy types and mechanisms applicable across the entirety of the GenAI lifecycle, organizations risk missing opportunities for meaningful intervention and create further devastating impact. In practice, this can lead to unbalanced development efforts, inefficient allocation of resources, and the perpetuation of environmentally harmful practices.

As GenAI continues to scale, embedding sustainability into its lifecycle is not merely desirable—it is essential for long-term responsible innovation.

Research Questions

This thesis addresses the gap by developing a conceptual foundation for environmental sustainability in GenAI development and further explores how this conceptual foundation can be utilized to inform practice in the business realities. To achieve this, it is guided by five core research questions:

RQ1: Which lifecycle stages does a GenAI model experience?

This research question serves to outline the entirety of the lifecycle. This is needed to ensure, that the further findings are able to address the entirety of the process and not miss on any of the stages.

RQ2: *Which sustainability strategy types can be applied to the GenAI lifecycle?*

This question returns the core of this research. The goal is to find an exhaustive collection of abstract mechanisms into which all applicable sustainability approaches can be categorized into. Each mechanism serves as the definition for a strategy type, that together provide an exhaustive overview.

RQ3: *In which lifecycle stages can which sustainability strategies be applied?*

Linking the strategy types to the lifecycle stages is valuable to both effectively identify suitable approaches per lifecycle stage and create a richer and more informative mapping of the landscape. This research questions merges the findings of RQ1 and RQ2.

RQ4: *How mature is the scientific research landscape around the sustainability strategy types for GenAI?*

To understand how much research has been conducted across the various strategy types and lifecycle stages returned from RQ3, the current research landscape is examined. RQ4 allows to identify both the presence of sustainability approaches across the lifecycle, as well as their absence. The absence presents research gaps, while the present ones can be used to inform practice.

RQ5: *How can the knowledge collected from RQ1 to RQ4 be applied to industry practice?*

The final research question examines how the conceptual framework and the insights derived from the preceding research questions can be translated into industry practice. This involves engaging with the realities of a specific organizational context and developing a sustainability blueprint tailored to that setting. The process serves a dual purpose: it validates the applicability of the framework as a foundation for practical implementation and illustrates how such implementation can be operationalized in practice.

Together, these questions aim to build a structured understanding of GenAI sustainability and provide explorations on how this conceptual knowledge can be translated into action.

Process & Deliverables

To systematically address the research questions, the research follows a multi-stage process (see Figure 1).

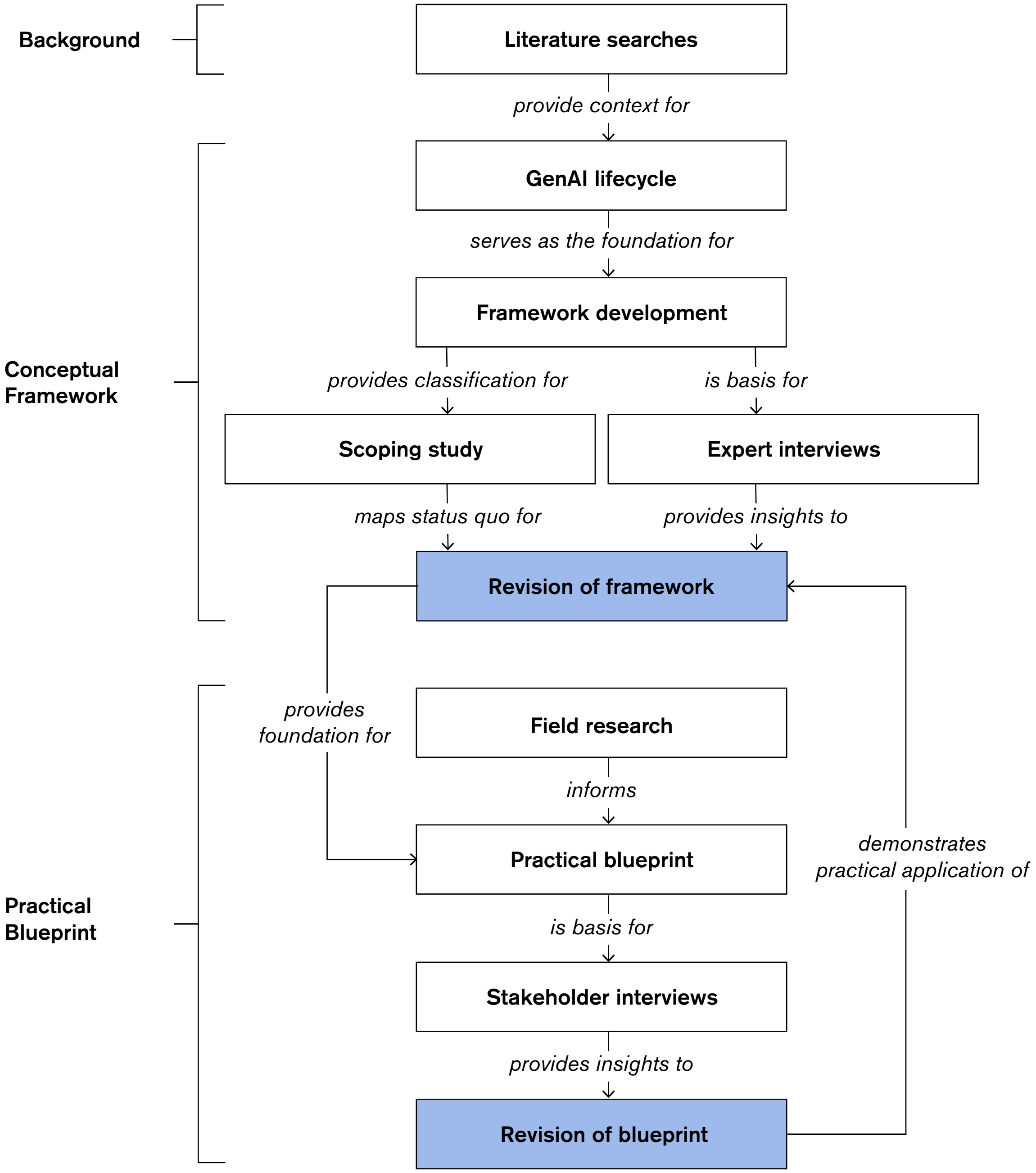


Figure 1: Project Overview

The first stage is the background stage. Here the research landscape is searched to provide contextual knowledge on the technology's application in businesses and the resulting environmental impacts of its use. This stage serves to explore and describe the environment of this research and provide contextual knowledge to generate sufficient understanding to adequately comprehend the work presented. For readers that are novices to the field of GenAI, Appendix A provides an overview of the technology.

The second stage aims to produce the first deliverable, a contextual framework defining the various sustainability strategy types applicable and their allocation to the various lifecycle stages of GenAI. Within this stage, an adapted lifecycle model is presented, based on preexisting models in research. This section addresses the research question 1. Through the activities incorporated in this lifecycle model, the first iteration of the framework is presented, addressing research question 2 and 3. On the basis of this framework, expert interviews are conducted, to gain insights and further ideate on the developed framework. Additionally, a scoping study is conducted, exploring white and grey literature to identify and allocate existing approaches to the strategy types and the corresponding lifecycle stages, addressing research question 4. From the insights of the scoping study and the expert interviews, the framework is revised. The revised framework serves to fully address RQ2 and RQ3 as well as serving as the first key deliverable of the project, while the results of the scoping study present the second deliverable. This conceptual framework showcases the different environmental sustainability strategy types which can be applied to the development of GenAI based applications. Due to its conceptual nature, it can be used to categorize existing specific sustainability approaches but also highlight gaps for future research. It showcases the foundational mechanisms of the field and therefore can be used as a foundation to then derive practical sustainability strategies for GenAI projects.

The third section explores the practical application of the framework and by this addresses research question 4. This section begins with field research, detached from the concrete framework. The aim of this field research is the characterization of the practical reality in a corporate context from the perspective of introducing a sustainability strategy to this reality. Both the characterization as well as the previously developed framework served as the inputs to derive a first, context specific blueprint for sustainability measures. Based on this blueprint, stakeholder interviews were conducted, to generate feedback and insights on the blueprint. From these insights, a revised practical blueprint was created, serving as the third deliverable. This practical blueprint demonstrates how the framework can be translated into practical action. The blueprint is a governance blueprint designed for a concrete case and setting, in a large international professional services company. It serves as an explorative example that demonstrates the applicability of the framework.

Background

3.1 Generative Artificial Intelligence in the Enterprise

Introduction

The rapid adoption of GenAI in the business landscapes occurs across industries. While sheer endless opportunities for use cases are identified and revolutionary transformations prophesied, so are risks and barriers emerging. This chapter explores the business implications of GenAI by reviewing relevant academic literature and industry reports. The aim is to provide an overview of the GenAI adaptation in the business landscape and the projected impacts. Further we explore the drivers and barriers influencing this uptake and the emerging risks.

Method

The literature was searched on Google Scholar using the following search string: „business“ AND („generative AI“ OR „GenAI“ OR „generative artificial intelligence“). The search criteria were focused on identifying papers that directly address the impact of Generative AI across multiple industries and the business landscape as a whole. To complement the research papers, industry reports were consulted for quantitative forecasts and projections related to the adoption of GenAI in business contexts.

Only works published between 2023 and before December 2024 were included, with 2023 marking a pivotal year in the adoption of GenAI in business organizations (Chui, Hall, et al., 2023).

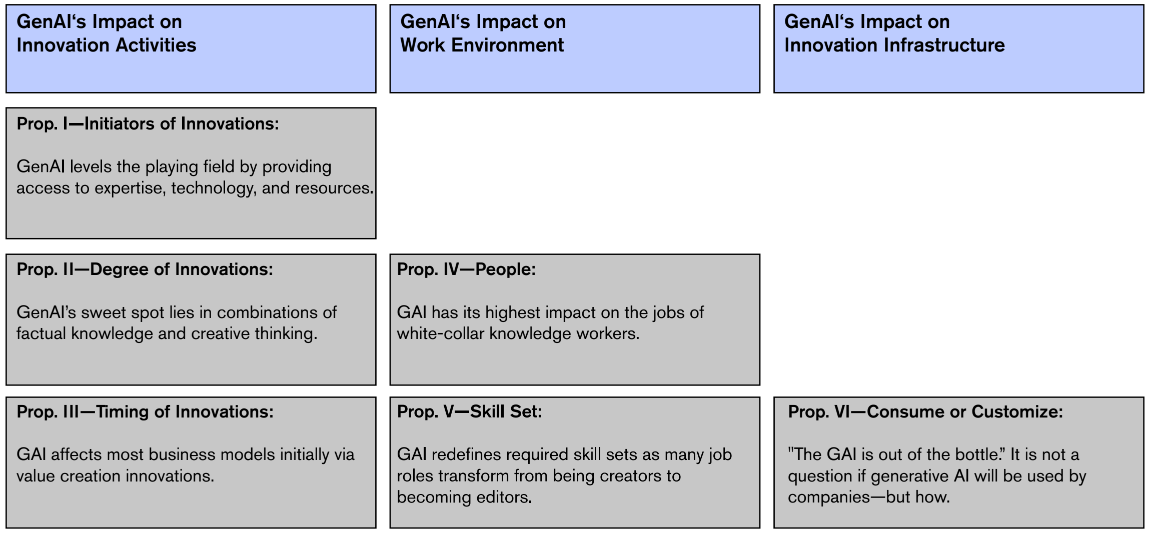
To put abstract values into perspective, equivalents were found in the context of the Netherlands. These examples were referenced from Statistics Netherlands (CBS).

Results

The ability of generative artificial intelligence to effectively address business tasks resulted in a high interest in the technology by the business world. Capable of producing creative outputs such as text, images, code, and simulations, GenAI shifts traditional paradigms of automation, creativity, and decision-making. Businesses across industries leverage GenAI to optimize operations, enhance customer experience, and develop innovative products and services. It is prognosed, that GenAI will be applied across all industries for a high variety of tasks (Chui, Hazan, et al., 2023). The impact of GenAI on the business landscape will be tremendous. Goldman Sachs (2023) predict a 7% increase of the global GDP via GenAI and Chui, Hazan, et al. (2023) prognose, that GenAI could add an equivalent of 2,6 – 4,6 trillion USD to the global economy annually. For comparison, the annual GDP of the Netherlands was around 1,1 trillion is 2023 (CBS, 2024). In a study conducted by Boston Consulting Group (2024) 85% of executives worldwide plan on increasing their AI/GenAI investments, showcasing, that GenAI does not only have the potential to transform businesses globally, but that this development is foreseeable.

The most central business functions which are affected by GenAI are Marketing and Sales, Customer Operations, Product R&D, Software engineering and Supply Chain and operations (Chui, Hazan, et al., 2023). Business value that is generated in these fields is among other factors the improved quality of content, competitive advantage, scaling employee expertise, expanding organizational capabilities, improvement in sales and revenues and enhanced customer experiences (Dencik et al., 2023). This value is generated through the activities that can be supported by the technology. Activities that GenAI can perform at the example of marketing and sales are the gathering of market trends and customer information from unstructured data; Drafting marketing and sales communications; Customizing marketing efforts to customer needs (e.g. language, demographic); Additional content for better informing purchasing decisions, such as virtual “try-ons” just to name a few (Chui, Hazan, et al., 2023). The possibilities of applying the technology seem sheer endless.

Besides these applied examples, GenAI also exerts fundamental impact on business transformation overall. Kanbach et al. (2024) present six propositions for GenAI’s impact on businesses divided into impacts on innovation activities, impacts on work environment and impacts on information infrastructure (see Figure 2). These fundamental impacts manifest themselves in measurable advantages. Boston Consulting Group (2024) measures, that utilizing GenAI in everyday tasks increases productivity by 10-20 % and reshaping critical business functions with GenAI results in efficiency gains of up to 50%. This demonstrates, that to remain competitive, many businesses will find no way around the uptake of GenAI.



GenAI‘s Impact on Innovation Activities	GenAI‘s Impact on Work Environment	GenAI‘s Impact on Innovation Infrastructure
Prop. I—Initiators of Innovations: GenAI levels the playing field by providing access to expertise, technology, and resources.		
Prop. II—Degree of Innovations: GenAI’s sweet spot lies in combinations of factual knowledge and creative thinking.	**Prop. IV—People:** GAI has its highest impact on the jobs of white-collar knowledge workers.	
Prop. III—Timing of Innovations: GAI affects most business models initially via value creation innovations.	**Prop. V—Skill Set:** GAI redefines required skill sets as many job roles transform from being creators to becoming editors.	**Prop. VI—Consume or Customize:** "The GAI is out of the bottle.” It is not a question if generative AI will be used by companies—but how.

Figure 2: Propositions for GenAI‘s impact on business
Adapted from Kanbach et a. (2024)

But there are also roadblocks: A survey by Dencik et al. (2023) revealed, that a majority of executives have concerns regarding cybersecurity and privacy when it comes to GenAI implementation in business organizations, additionally a significant portion sees inadequate financial justification and considers the required investment too high. Further barriers are concerns about data accuracy and bias, regulations and compliance and others (Dencik et al., 2023). In a survey from McKinsey and Company with 913 companies which have adopted GenAI, only 11 organizations consider the environmental impact of GenAI a risk and only 5 companies work actively on mitigating said risk (Chui, Hall, et al., 2023). While these hurdles are slowing down company-wide adoption, the advantages from a business standpoint are significant.

Discussion

While there is a strong likelihood that GenAI will experience widespread adoption across many industries, the timing and extent of this adoption are still subject to the resolution of several challenges. Businesses will need to navigate ethical, technical, and regulatory hurdles, as well as invest in reskilling their workforces. While the insinuation of a lack of reflection and one-dimensionality might be appropriate regarding the propositions prognosed to the impacts of GenAI, like the one presented from Kanbach et al. (2024), the existence of business value through GenAI is apparent. The widespread adoption of GenAI seems probable in the medium to long term, but it may not be uniform across all sectors or regions. While these hurdles pose significant challenges, the transformative capabilities of the technology are undeniable. Ultimately, the question regarding the adoption of GenAI in business organizations is not if, but when, where, and how. The prognosed impact by GenAI on the business-landscape would require a drastic scale up of the technology and its corresponding infrastructure. With a majority of companies nowadays pursuing sustainability goals, it raises questions as to why such few organizations consider environmental sustainability as a concern for implementing GenAI.

3.2 Environmental Impacts of Generative AI

Introduction

The widespread integration of Generative AI (GenAI) into business operations, alongside its increasing presence in everyday life, is not without consequences for the environment.

Beyond the general societal interest in a sustainable future, many companies are now incorporating environmental considerations into their corporate responsibility strategies. According to KPMG (2022), 96% of G250 companies report on sustainability, but 30% are falling short of their Scope 1 and 2 emission targets, with nearly half failing to meet their Scope 3 goals (Bain & Co., 2024). As the adoption of GenAI in business continues to grow, it poses a risk of undermining corporate sustainability efforts, potentially hindering companies‘ ability to achieve their environmental objectives.

The aim of this chapter is the identification of the different dimensions in which GenAI exerts impacts on the environment. Further than that, quantitative estimations for a selection of dimensions have been consulted to understand the scale of that environmental impact across different lifecycle stages. The goal is to understand the nature, magnitude, and location of the environmental effects associated with GenAI, making the case for why companies should be concerned about the environmental ramifications of GenAI use.

Method

The literature was searched on Google Scholar, using the following search string: („Environmental Sustainability“ OR „Sustainability“ OR „Environmental Impacts“) AND („generative AI“ OR „GenAI“ OR „generative artificial intelligence“ OR „AI“ OR „Artificial Intelligence“).

As the search was divided in two areas - a qualitative mapping of the environmental impacts and a quantitative analysis of those impacts - two different selection criteria lists were used. The first area included works that address the different categories and dimensions of environmental impact by AI and GenAI on a general level, specifically those presenting an overview of multiple dimensions. The second area included works that quantify the environmental impacts of GenAI across one or more of the identified dimensions, for more than one lifecycle stage. For both areas, only papers between 2018 and December 2024 have been selected, as December 2024 marks the current time of writing and 2018 marking the release of GPT-1 (Radford et al., 2018) and BERT (Devlin et al., 2018) which are considered the early milestones of GenAI.

From these papers, snowballing was conducted to find works, which explain underlying frameworks and concepts, in order to provide novice readers with enough contextual information to grasp the presented content.

Additionally, abstract values and scales were put into a more understandable context by referencing examples or equivalents in the Netherlands. These examples were referenced from news articles and from Statistics Netherlands (CBS).

As the goal of this chapter is not an exhaustive mapping of the research landscape, rather than an exploration and overview of a specific concept, the search was stopped when sufficient material has been found to allow for an exhaustive overview of the addressed concepts.

3.2.1 The dimensions of environmental impact

The vast infrastructure required to develop and run GenAI systems affects the earth’s environment across a multitude of dimensions, beyond just the emissions from energy consumption.

The stability and resilience of the Earth system as a whole is upheld by nine critical processes (Richardson et al., 2023). The widely established planetary boundary framework (PB) aims at delignating these processes and making humanities impact upon each measurable (Rockström et al., 2009). All of boundaries have been disturbed by human activity, with six out of the 9 being transgressed, meaning the current status of these is not within safe limits (Richardson et al., 2023). At the current point in time, it is therefore of existential importance to both bring the transgressed planetary boundaries back into safe limits and change the trajectory of those which are expected to be transgressed. This applies to all human activities which enact influence on the nine processes.

Planetary Boundary	Parameter	Status
Climate Change	Atmospheric CO2 concentration	transgressed
Freshwater Change	Blue & Green water depletion	transgressed
Biosphere Integrity	Genetic diversity	transgressed
Land-System Change	Area of forested land	transgressed
Ocean Acidification	Carbonate ion concentration	not transgressed
Novel Entities	% of synthetic chemicals released into the environment without adequate safety testing	transgressed

Table 1: Affected Planetary Boundaries, adapted from Falk et al. (2024) and Richardson et al. (2023)

Research from Falk et al. (2024) has shown, that the development and deployment of artificial intelligence negatively impacts six out of the nine boundaries, with five of them being already transgressed and one being soon to be transgressed (see table 1). While the majority of research is focused on energy related emissions (Brevini, 2020; Crawford & Joler, 2018; Falk & van Wynsberghe, 2023; Kaack et al., 2022; Ligozat et al., 2022), environmental consequences across all of the six identified dimensions occur and therefore need to be considered (Falk et al., 2024).

Note, that the planetary boundary framework does not directly capture the depletion of abiotic resources (metals, minerals and fossil-based resources), as this variable - on its own - cannot destabilize the earth system and only has socio-economic consequences (Paulillo & Sanyé-Mengual, 2024). However, the activities that come with the extraction of abiotic resources, come with environmental costs that affect other PBs, such as land system change or climate change (Paulillo & Sanyé-Mengual, 2024). Therefore it is not listed as an additional dimension, while it is explored later on.

Climate Change & Ocean Acidification:
Climate change and ocean acidification are affected by energy related carbon emissions. These emissions stem mostly from the mining of resources for hardware, the production of hardware systems, running the systems, e.g. in training and inference and lastly disposing the systems. To understand the scale, GPT-3 (175B) is estimated to have contributed 502 metric tons of CO2 equivalent, Gopher (280B) 352 tons and Llama 2 (70B) 291,42 tons (Stanford University, 2024), just to name a few. For comparison, 502 metric tons of CO2 equivalent equates to taking 213 round-trip flights between New York and Singapore (International Energy Agency et al., 2023).

Freshwater Use:
Vast amounts of freshwater is used in the manufacturing and cooling of datacenters (Falk et al., 2024). A small, 1-MW datacenter alone, can consume around 25.5 million liters of water per year (Mytton, 2021). For comparison, in 2022 the global hyperscale data center capability was estimated to reach 13177 MW (Structure Research, 2023) with the trend continuously growing (Synergy Research Group, 2023). Additionally, the production of hardware systems, such as chips, is resulting in overexploitations of freshwater resources, for example in Taiwan where the world's largest microchip producer relies on regional water basins (Roussilhe et al., 2024). Further, water is used in the production of energy that powers the system. While some renewable energy like solar or wind energy require little water, others such as nuclear power and biofuels consume a lot (Falk et al., 2024).

Biosphere Integrity:
The intensive land developments that occur with the construction of the hardware impacts, divides and destroys natural habitats (Falk et al., 2024). The land use of datacenters is vast. For example, a single data center planned by Meta in the town of Zeewolde in the Netherlands would have consumed farmland with the length of 245 American Football fields (Brown Hamilton, 2022). This plan was canceled due to concerns about the effects of such a facility on the natural environment, but many others are being constructed, such as the Meta data center in Altoona, Iowa taking up over 114 acres of land (City of Altoona Iowa, 2021). Besides the land development, the intense water-use further affects the biosphere integrity by lowering ground water and reducing surface water (Falk et al., 2024).

Land-System Change:
The large-scale extraction of resources such as aluminum, steel, copper, indium and lithium which are required for the production of the GenAI infrastructure results in the reduction of forested land, effecting the boundary land-system change (Falk et al., 2024).

Novel Entities:
The resource extraction and manufacturing of hardware systems as well as the frequently occurring unproper disposal of these systems results in the release of toxic and hazardous elements into the environment without adequate safety measures (Falk et al., 2024).

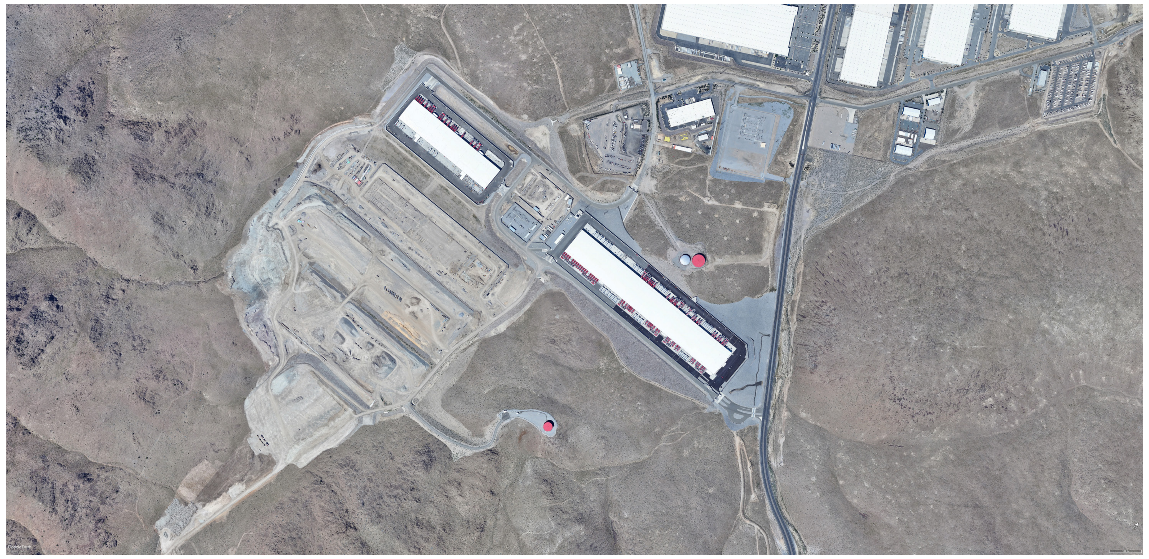


Figure 3: Construction of the Citadel Campus Data Centers
Source: Google Earth Pro (2023)

3.2.2 The quantification of environmental impacts to the GenAI lifecycle

Specifying the environmental impacts of the lifecycle stages of GenAI depends on the model, context and setup. It is therefore not possible to assign specific impacts from a generalized perspective. Various assessments have been reported that assign the environmental impacts to AI lifecycle stages (Berthelot et al., 2024; P. Li et al., 2023; OECD, 2022; Strubell et al., 2019; Wu et al., 2021). This chapter specifically focuses on the environmental impacts of GenAI, as the emission of other ML types differ significantly (Luccioni & Hernandez-Garcia, 2023), due to the drastically larger resource consumption of most GenAI models.

Berthelot et al. (2024) conducted a life cycle assessment (LCA) for the GenAI model Stable Diffusion, a text-to-image generative deep learning model accessible online. The stages considered in the LCA are model training, inference, hosting the service on the web, the network communications and the end user terminals utilized for accessing the model. The LCA was conducted for two functional units: The first one, FU1 representing the average impact of a single user visiting the website and submitting one prompt leading to the generation of four images; The second one, FU2 accounting for the cost of service of one year overall, over versions v1-4 (2 months) and v1-5 (10 months). From the perspective of businesses hosting GenAI platforms, FU2 is a more representable unit as it is from the hosts perspective. Therefore, the results for FU2 are of higher relevance for this work. In FU2, the training of v1-4 and v1-5 is based on the training of v1-0, v1-1 and v1-2. Because of that, the LCA conducted by Berthelot et al. (2024), presents a full legacy scenario containing also the training of the variants v1-0 until v1-2. For the full legacy scenario, the LCA analyses the abiotic depletion potential (kgSb eq), the global warming potential (kgCO2 eq) and the primary energy use (MJ). The result of the LCA (see Figure 4) shows, that End user Terminals contribute the most to both abiotic depletion and global warming potential. Networks and Training contribute similarly to the abiotic depletion potential followed by Webhosting and lastly the inference. The inference has the second largest contribution to the global warming potential, followed by Training and lastly networks and webhosting. Further it can be seen, that Inference draws the largest amount of energy followed by the End user Terminals, Training, Networks and lastly Webhosting.

The large impact of the End user terminal can be explained by the explored model (Stable Diffusion) which is accessed by a vast amount of End user terminals, due to the large number of users. While both the Webhosting as well as the End user terminals are not part of the GenAI model itself, it becomes clear that the adoption of GenAI contributes the growth of IT usage and its footprint (Berthelot et al., 2024). For the GenAI system itself, the results outline, that training, inference and network communications come with significant impacts. Overall, the climate change potential of the FU2 is 360000 kgCO2 eq, equating to 153 round-trip flights between New York and Singapore (International Energy Agency et al., 2023).

Other studies such as research by P. Li et al. (2023) also consider the training and inference as the key stages from the GenAI lifecycle in which environmental impacts occur. P. Li et al. (2023) estimate the water consumption footprint of the training and inferences of GPT-3 based on various locations. They estimate, that training GPT-3 in the Netherland would have consumed 5,237 million liters of water. For comparison this equates to over 111 years of water use by the average Dutch citizen (CBS, 2023). Every 61 inferences – if computed in the Netherlands – is estimated to consumer around 1 liter of water (P. Li et al., 2023). With ChatGPT, a tool based on a GPT model, scoring over 650 million users in just a single month of use (Semrush, 2024), it becomes clear how significant the water use of GenAI is.

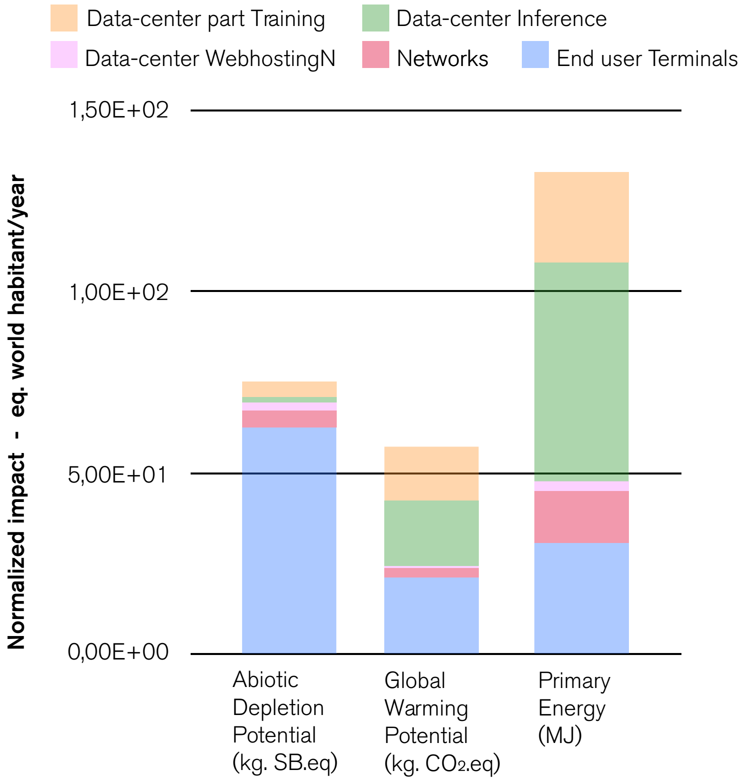


Figure 4: Full legacy scenario impact of Stable Diffusion Adapted from Berthelot et al., 2024

Discussion

The increasing uptake and prognosed scaling of GenAI doesn't come without consequences. Research shows, that a scale-up of the technology exerts vast impacts across our environment. While the problem has been identified in academia, the general understanding and awareness remains limited. To this date, the research landscape around the sustainability of AI focuses mostly on the energy efficiency, with little attention being paid across other environmental dimensions. With GenAI exerting negative influence on the majority of earth systems it becomes clear, that a focus solely on energy consumption and the resulting carbon emissions is insufficient. Instead, I align with (Falk et al., 2024) in that the planetary boundary framework is an appropriate framework in attributing AI's environmental impact over a broader set of dimensions. While the PB framework does not directly address certain dimensions that could be considered environmental impacts, such as abiotic depletion potential, it encompasses related effects indirectly through categories like land-system change and novel entities. Because these impacts are captures indirectly and the PB being a widely established and used framework, I advocate for its use in this context.

While the two presented quantitative studies cannot be generalized to all types of GenAI applications, the quantifications outline how significant the environmental costs can be. Especially the phases of training and inference contribute to the environmental impacts.

As discussed earlier, GenAI is expected to be scaled up further with an increasingly large infrastructure. This underlines the relevance of addressing environmental impacts to mitigate further damages and the contribution of GenAI to the destabilization of the Earth System.

It is therefore part of corporate responsibility to consider the environmental consequences of taking up and scaling GenAI applications in businesses.

3.3 The field of GreenAI

Introduction

Green AI is an emerging field of scientific research in response to growing concerns about the environmental impact and inclusivity of AI technologies (Schwartz et al., 2020). It is in contrast to Red AI which describes AI research aiming to improve accuracy through vast computational power. In the field of Red AI, linear performance gains are achieved via exponentially growing models, a development that is inherently unsustainable (Schwartz et al., 2020). As a relative to the field of sustainable software engineering Green AI addresses environmental concerns of AI. Researchers and industry professionals are beginning to propose strategies aimed at reducing the ecological footprint of these technologies through sustainability practices. This chapter provides an overview of the current state of research in Green AI. In addition to mapping the research landscape, the chapter seeks to identify research gaps and potential areas for improvement in the development of more sustainable AI systems.

Method

The literature was searched on Google Scholar with the following search terms: („Green AI“) OR („Environmental Sustainability“ OR „Sustainability“) AND („generative AI“ OR „GenAI“ OR „generative artificial intelligence“). The search focused on identifying papers that address approaches to improving the environmental impact of AI and GenAI. Studies were selected based on their relevance to strategies aimed at mitigating the environmental footprint of these technologies. Specifically, only papers that provided insights into means towards sustainability improvements in GenAI were included in the review. Papers that did not address these aspects were excluded from consideration. Only papers between 2018 and December 2024 have been selected, as December 2024 marks the current time of writing and 2018 marking the release of early milestones of GenAI.

Results

As the environmental footprint of AI technologies is gaining increasing attention in academia and industries the new field of “Green AI” is emerging. Green AI is positioned at the intersection of sustainable software engineering and AI engineering (Yarally et al., 2023). Most literature in the field of Green AI focusses on the energy efficiency, with only a few addressing the carbon and overall ecological footprint (Verdecchia et al., 2023). In their systemic literature review of the field, Verdecchia et al. (2023) identified that out of 110 publications, 81 publications defined Green AI as addressing the energy efficiency of AI with only 9 publications considering the overall ecological footprint of the technology.

The key topics identified in the publications studied by Verdecchia et al. (2023) heavily focus on the technical set-up of AI systems with little attention being paid to the surrounding context and application. Specific attention is paid to the training phase with a majority of all papers studied in mentioned review exclusively considering the training phase (Verdecchia et al., 2023). The most frequently addressed topics (with mentions in 10 or more publications) are footprint monitoring, hyperparameter tuning, model benchmarking, deployment, precision-energy trade-off and algorithm design. All of these falls under the technical set up, with contextual topics such as policy or ethics are only mentioned three times each in the reviewed literature (Verdecchia et al., 2023).

Within the domain, solutions are proposed to develop greener AI. Such solutions are for example the use of adaptive backpropagation in fine-tuning LLMs (Huang et al., 2023), research on the energy efficiency of hyperparameter optimization strategies in deep learning training (Schwartz et al., 2020; Yarally et al., 2023) green federated learning (Thakur et al., 2024) and research and optimization strategies for energy efficient neural network architecture (Schwartz et al., 2020; Yarally et al., 2023). Little solutions are proposed beyond those that make the systems more efficient. More efficient models demand less resources for a similar or better result than less efficient models, but it doesn't stop there. Historic developments have shown, that with increasing efficiency of a technology, the resource demand of that technology often also increased due to the growing volume of consumption – a phenomenon called the Jevons' paradox (Shumskaia, 2022). This phenomenon can be observed with AI. The efficiency and performance have fundamentally improved, but so did the overall resource consumption as it empowered larger model sizes and an increased consumption volume. While in the past, most natural language processing tasks (NLP) could be developed and trained on laptops and servers, they now require vast computing resources, such as the hyperscale datacenters discussed earlier (Shumskaia, 2022). While advancements in efficiency are generally beneficial for the resource demand, they need to be complemented with other strategies, such as sufficiency strategies.

One such sufficiency strategy presented is the benefit-cost evaluation framework for GenAI applications by Bashir et al. (2024). The framework offers means to evaluate the environmental benefits and costs of utilizing GenAI applications instead of a baseline - the business-as-usual way of solving the task at hand. Utilizing this framework, one can consider whether the use of GenAI for a particular use case is justifiable in relation to the resulting environmental consequences (Bashir et al., 2024). Little attention has been paid to solutions beyond efficiency improvements.

The authors of works in the domain of Green AI are to a large proportion from an academic background (Verdecchia et al., 2023). Out of the works reviewed by Verdecchia et al. (2023) 75 papers were written in a purely academic context, with only 3 being exclusively written by industrial authors. With the model developments frequently demanding vast resources, many experiments have been shown to be far beyond the reach of academic researchers (Schwartz et al., 2020). For example, researchers by DeepMind evaluated 1500 hyperparameter settings for demonstrating the potential of their model and researchers at Google trained over 12800 neural networks in a neural architecture search for improving the performance; while demonstrating the potential of their models, a full and fair comparison to competitive models or comparable research in an academic context would be unaffordably expensive (Schwartz et al., 2020).

Discussion

The current research landscape of Green AI pays little attention to environmental impacts of GenAI beyond its energy use. This underrepresentation of other impacts, such as its biotic resource, water and land-use provides a research gap to be filled. It has to be noted, that a single solution strategy can address multiple environmental impacts. For example, reducing the computational load of training a model on a GPU reduces the model's occupation time of that GPU. Therefore it demands less energy and at the same time accounts for a smaller proportion of the physical resources required to produce and run the GPU – such as metals and water. Nonetheless, research shows that impacts beyond energy use are significant, highlighting a relevant research gap that needs to be addressed, as energy efficiency solutions alone are insufficient to ensure environmental sustainability.

Furthermore, the current research leans towards focusing on technological development. Other lifecycle phases such as documentation or the business understanding underlying the model deployment are rarely addressed. This research gap presents room for opportunities in which novel or refined sustainability strategies might offer leverage.

Besides the addressed impacts and lifecycle phases, also the type of presented solutions strategies display research gaps. At this point in time, the nature of strategies presented leans heavily towards efficiency strategies. While valuable, other strategies remain largely unaddressed. This may be because efficiency strategies also provide economic benefits to the industry, whereas for example sufficiency strategies potentially reduce the demand for GenAI. Additionally, experts in the field may focus on efficiency improvements, as these are already deeply ingrained considerations in software development for economic and feasibility reasons. Different strategy types beyond efficiency strategies therefore present research gaps. As efficiency strategies on their own only have a limited ability in building sustainable AI, I advocate for the development, consideration and combination of a diverse set of solution types.

With limited contributions from industry and insufficient resources for large-scale experiments in academia, the development of sustainability strategies for AI remains significantly constrained. A possible explanation for the varying level of interest in academia and the industry are differences in worldviews. While academia might tend towards a social constructivist worldview, industry players are likely more inclined to adopt a technologically deterministic perspective. An explanation for that is the neutrality in academia which supports the idea that technological developments should ultimately benefit the greater good and be shaped towards that. Contrary, industry players are likely biased towards improving their position in a competitive environment, making their technological advancements a tool to stay ahead of competition, which incentivizes the creation of larger and more capable models (Red AI). Therefore, only a limited variation of potential development trajectories is considered, at the expense of sustainability improvements. This contrast creates the need for closer collaboration and awareness creation across the different domains.
I therefore advocate for the cooperation of academic researchers with industry for developing practical solutions and awareness creation.

Conceptual Framework

4.1 The Generative AI Lifecycle

Introduction

The first step in developing the framework is to map the overall lifecycle of a GenAI model and its associated stages. A lifecycle model makes the various actions surrounding the model visible and structured, allowing them to be systematically addressed through the framework. In this context, the purpose of the lifecycle model is not to provide an accurate representation of real-world development processes - which are often highly iterative and non-linear - but rather to offer an exhaustive categorization of the different phases that can occur. The occurring actions should be grouped into distinct phases that can be addressed independently within the framework.

Translating the technological mechanisms of GenAI to a real life use case is a multi-stage process. These stages make up the different phases of the GenAI lifecycle. Strategies, measures or approaches targeting GenAI applications inherently address specific phases within this lifecycle. Understanding these stages is crucial for assessing where and how changes can be made.

The goal of this chapter is therefore the presentation of a lifecycle framework, with the aim of creating an understanding about the sequence and nature of the phases on a general level. While real world cases likely differ from each other to some extent, to goal is not the creation of a precise representation of a real life lifecycle, but rather an abstraction that allows for a categorization and an understanding of the generalized process flow.

The resulting phases and associated actions are then utilized to understand which sustainability mechanisms are required in the framework, in order to address the entirety of the GenAI lifecycle.

Method

The literature was searched on Google Scholar using the following search string: („lifecycle“ OR “life cycle” OR “life-cycle”) AND („generative AI“ OR „GenAI“ OR „generative artificial intelligence“ OR “AI” OR “Artificial Intelligence”).

From these papers, snowballing was conducted to find works, which explain underlying concepts, in order to provide novice readers with enough contextual information to grasp the presented content.

As little suitable literature was found in an academic context, a Google search was conducted for the same word search. The background of the authors was checked based on their industry experience and suitable publications chosen.

Only works between 2018 and December 2024 have been selected, as December 2024 marks the current time of writing and 2018 marking the release of GPT-1 (Radford et al., 2018) and BERT (Devlin et al., 2018) which are considered the early milestones of GenAI.

As the proposal and development of a new, scientifically proven, lifecycle framework is outside of the scope of this work, a preexisting framework for the overarching domain of ML was chosen from the scientific literature, founded on its widespread adoption in the field. Based on this, the characteristics of GenAI were specified, without changing the overall steps of the framework. The characteristics were derived from a non-scientific GenAI lifecycle framework, proposed by an author with significant industry expertise. All specifications added are marked as such.

4.1.1 Overview

Little scientific research has been done on the specific lifecycle of GenAI models at this point in time. GenAI falls under the category of ML, a field in which scientific studies have been conducted regarding lifecycle frameworks. Therefore I propose the specification of a scientific ML lifecycle framework to the characteristics of GenAI. Haakman et al. (2021) propose a series of ML lifecycle frameworks, among others the Cross-Industry Standard Process for Data Mining (CRISP-DM) which originated from Shearer C. (2000). The CRISP-DM process remains one of the most widely adopted frameworks among data scientists today (Saltz, 2024a). It provides a generalist structure that is adaptable and emphasizes business context, making it a suitable choice for this case. As shown in Figure 5, the CRISP-DM framework by Haakman et al. (2021) revised for ML, separates the overall lifecycle into nine key stages: Business Understanding, Data Collection, Data Understanding, Data Preparation, Modeling, Documentation, Evaluation & Risk Assessment, Deployment and Model Monitoring. It must be noted that the adapted CRISP-DM model for ML also applies to GenAI, as GenAI is a subset of ML. Nonetheless, the lifecycle of GenAI comes with specific characteristics and phases compared to other ML applications, which are not captured by the more generalist CRISP-DM model proposed by Haakman et al. (2021).
While to my best knowledge, no scientifically proven GenAI-specific lifecycle frameworks exist, such frameworks have been published in other media (Arsanjani, 2023; Microsoft Cooperation, 2024; Saltz, 2024b). The director of applied AI engineering at Google, Ali Arsanjani (2023) proposes a GenAI-specific lifecycle. While the layout and granularity differ, the key phases presented in his proposal can be found in the adapted CRISP-DM framework (Haakman et al., 2021).

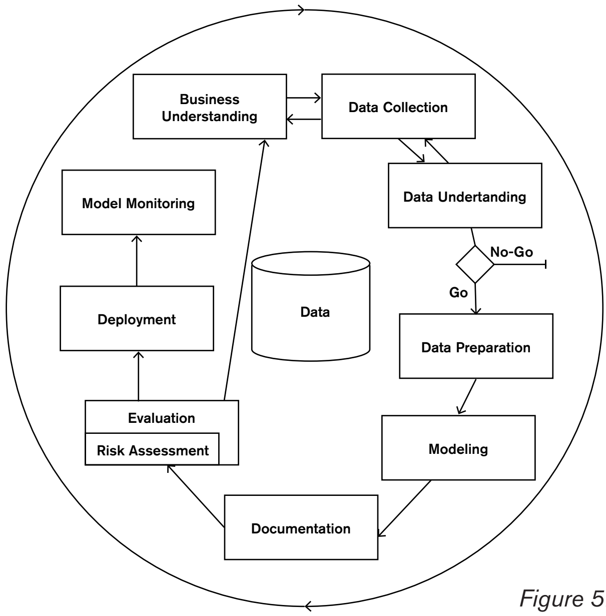


Figure 5: Adapted CISP-DM for ML Based on Haakman et al. (2021)

4.1.2 An adapted CRISP-DM framework for GenAI

I propose an adapted CRISP-DM framework based on the Generative AI Lifecycle as presented by (Arsanjani, 2023).

Data: The heart of the CRISP-DM framework is the data, which is required to successfully run the overall project. While both Haakman et al. (2021) and Shearer C. (2000) don’t further describe the type of data required, Arsanjani (2023) describes three categories of data in the GenAI lifecycle: Data which is used as the basis for training the foundational model, domain-specific data which is used for adapting the model for use-case specific tasks and input data, which is data that is fed into the model during inference.

Business Understanding: This phase, as described by Shearer C. (2000), serves as the initial phase of the project. The goal is the understanding of the objectives from a business perspective, meaning understanding the background, the business objectives and the business success criteria. From there on, the situation is assessed: The Inventory of resources is laid out; Requirements, assumptions and constraints are formulated; Risks and contingencies are identified; Terminology is defined; Costs and benefits of the project are analyzed. After that the technical goals and success criteria are formulated and lastly a project plan is produced (Shearer C., 2000). While the framework

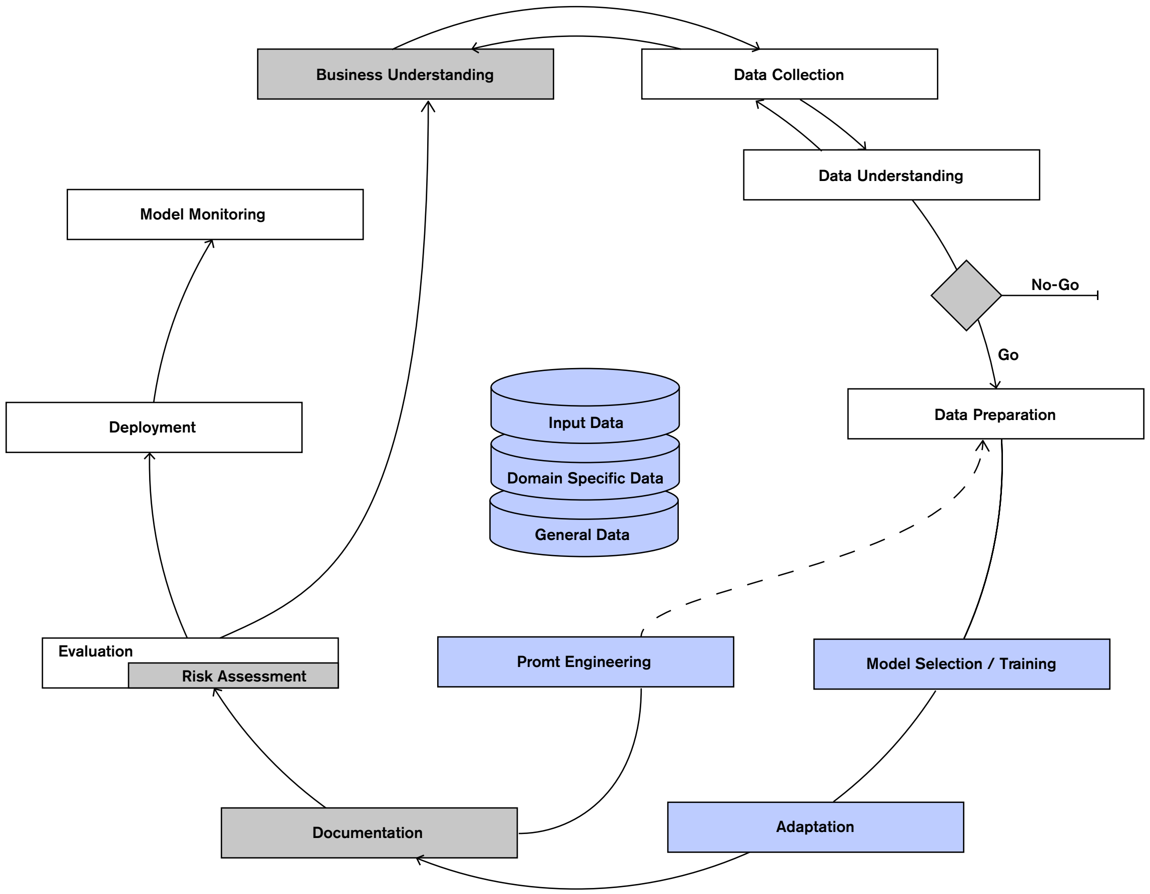


Figure 6: Adapted CISP-DM for GenAI
Based on Haakman et al. (2021) & Arsanjani, (2023)

of Arsanjani (2023) does not explicitly consider the organizational context of the GenAI lifecycle, it does propose a phase of problem formulation and preliminary research that broadly encompasses similar activities. As the key steps of the CRISP-DM framework have been proven to be relevant, and the focus of this thesis being GenAI in the context of businesses, this step remains part of newly adapted CRISP-DM framework.

Data Collection: This step is newly proposed by Haakman et al. (2021), as it is of significant relevance for ML projects. Data of different criticality levels must be accessed and the accurate representation by the collected data of the problem in question ensured. This step is also found in the framework of Arsanjani (2023). Additionally, it is highlighted there, that besides being representative of the problem at hand, the collected data shall further represent a diverse range of perspectives, backgrounds and sources. In some cases, Arsanjani (2023) proposes the generation and use of synthetic data, in case this improves the desired performance metrics.

Data Understanding: The goals of this step is the creation of understanding about the data, as well as ensuring that the data meets the desired quality level via a data validation process (Shearer C., 2000). This contains reporting a data description, exploration and quality. While depending on the data the depth of this step might vary, Arsanjani (2023) also highlights the need to validate the data after collection.

Go-/No-Go Point or Feasibility study: After the data understanding phase, it is essential to check, whether the concept of the project is able to deliver long-term expectations (Haakman et al., 2021) If not, the project should be discarded. If it can be assumed at this stage, that the project is feasible and able to meet all demands, the next stage is entered. Note, that the overview by Arsanjani (2023) does not specify this step. As this step can propose the end of the project before GenAI would be implemented, it might not be understood as a part of the GenAI lifecycle. Nonetheless research proofed the relevance of this step (Haakman et al., 2021).

Data Preparation: The step of data preparation entails the preparation of data for its use in the ML model. This consists of selecting the suitable data, cleaning the data and processing the data to arrive at standardized, merged and curated data sets. This step is both described by Shearer C. (2000) and (Arsanjani, 2023).

Modeling: The modeling phase as described in the CRISP-DM model contains the steps from selecting the model, creating a test design, building the model and assessing it (Haakman et al., 2021; Shearer C., 2000). This step entails all development required before the model evaluation followed by the deployment. Arsanjani (2023) splits this step into three core parts: The selection/ training of the model, the adaptation of the trained model to a specific domain and the prompt engineering.
Selection/ training of model: This step aligns with the traditional understanding of "modeling" in the CRISP-DM model. The model is selected and trained with the data prepared for this task and assessed based on its performance.

Adaptation: Here a pretrained foundational model is adapted, based on domain-specific data and fine-tuned to perform a specific task based on the new data provided. This adaption allows not only the finetuning of a foundational model, but also distilling the foundational model to maintain the functionality required for the specific problem, while downsizing the model for improved usability or efficiency (Arsanjani, 2023).

Prompt Engineering: Prompt Engineering takes a key role in the use of GenAI, as different prompts will result in different outcomes. In this stage, the prompts are designed, generated, tested and iterated to achieve the desired output. Arsanjani (2023) proposes, that this stage can even occur before Data Preparation, to focus on the relevant data for the task and overall streamline the development process.

Due to the specific characteristics and the importance of the different sub steps that fall under the step “modeling”, the substeps are explicitly stated in the adapted CRISP-DM model (see Figure 6).

Documentation: The documentation stage as proposed by Haakman et al. (2021) is the phase in which the structure and related activities of the model are documented. This entails the design, evaluation, testing and so forth. The documentation of a model extends beyond just the code and entails the full system. It serves as a knowledge base to evaluate, maintain, debug and track all decisions in relation to the model. While this step is not explicitly part of the GenAI lifecycle as proposed by Arsanjani (2023), it’s benefit for ML application has been demonstrated by Haakman et al. (2021).

Evaluation & Risk Assessment: In the evaluation stage, the overalls model’s performance is evaluated regarding the criteria defined in the “Business Understanding” phase of the lifecycle and associated to the model are assessed. Specifically, Arsanjani (2023) highlights the assessment of the model fairness and bias when working with GenAI. This phase is present in both the CRISP-DM framework from Haakman et al. (2021) and the GenAI lifecycle from Arsanjani (2023).

Deployment: In this phase, the deployment is executed. This entails the planning of the deployment, the preparation of the model for deployment (e.g. connecting it to an endpoint and testing that endpoint) and the actual deployment (Arsanjani, 2023; Haakman et al., 2021; Shearer C., 2000).

Model Monitoring: Once the model is deployed its performance is continuously monitored to ensure, that it is behaving as expected. Arsanjani (2023) highlights for GenAI models, that if changes occur, such as data and model drift, they need to be managed accordingly.

It must be noted that neither the CRISP-DM framework, nor the mentioned lifecycle frameworks of GenAI account for the end-of-life of the models (Arsanjani, 2023; Haakman et al., 2021; Microsoft Coorperation, 2024; Saltz, 2024b; Shearer C., 2000). A reason why the end-of-life might not be further discussed is the recency of the widespread application of these technologies. At the current point of time, this step has not shown to be relevant yet when considering the lifecycle of GenAI, but this can potentially change in the future. Likely after their end-of-use, models hibernate on storage systems until re-used or deleted.

Discussion

Due to the lack of suitable existing suitable scientific GenAI lifecycle frameworks, I propose an adaptation of a lifecycle framework widely established among data scientists.

While CRISP-DM offers a well-established structure for ML projects, its generalist approach benefits from adjustments that address GenAI-specific characteristics. Key enhancements include a clearer emphasis on diverse data types, such as foundational, domain-specific, and input data. Additionally, GenAI’s unique steps like model adaptation and prompt engineering expand the modeling phase beyond traditional ML tasks. The resulting phases provide a holistic overview over the application of GenAI lifecycle in general phases.

It has to be noted, that in real life scenarios, the process is likely to be drastically messier, iterative and not as linear then presented in the adapted CRISP-DM framework. The central difference between real life processes and the abstract CRISP-DM is the presence of iteration and the multitude of interactions between lifecycle stages but also projects entirely. For example, monitoring a deployed model will likely inform the overall business understanding, as learnings are made here that will inform the next model, but the adapted CRISP-DM framework does not showcase an arrow between “Model Monitoring” and “Business Understanding”. The aim of the presented framework is not an accurate depiction of a real life development and implementation process, but rather an abstract depiction of all the overarching stages. The goal is therefore not to map a precise representation of the reality but rather to showcase a selection of lifecycle stages, that effectively classify groups of actions occurring within the lifecycle, and a somewhat generalizable sequence of these stages.

The presented framework does also not consider the actors involved in each phase. Likely each actor is only involved in a certain part of the overall lifecycle. For example, a company might acquire access to a pretrained foundational model and choose to adapt it for its domain specific use case, leaving the actual model development and training stage in the hands of a third party, whilst conducting the adaption stage themselves. While this is not of relevance at this stage, a more focused and customized perspective might need to be chosen when conducting specific case studies.

Note that wording used in the proposed CRISP-DM framework for GenAI might differ from other frameworks in literature. For this project, we will utilize the wording proposed in the sense of the corresponding descriptions of each phase.

The presented CRISP-DM framework for GenAI offers a coarse overview of the stages occurring in a GenAI lifecycle. We will utilize this the proposed phases as potential areas, for the implementation of strategies and approaches further down the line.

4.2 Framework Development

Introduction

In order to uncover the different types of strategies able of addressing the environmental impacts of GenAI, it is helpful to explore the solutions used across other domains. The central environmental impacts - such as material use, energy use and its associated carbon emissions, land use, and the uncontrolled release of toxins occurring in the mining process of materials (Falk et al., 2024)– can be described as consequences derived from the overall resource use of GenAI. The resource demands linked to the lifecycle of artifacts, from sourcing to disposal, pose a significant sustainability challenge for humanity across various domains.

This issue is addressed with the concept of circular economy. Circular economy is a concept, that - while existing for a while - has recently gained a vast amount of traction by policy makers, in academia and the business landscape (Geissdoerfer et al., 2017). The most prominent definition proposed by the Ellen MacArthur Foundation (2013, p.14) is, that the circular economy is “an industrial economy that is restorative or regenerative by intention and design”. In this system, resource input and output of a system is minimized and those resources that do enter the system are aimed to remain in that system. Various strategies have been proposed to increase the circularity of products/systems called the R-strategies (Potting et al., 2017). These strategies are particularly interesting as they are focusing to reduce resource and material consumption of product chains (Potting et al., 2017). While GenAI applications are not products in the traditional sense, the nature of environmental impacts and associated activities bare close resemblance. Different types of R-strategies, focused on physical products have been proposed, modified and adapted over time (Potting et al., 2017).

In this chapter, I therefore propose the development of a comparable set of strategy types, adapted to GenAI applications, based on the preexisting R-strategies.

Method

For finding a suitable and exhaustive set of R-strategies to be used as the baseline for this process, a literature review was conducted. Google Scholar was searched with the following search string: “R” AND “strategies” AND “circular”. Papers were chosen based on their relevance, in presenting configurations of R-strategy types. From there on snowballing was conducted to uncover the publications in which the utilized type of R-strategies was originally proposed.

From the different frameworks found, a suitable R-strategy type of chosen. The suitability was determined by its exhaustiveness and its level of adoption.

After a suitable R-strategy framework was determined, the definition of each strategy was abstracted to its underlying core mechanism. This abstraction was done by generalizing the descriptions – away from their focus on physical products towards a level of artefacts overall.

The resulting selection of abstract strategies is then compared to the actions performed in each of the GenAI lifecycles. If the action allows for the application of the abstracted strategy, then the strategy is relevant for this phase. A strategy might apply to none, one or multiple lifecycle stages. If applicable to none, the strategy is discarded as no use case has been identified in the context of GenAI.
Lastly, the abstracted strategies are fitted to the context. This can include the merging of strategies in the case of overlaps or the renaming/specification of strategies to better fit their new context. This resulted in a selection of key strategies for environmental sustainability across the GenAI lifecycle.

4.2.1 Selection of R-Strategy

Various R-Strategies have been identified (CE & MVO Nederland, 2015; Council for the Environment and Infrastructure Netherlands, 2015; Ellen MacArthur Foundation, 2013; Ranta et al., 2018). With varying levels of granularity, the 10R strategy by Potting et al. (2017) is a widely adopted framework of strategy dimensions. In addition to its widespread adoption, it further boasts a high dimensionality, with 10 strategy categories identified. Therefore the 10R framework by (Potting et al., 2017) will serve as the base for developing GenAI specific strategy categories.

The 10 Rs proposed as seen in table 2, are divided into the strategies R0 - R9. The lower the number, the higher the “level of circularity”, meaning the higher the prognosed impact of the strategy. The first three strategies (R0 Refuse, R1 Rethink, R3 Reduce), have therefore the likely the highest impact and fall under the category “smarter product use and manufacture”, followed by the next five categories (R3 Re-use, R4 Repair, R5 Refurbish, R6 Remanufacture, R7 Repurpose) in the category “extend lifespan of product and its parts” and the last two strategies (R8 Recycle, R9 Recover) in the category “useful application of materials’ (Potting et al., 2017). The naming of both categories and strategies makes their focus on physical product fairly clear, requiring a revision for their application to GenAI.

Category	*Strategy*	*Description*
Smarter product use and manufacture	**R0 Refuse**	Make product redundant by abandoning its function or by offering the same function with a radically different product.
	R1 Rethink	Make product use more intensive (e.g., through sharing products or by putting multi-functional products on the market).
	R2 Reduce	Increase efficiency in product manufacture or use by consuming fewer natural resources and materials.
Extend lifespan of product and its parts	**R3 Re-use**	Re-use by another consumer of a discarded product which is still in good condition and fulfills its original function.
	R4 Repair	Repair and maintenance of a defective product so it can be used with its original function.
	R5 Refurbish	Restore an old product and bring it up to date.
	R6 Remanufacture	Use parts of a discarded product in a new product with the same function.
	R7 Repurpose	Use parts of a discarded product in a new product with the different function.
Useful application of materials	**R8 Recycle**	Process materials to obtain the same (high grade) or lower (low grade) quality.
	R9 Recover	Incineration of materials with energy recovery.

Table 2: 10-R Strategies
Based on Potting et al., 2017

4.2.2 Abstraction of R-Strategy

To abstract each of the ten strategies and uncover its underlying core mechanism, the definitions were divided into their core elements. Further the following concepts were used in the general statement:

Unit: With “unit” the artefact in question is meant. In the context of physical products an example could be one chair or in the context of this work an artefact could be one application of GenAI.

Function: Function describes the task to be performed by the unit. In the example of the chair, the function might be the need for seating, or the need for a decorative object. In the realm of GenAI a function could be for example the automatic generation of damage reports in a production facility.

Functionality: This concept describes the ability of the unit to perform the desired function.

Resources: The resources required to provide the unit.

Specifics: Parts of the definition that clarify the chosen context. As the goal is the generalization, these parts are removed from the definitions.

Further the method of abstracting is showcased at the example of the first R strategy (R0 Refuse):

Original definition:

Make product redundant (core element 1) by abandoning its function (core element 2) or by offering the same function with a radically different product (core element 3).

The original definition can be divided into three core elements. The abandonment of the function the product serves to fulfill and the fulfilllment of that function in a different way serve as the means for the end of making the product redundant. These three core elements can be simplified and abstracted to the new definition:

Function abandoned or solved another way - no unit used

In the same manner, all other R strategies have been abstracted. The resulting definitions can be seen in table 3.

Strategy	*Abstracted description*
R0a Refuse	Function is abandoned or solved another way - no unit is used
R1a Rethink	Increasing the functionality/unit ratio
R2a Reduce	Less resources per unit with same functionality
R3a Re-use	Repetition of function per unit under different stakeholders
R4a Repair	Repair of unit to maintain original functionality
R5a Refurbish	Adaptation of unit to accommodate changing environments
R6a Remanufacture	Incorporation of unit components after end-of-life, in different unit with same functionality
R7a Repurpose	Incorporation of unit components after end-of-life, in different unit with differing functionality
R8a Recycle	Material extraction through dissolution of the unit
R9a Recover	Energy extraction through the dissolution of the unit

Table 3: Abstracted 10-R Strategies
Adapted from Potting et al., 2017

4.2.3 Assigning the strategies to the GenAI lifecycle

The first stage of the GenAI lifecycle is the “business understanding”. In this initial stage, the objectives of the project are to be understood from a business perspective, meaning understanding the background, the business objectives and the business success criteria. In this stage, the overall project is planned, allowing for the decision on how those objectives might be reached or the abandonment of the objectives overall. These actions allow for the use of the mechanisms underlying R0a. As deciding the means for reaching the objectives also entails the decision on budgets (e.g. potentially emission budgets) and the overall concept of the means, R1a is applicable. All other strategies (R2a, R3a, R4a-R9a) are based on a model being developed or already existing in the organization, therefore none of these apply at this stage.

The second stage, data collection, involves accessing and preparing data necessary for training, fine-tuning, and running the model. Although this stage plays a critical role in shaping subsequent phases of the lifecycle, its influence on the GenAI application is indirect, as the resulting model is not directly affected in this stage. Consequently, none of the proposed strategies, which all contain mechanisms that directly affect the model, are directly applicable to this phase.

The third stage, data understanding, contains the exploration and validation of the collected data. Here, the potential functionality of the model is determined. It can be explored, if the data is suitable to increase the functionality of the model (e.g. increasing the model lifespan by frequently feeding it with new data or designing it to serve a broader range of tasks), this makes strategy

R1a applicable. As neither the organizational context of model use (addressed by R0a), the technological set up of the model itself (addressed by R2a and R3a), the maintenance of a model (addressed by R4a and R5a) or the end of life (EOL) of a model (addressed by R6a – R9a) is addressed in this stage, R1a is the only directly applicable strategy.

The following step is a Go-/No-Go Point. Here the decision is made, whether the available data is suitable for achieving the desired objectives. While this can mean the abandonment of the technology (as described in R1a), sustainability impacts are not accounted for here. As the consideration of environmental impacts in a Go-/No-Go Point would require the consideration of the business context, a refuse strategy would be allocated to the Business Understanding phase. Therefore, R1a does not apply to this step. As no other actions apply, no other strategy can be utilized here.

If the project is continued, the data is prepared to be used for the training, finetuning and inference of the model. Here, data is selected, cleaned and processed to create suitable data sets for the next phases. The nature of the data sets influences among other impacts, the computing power required for training, tuning and inference, directly affecting the resource demand of the model, making strategy R2a applicable. The derived datasets impact the functionality of the model, making R1a also applicable. As data sets can be re-used for the training of new models R6a applies. Note that R7a describes the reuse of components in a unit of different functionality, but as in this scope the unit is always a GenAI model, its functionality remains the same. Therefore R7a does not apply. As neither the maintenance of the model (addressed by R4a and R5a), it's end of life (addressed by R8a – R9a) or the broader organizational context are directly addressed in this stage (addressed by R0a and R3a), no other strategy can be utilized here.

The phase “model selection and training” focuses on the development of the model. In the selection process of the algorithm, one might choose a preexisting model, making R3a applicable. R1a is not applicable, as the functionality and the overall setup are already decided. Further the technical setup of the system is designed here, defining the resources required to develop and operate the model, making R2a applicable. R6a can be applied, as the reuse of model components from a previous model might occur, if the new model is selected and trained to perform a function that the previous model failed to deliver. While components from a different project could be reused here (e.g. a foundational model) – as addressed by R7a, these actions would either be covered by R3a or serve to reduce the resource demand, meaning they are already covered by R2a. Therefore, I propose to disregard R7a in this phase. R0a is not applicable as the means to achieve the objective have already been decided. Lastly strategies R4a, R5a, R8a - R9a are not suitable, as no model, capable of performing the desired function was or is available in this phase. Similarly to the previous stage, the phase “adaptation” defines the technological design and therefore the resource demand for finetuning and operating the model, affecting R2a. R1a is not applicable here, as the functionality and the overall setup are also already decided. As the model is already chosen, R3a is not applicable. In this case a foundational model is already existing, which is fitted to a domain specific context, either because such a fitted model doesn't exist yet, or because an existing model does not perform its function in the desired manner (e.g. due to being outdated - making R5a applicable). In the case of a non-functioning previous model, elements of that (e.g. the data sets) might be incorporated in the revision (e.g. by training a different model type with the same data sets), making R6a also applicable. R0a is not applicable as the means to achieve the objective have already been selected. As R4a only addresses repairs (e.g. bug fixes) and not a revision, it is not applicable. Lastly R8a and R9a are not of relevance in this stage, as no EOL management is undertaken. R7a is disregarded for the same reason as in the previous phase.

The stage prompt engineering influences the result the model delivers and therefore affects the efficiency of the model. It impacts the way and intensity of model use, affecting the resource demand of the model during inference (allowing for R2a). R1a is not applicable, as the functionality and the overall setup are already decided. None of the other strategies is directly targeted within this action space, making R2a the only applicable mechanism.

In the phases documentation and evaluation and risk assessments, no actions are undertaken that directly change the GenAI model, as all actions in these steps serve the analysis and reporting of the developed model to inform future phases or projects. Therefore, none of the 10 strategies can be applied to these phases.

The deployment phase focuses on the way inference is conducted. As the way the model is used, has an impact on the resource demand of inference, strategy R2a is applicable. In this stage the functionality of the model and its structure are not affected, making R1a not usable here. Beyond that, the maintenance of the model during inference is not addressed (ruling out R4a and R5a), neither is the organizational context of the model use (targeted by R0a and R3a) nor the EOL of the model (affected by R9a to R9a).
Lastly the model is monitored during its use in the phase “model monitoring”. In case the model does not deliver the desired functionality, measures are taken. Such measures are for example the adaptation of a model to a changing environment or fixing of previously unidentified issues. Within the occurring actions, the mechanisms of both R4a and R5a can be therefore used. In the case that changes to the model (addressed under R4a) are not sufficient,

the model might need to be revised overall, and the old model stops being utilized. No development is being conducted in this stage, making R6a and R7a not usable here. Beyond that neither of the other strategies are suitable, as the only actions conducted in this phase, are those, targeted at making a preexisting model perform its predefined functionality, ruling out the EOL management of a model (R8a and R9a), the business context (R0a and R3a) and the mode of operation of the model beyond its ability to perform the desired function (R1a and R2a).

As no lifecycle stage contains actions targeted at processing models after their EOL, both R8a and R9a are discarded. And while a model not able to perform its function, might be retired and have its elements reused in the same project (making R6a applicable), R7a addresses a reuse of components in a different project. As R7a is theoretically a valid strategy in the field of GenAI, one might argue that it should be considered within this framework. Against that, I argue that out of the perspective of the project from which the EOL-model derived from (project A), R7a is not applicable, as its actions occur in another project (project B). In the perspective of the other project (project B), the motivation behind these actions is either the re-use of a model (R3a), or the increase of efficiency, pointing towards the use of R2a. Therefore, the actions derived from R7a are already found under R2a and R3a. R7a can be discarded.

4.2.4 The Rs of Generative AI

R0 Refuse

The first abstracted strategy “Refuse” has been shown to be applicable in the GenAI lifecycle. In the early phases of a project, it must be considered to either abandon the functionality that GenAI is believed to deliver or compare other means to solving that problem, for example traditional ML approaches or manual work. The decision to use GenAI for a use case must be justified and the negative impacts weight with the positive values created. This decision is of significant importance, as the abandonment of GenAI use, also means the full removal of the environmental impacts that could have occurred through the use of this technology.

Proposed definition: *The function that GenAI is planned to perform is abandoned or performed by other means - no GenAI is deployed.*

R1 Reframe

The strategy “Reframe” considers the overall setup of the system in which the technology is used. The goal of this strategy is to use of minimal resources for a certain use case - in the overall system, beyond the technological set up. Examples for targeted areas are governance (e.g. by setting resource or emission budgets for a use case), the project set-up (e.g. fusing multiple required functionalities into one project instead of running multiple projects in parallel) or the system/application design (e.g. deciding on who will use it, how often, where, when and how). This strategy is related to R2 Reduce, as both aim to lower the resource demand. The difference is that R2 targets technological processes and mechanisms, while R1 focuses on the system and environment, in which the technology is embedded in. This separation is necessary, as the actors and type of actions grouped under the two strategies differ from each other and should therefore be addressed separately. The name “reframe” was chosen as it aligns with the idea of fitting the technology into a new context.

Proposed definition: *Reducing the resources required to fulfilll a specific use-case, by optimizing the system and the environment that the GenAI model will be embedded in (focus on strategy, organizational set-up, governance and design).*

R2 Reduce

The strategy “Reduce” aims to reduce the resource demand of a GenAI application by optimizing the technological set-up, processes and mechanisms. This strategy occurs during the phases of technological development and use. Examples are approaches that reduce the resources required for training the model (e.g. energy-aware hyperparameter optimization), finetuning the model (e.g. adaptive backpropagation), prompt engineering (e.g. optimized prompt structures) among others. This strategy has been addressed frequently as it comes hand-in-hand with cost savings and a reduction in hardware needs. It must be noted that this strategy can increase the environmental impacts, as it may result in an increased amount of use cases and use intensity – a phenomenon called Jevons paradox. Therefore it alone is not a sufficient response to the vast environmental impacts of the technology.

Proposed definition: *Optimizing the technological processes and mechanisms to reduce the required resources for development and operation of the technology (focus on technological process and mechanisms)*

R3 Reuse

The strategy “Re-use” describes the reuse of a GenAI model as a whole in its intended functionality. The context of use, the purpose, and the task it performs might vary, but has to be re-used in its original functionality, including inference. This means, that approaches which use a model to train another one (transfer learning), are not included by this strategy - as this would present a reuse for another functionality than original intended. This strategy occurs frequently in the field, as large foundational models are applied to all sorts of contexts – finetuned or not. The goal of this strategy is the minimization of new developments which in the end may minimize environmental impacts, such as the emissions from training. The risk of this strategy is "taking a sledgehammer to crack a nut": Potentially, small problems are tackled with overly powerful solutions, such as LLMs - problems that could also be solved with more compact and smaller solutions. Therefore, the problem-solution-fit must be well considered when applying this strategy.

Proposed definition: *Leveraging preexisting models instead of creating new ones.*

R4 Release

As a fusion of strategies R4a and R5a, I propose the overarching strategy "Release". Both "Refurbish" and "Repair" are applicable strategies and valid considerations. Both of them describe approaches that allow a GenAI applications that fails to perform their desired function to achieve this function again. The only difference between the two is that this failure is created in "Refurbish" by the model or the underlying data being outdated, and in "Repair" by the system being faulty. As the approaches falling under R4a and R5a are often similar, with similar actors involved, a clear differentiation is not necessary. Instead, I propose the new strategy "Release" that summarizes both. The name change is based on "Releases" describing the release of new software versions, for example containing bug fixes (approaches that originally fall under "Repair") and updates (falling under "Refurbish").

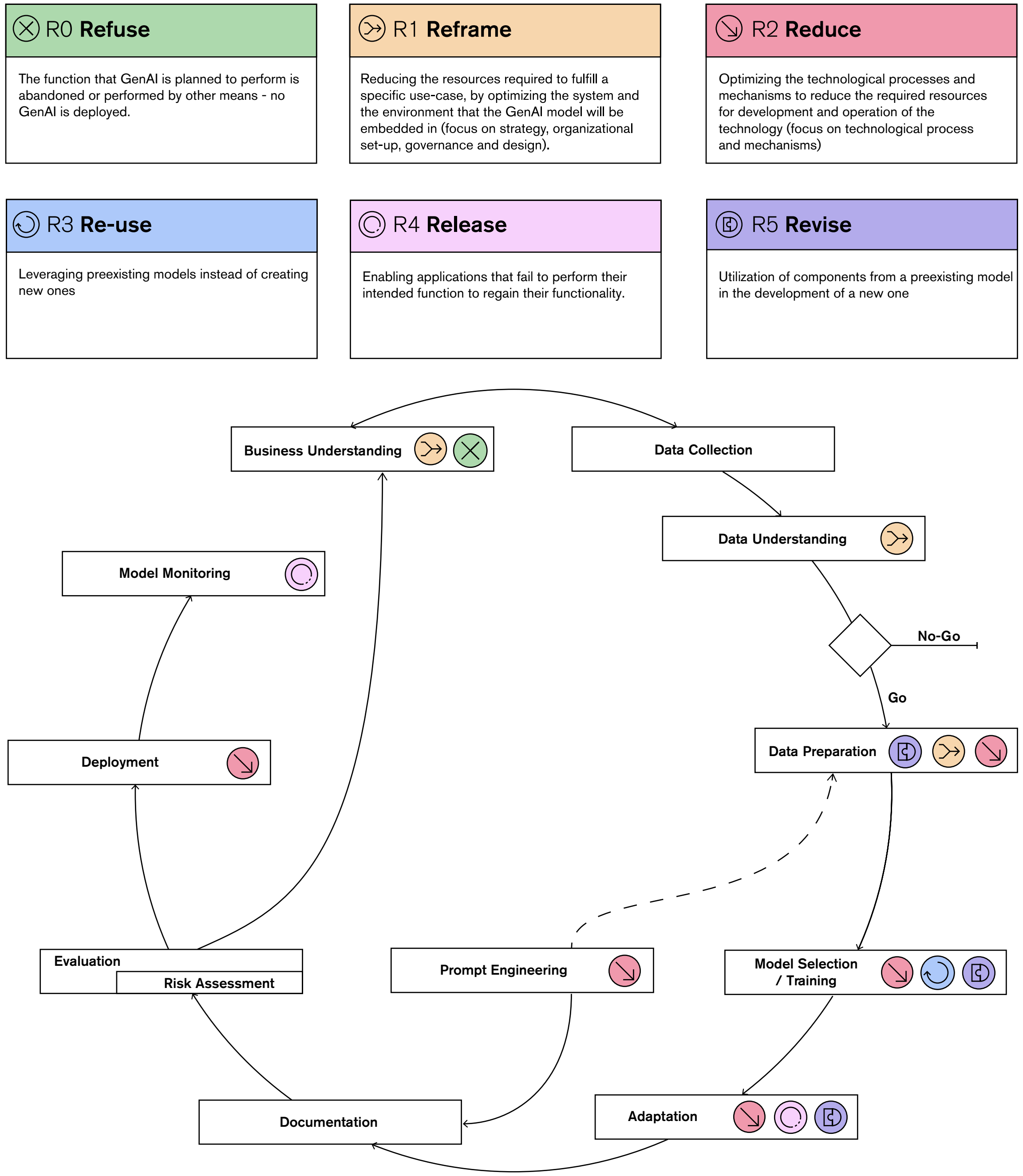
Proposed definition: *Enabling applications that fail to perform their intended function to regain their functionality.*

R5 Revise

The underlying abstracted strategy "Remanufacture" is useful in the context of GenAI in the phases in which the model is developed. When a model does not meet the necessary criteria, a new one can be developed reusing components of the previous one. Such components are for example learned knowledge structures that are transferred to a new model or datasets that can be reused for training a new model. The difference to R3 is, that in R3 that the model gets reused in its intended functionality, while in R5 only components get reused. Furthermore, I propose a name change from "Remanufacture" to "Revise", as "Remanufacture" describes a process in the domain of physical products, while "revision" describes the mechanism of creating a new version based on the parts of an old one, applying more closely to this context.

Proposed definition: *Utilization of components from a preexisting model in the development of a new one*

4.2.5 R-Strategies across the GenAI lifecycle



R0 Refuse
The function that GenAI is planned to perform is abandoned or performed by other means - no GenAI is deployed.

R1 Reframe
Reducing the resources required to fulfill a specific use-case, by optimizing the system and the environment that the GenAI model will be embedded in (focus on strategy, organizational set-up, governance and design).

R2 Reduce
Optimizing the technological processes and mechanisms to reduce the required resources for development and operation of the technology (focus on technological process and mechanisms)

R3 Re-use
Leveraging preexisting models instead of creating new ones

R4 Release
Enabling applications that fail to perform their intended function to regain their functionality.

R5 Revise
Utilization of components from a preexisting model in the development of a new one

Figure 7: The 6R Framework of GenAI (Version 1)

4.2.6 The applicability to AI/ML models beyond GenAI

This framework focuses on GenAI models, as they exhibit distinct characteristics compared to other AI model types. For example, the presence of a prompt-engineering stage in GenAI introduces additional lifecycle steps and influences the specific sustainability strategies that can be applied. Moreover, the current mapping of R-strategies to lifecycle stages emphasizes development and training phases - reflecting the high environmental impact of these stages in GenAI. In contrast, traditional ML models often incur their greatest environmental cost during large-scale inference.

To ensure applicability across a broader range of AI systems, the framework would require targeted adaptations. However, the strategy types are defined at a high level of abstraction and remain relevant beyond GenAI. They offer a conceptual structure that can guide sustainability efforts across diverse AI model types when appropriately contextualized.

Discussion

The development of a set of GenAI specific R-Strategies allows classifying the multitude of approaches into their core mechanisms and connecting these mechanisms to the various lifecycle stages. The benefit of such a conceptual map is the arising systematics, that allow for a more strategic and targeted development of measures for improved sustainability of GenAI. The framework connects concrete approaches to the various lifecycle stages and through its classification allows for a more targeted search of approaches within these stages.

While the modus operandi of the 6R framework aligns with the R strategies from other domains, the exact meaning differs mostly. The first R, R0 Refuse, is almost identical to the refuse strategy of the 10R framework. While being a potent and effective strategy, it does not align with the techno-optimist worldview prevailing in the domain, likely making it a somewhat overlooked mechanism. I argue that the R0 Refuse strategy may currently be the most critical sustainability approach in GenAI, as the uncertainty surrounding its true environmental impact makes its use inherently risky. The strategy R1 Reframe focuses on the context and environment of the model, showcasing parallels with traditional "rethink" strategies. This strategy aims to set the stage for the technology in a responsible way. As the focus of most domain experts is on the technology itself, this field – while being highly impactful – is seldom addressed. With increasing use of GenAI in the context of business organizations, this topic deserves more attention. R2 Reduce has a clear efficiency focus for the technology. I suspect that besides the multiple other benefits this strategy frequently offers (cost reductions, decreasing hardware demands etc.) it is also a way to achieve quick wins, that can be used to easily respond to sustainability concerns, without having to actually address the underlying, systemic issues. For example, the CEO of OpenAI, Sam Altman, claiming that the way forward are more climate-friendly sources of energy (Reuters, 2024). R2 Reduce approaches only focus on processes and mechanisms within the system and do not address the system as a whole. Because of that - while the R2 Reduce strategy is important - I argue that it itself is insufficient to address sustainability concerns alone, and any claims that do that should be taken with a grain of salt. R3 Reuse seems like a trivial approach, but it is still important to highlight it. Its benefits range likely beyond sustainability impacts, as for companies without the inhouse expertise or the capital to invest in external AI development, the reuse of preexisting systems can be the only way to utilize the technology in business operations. Nonetheless it offers sustainability impacts and therefore is a relevant part of this framework. The strategy R4 Release aims to both prevent the premature end-of-use of a model and extents the time of model usage. Therefore it decreases the quantities of model development in a timespan, reducing the impacts of development. Lastly, R5 Revise argues for the reuse of model components. Similar to most of the other proposed strategies its benefits extent beyond the reduction of environmental impacts, making it likely an attractive strategy for companies to also save on costs. It must be noted, that depending on the type of component that is reused its suitability for the new use case must be critically considered. For example, if a data set is reused, it must be rigorously evaluated, if it is fully suitable and a good representation for the new use case. A severe risk in that case, is that an unresponsible reuse of datasets might result in heavily biased or ill-preforming models.

In the next step, it is important to populate the framework. Consulting the literature not only allows to present examples for the different strategy types but also see if the framework is able to capture all approaches and validate its exhaustiveness. Beyond that, an exhaustive literature search can also provide insights into the maturity of the different strategies – how many approaches exist and how proven they are.
Furthermore, expert feedback will be collected to further iterate and challenge the framework.

4.3 Scoping Study

Introduction

The aim of the developed framework is the connection of specific approaches with the overarching process. The strategies formulated are abstract mechanisms that describe the overall categories into which concrete approaches can be clustered. In order to validate the selection of strategies proposed, a scoping study was conducted. The scoping study aimed to provide an exhaustive overview of approaches - presented in scientific literature, that directly positively impact the environmental sustainability of GenAI applications. Beyond testing the validity of the presented approaches, the resulting selection of approaches serve to populate the otherwise abstract framework with concrete knowledge. Further than that, a broad overview of collected and categorized approaches allows us to understand how developed the different strategies are. This knowledge points towards gaps in literature and can be utilized to inform future research, for a more targeted and strategic progress in this field.

It must be noted that the pace of new approaches being proposed is increasing quickly. Therefore this scoping study will likely be outdated fast. Nonetheless, it allows us to take a snapshot of the momentary status quo. Additionally, the study focuses on peer-reviewed articles, which restricts the scope to publicly accessible, academic sources. This approach may overlook potentially innovative and novel methods published elsewhere. However, this scope is chosen to ensure a high level of confidence regarding the validity of the identified approaches.

Method

The scoping study aims at answering the research question: How mature is the scientific research landscape around the sustainability strategy types for GenAI?

For that two main sources were searched. As most peer-reviewed paper in this field are published on IEEE Xplore, this platform was chosen to make sure the scoping successfully targeted the desired studies. In order to capture the broader landscape and make sure an exhaustive overview was generated, the digital search engine Web of Science was included. On both platform the same search strategy was used on abstracts of papers published before the 14th of January 2025. The search strategy contained four key components: First, the paper needs to be addressing the technology GenAI or AI; next the paper needs to address environmental sustainability as a whole or a sub-concept of it; further the paper needs to present a certain mode of operation, such as an approach, a technique etc.; and lastly a form of improvement towards environmental sustainability needs to be achieved. These criteria were captured in the following search string:

[„Generative AI“ OR „GenAI“ OR „generative artificial intelligence“ OR „AI“ OR „Artificial Intelligence“]
AND [„sustainab*“ OR „green AI“ OR „energy-efficient“ OR „eco-friendly“ OR „environmentally friendly“ OR „environmental impact“]
AND [„technique*“ OR „method*“ OR „approach*“ OR „framework*“ OR „strategy“ OR „optimization“ OR „optimisation“ OR „best practice*“]
AND [„reduc*“ OR „minimiz*“ OR „optimis*“ OR „optimiz*“ OR „improve*“ OR „eco-design“ OR „eco design“]

On the two platforms mentioned, this search string was applied to abstracts of papers from computer science, environmental science and directly related fields. Furthermore, only those available publicly, or via the licensing of the Delft University of Technology were chosen. All works selected were in English language.

From the resulted selection, all duplicates were removed which resulted in a selection of 941 papers (see Figure 8). The last inclusion criteria applied, was the presentation of an approach to make GenAI/AI more sustainable in itself. Works that utilized this technology to improve the environmental sustainability of applications outside of the direct context of AI/GenAI were excluded. For the core study, a focus on peer-reviewed-paper was chosen, to provide a clear scope and collect scientifically sound approaches. While this search returned sufficient suitable papers, it came with the limitation of not including any highly novel or innovative approaches – either because these could be to daring to reach a scientific consensus at this point in time, or because their development lies too close to the immediate present, meaning that a peer review process might not yet have been completed.

It is possible, that strategies and their allocation to lifecycle stages are valid and offer real opportunities, without any approach being presented under them at this stage. But in this case, this scoping study does not provide sufficient proof to validate their existence. To counter that, all strategies which couldn't be allocated with any approach from the initial scoping study are subject to a falsification approach as a follow up. For each of the strategies and the allocated dimensions, grey literature was searched specifically for the missing approaches. Finding a single approach under a strategy disproves the counterhypothesis - that no approach exists under this strategy - and therefore provides a raison d'être for this strategy. For this follow up approach a literature search was conducted. This was done by searching Google Scholar using an abstract search with the [„Generative AI“ OR „GenAI“ OR „generative artificial intelligence“ OR „AI“ OR „Artificial Intelligence“] AND [„sustainab*“ OR „green AI“ OR „energy-efficient“ OR „eco-friendly“ OR „environmentally friendly“ OR „environmental impact“] AND ["data preparation" OR "repair*" OR "bug fix*" OR "prompt engineering"]. Additionally, a paper of Bashir et al. (2024) was added which was uncovered in the background chapter “the field of green AI”, as it offered a strong fit with the “R0 Refuse” strategy in the phase “business understanding”.

The resulting selection of literature was then sorted across the corresponding R-strategies and the lifecycle stage targeted. Some papers presented multiple approaches and were therefore assigned multiple categories.

Results

The initial search returned 1538 papers, of which 509 were removed due to being duplicates. The resulting 941 papers were then screened based on the selection criteria from which only 71 fitted the scope of this study. After being categorized, those strategy/lifecycle phase combinations for which no approach was detected were populated with grey literature (see Figure 8). Overall, approaches were found for every strategy and every of its correlating lifecycle phases, proving that the framework presents at this point in time a factually correct representation of the different approaches towards environmental sustainability in GenAI applications. Due to the vast differences in the numbers of literature assigned to the different strategies (see table 4), it remains unclear to what extent the less populated strategies will be utilized in the future. Nonetheless, for now, the presence of approaches across all dimensions does indicate a relevance of the six strategies for green AI research and a correct allocation of strategies to the corresponding lifecycle stages. This uneven distribution of approaches across the R-strategies is also found in comparative studies, reviewing approaches of R-strategies in literature (Hoveling et al., 2024). This provides another argument, that an uneven distribution of approaches in itself is not reason enough to abandon a strategy, but rather a product of the context. Therefore all six identified and formulated strategies remain of interest.

Method

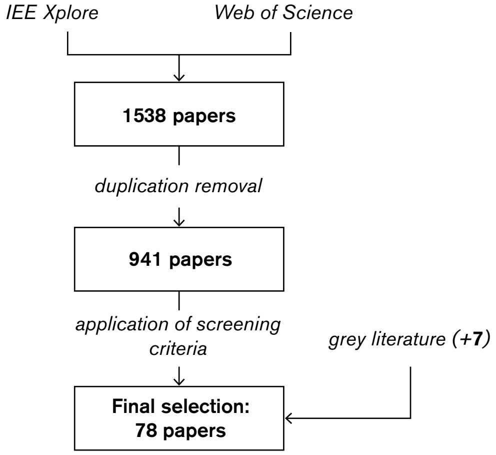


Figure 8: Literature Search

R0 Refuse Approaches

Refuse approaches are techniques that support the decision to abandon the use of GenAI technology by either replacing it with another means of performing the desired functionality or abandoning the functionality overall. Within the peer-reviewed literature that was dissected in this study, not a single refuse approach could be found. The follow-up search of grey literature revealed a refuse approach. As refuse approaches don’t necessarily lead to other benefits such as cost-savings, the key drivers are feasibility and sustainability. As only one refuse approach was identified, the nature of these approaches cannot be further summarized on generalized at this point in time. The approach found is the following:
Business Understanding: In the paper “The Climate and Sustainability Implications of Generative AI” the authors, Bashir et al. (2024) propose a comparative cost-benefit analysis of introducing GenAI applications versus a baseline scenario – meaning simply conducting business-as-usual. The framework presented is built on concepts borrowed from LCA procedures. The goal is weighing the decision of implementing GenAI on the basis of environmental costs versus the expected benefits which can result in abandoning the use of GenAI, making it a refuse strategy.

R1 Reframe Approaches:

Reframe approaches are the second most commonly found approach category from the scoping study, with 7 peer-reviewed articles out of 71 presenting them. While also reducing the impact of GenAI applications, reframe approaches focus on the set-up and design of the system and context into which the technology is implemented in, rather than the technology itself (this is addressed by R2 Reduce). Therefore the approaches are occurring in the steps before the actual technological development, such as business understanding, data understanding and data preparation. The approaches are mostly strategic or design processes that focus more on the overall understanding rather than the technical mechanisms. Exemplary approaches are:

Business Understanding: In the paper “Sustainability Budgets: A Practical Management and Governance Method for Achieving Goal 13 of the Sustainable Development Goals for AI Development”, Raper et al. (2022) propose the concept of sustainability budgets as a means of guiding the further design and set-up of the model. A fixed sustainability budget (based on the SCI (the software carbon intensity) rate, defined by the Green Software Foundation) is allocated to a project. The developers are then informed about that limit and the development is conducted in a way, so that the resulting application stays within the limit.

Data Understanding and Data Preparation: While many approaches are presented in the paper “Eco-Friendly AI: A Guide to Energy-Efficient Techniques Across the AI Life-Cycle” by Meemken & Poth (2024), one of them focuses on a reframing approach for data understanding and preparation. They outline, that data loses its predictive value over time (e.g. natural language data is reaching its half-life of predictive value after around 7 years). This means also, that the use of that data will continuously loose it’s value with growing age, while the same computation effort is required to digest it during the model (re)-training or inference. They therefore highlight the need (together with other reduce approaches) for making the effort of understanding the pace, in which the data in question is losing its predictive value and adjust the data used in the model accordingly.

R2 Reduce Approaches

Exploring the peer-reviewed literature landscape, displays that “Reduce” strategies have received a vast amount of attention in the past. Out of 71 selected peer-reviewed papers, 61 propose a reduce (see table 4) approach. Besides sustainability as a driver, reduce approaches also lead to the reduction of costs and the increase of potential applications. Reducing the computing power required to train and operate a model means, that less energy is consumed, and the model can be run on smaller hardware, such as edge devices. What sets reduce approaches apart from reframe approaches, is their strict focus on technical processes and mechanisms. Therefore these approaches occur only within the lifecycle stages that influence the technical setup of the application/model (see Figure 9). Exemplary reduce approaches across the lifecycle stages are:

Data Preparation: Exemplary reduce approaches that can be utilized in the lifecycle stage of data preparation are presented in the paper “Data-Centric Green Artificial Intelligence: A Survey” by Salehi & Schmeink (2024). Some techniques outlined in the work are “active learning”, “dataset distillation”, and “curriculum learning”. In active learning, the most informative samples of the data set are identified and labeled. The training set is then iteratively updated with the selected samples, minimizing the training dataset size and epochs required, thereby reducing energy consumption. In dataset distillation, a small synthetic data set is created that retains the information of the original, large dataset. This is done via techniques such as gradient matching or meta-learning. This way, the model learns just as effectively but much faster, because it works with fewer examples. In curriculum learning, the model is trained first using simpler examples from the data first and then gradually introducing more complex ones. This approach demands reduces the iterations in training required. These few examples show how the computational load of training can be reduced by techniques in the data preparation.

Model Selection / Training: The paper “A Lightweight Spiking GAN Model for Memristor-centric Silicon Circuit with On-chip Reinforcement Adversarial Learning”, authored by Tian et al. (2022) presents an approach to improve the efficiency in the training of Generative Adversarial Networks (GANs). The set up presented utilizes spiking GANs, a lightweight form of GANs based on Spiking Neural Networks (SNN). While SNNs continuously compute activations for all neurons, even when some computations are not necessary, spiking neurons only act when needed, resulting in sparse activity and therefore significant energy savings. In the approach presented by Tian et al. (2022), spiking GANs are then further ran on energy saving hardware and trained utilizing reinforced adversarial learning, a combination between reinforcement learning (see Appendix A) and adversarial learning (see Appendix A). Reinforcement rules in this set up particularly reduce the energy required in training by simplifying it and removing the need for backpropagation. Many approaches found in this study are reduce approaches focusing on the model selection / training stage. Therefore, many options are already known which can be implemented here.

Adaptation: One example for a reduce approach found in literature for the phase of adaptation is presented in the paper “Towards Green AI in Fine-tuning Large Language Models via Adaptive Backpropagation” by (Huang et al., 2023). With LLMs traditionally being trained utilizing back propagation, this means that for every training and finetuning cycle, every parameter is readjusted. This process leads to significant computational overhead and energy consumption, particularly in large-scale models with billions of parameters. Huang et al. (2023) propose adaptive backpropagation, a method that selectively updates only the most relevant parameters during training or fine-tuning, instead of recalibrating all parameters. By identifying and focusing on critical layers or weights, this approach reduces unnecessary computations, leading to lower energy usage while maintaining model performance.

Prompt Engineering: Rubei et al. (2025) present in their paper “Prompt engineering and its implications on the energy consumption of Large Language Models” an approach to reduce the computing power required via prompt engineering of LLMs. They present structured ways to craft zero-shot (meaning prompts that provide no examples), one-shot (providing a single example) and few-shot (providing multiple examples) prompts in combination with custom tags (such as <code> or <incomplete>) and explanation for those tags. Their findings demonstrate that the use of structured and tagged custom prompts can significantly reduce energy consumption while simultaneously improving accuracy in code completion tasks.
Deployment: Next-Word Prediction: The paper “A Perspective of Energy-Aware Distributed Inference” by S. Wang et al. (2024) presents an exemplary reduction approach during inference. The presented system utilizes multiple

trained NLP models deployed on edge servers and user devices. The user input is distributed across a selection of models for inference and the output of these models is then aggregated. The selection of models is based on an algorithm that balances prediction accuracy, latency and energy consumption (by activating models with low energy demands). This process reduces energy consumption compared to traditional centralized approaches.

R3 Reduce Approaches

Reuse approaches leverage preexisting models to fulfilll a functionality rather than developing new ones. This means that an already trained model can be useful across other functionalities beyond its originally intended purpose. It must be noted that the literature separates between three types of reusing deep neural networks: Conceptual reuse, describing the reuse of existing theory; Adaptation reuse, describing the modification of models; Deployment reuse, which describes the reuse of an existing model in a novel environment (Davis et al., 2024). In this framework, only deployment reuse is part of the R3 Reuse strategy, as it is the only one reusing the entire model within its originally intended purpose. The scoping study of peer-reviewed articles did not deliver any reuse approach. Approaches were found in grey literature:

Model Selection / Training: The paper "The Deep Learning Compiler: A Comprehensive Survey" by M. Li et al. (2020) presents deep learning (DL) compilers (see Appendix A), which are software tools designed to optimize the process of running deep learning models on various hardware platforms. This is important as reusing usually means a change of environment and therefore a change of hardware. As deep learning models grow in complexity and hardware becomes more specialized, DL compilers play a crucial role in ensuring that models are portable across hardware, which is especially important in the context of deployment reuse.

R4 Release Approaches

Approaches clustered in the category R4 Release aim at making a model, that does not perform its function functional again. This can be either because the application or the data it was trained on got outdated or because of the presence of bugs. In the peer-reviewed literature no approach was identified that utilized this to improve the environmental sustainability. While the follow-up research also did not result in an approach that specifically mentioned the use of such a mechanism for sustainability, an approach was identified that still fulfillled the criteria. It must be noted that the connection between release approaches and sustainability is less logical and clear than with other strategies (e.g. reduce or refuse strategies), which can influence the appearance of sustainability related terms in the papers. The approach identified is the following:

Adaptation & Monitoring: The paper "Repairing deep neural networks: fix patterns and challenges" by Islam et al. (2020) proposes the use of automated bug repair tools for deep neural networks. It is identified that deep neural network bug fix patterns are unique compared to traditional software bugs, primarily involving fixes related to data dimensions, network connectivity, layer adjustments, and optimization. In their study they identify specific bug fix patterns for DNN and propose automizing the repair process.

R5 Revise Approaches

The idea of revise approaches is the reuse of system components, once an AI system is either decommissioned or not usable for the functionality in question. Two approaches from the initial literature search were found that fit this type of strategy. Revise strategies are applicable across all the technical development steps, such as data preparation, model selection/ training and adaptation. The reuse of a component means, that this component does not need to be developed from scratch, saving on the resources that this new development would need. Note that in the classification of neural network reuse approaches, adaptation reuse and conceptual reuse both describe approaches that fall under R5 Revise, as in both only elements of an old model are found within the new ones. The approach identified is the following:

Data preparation, Model selection/training and Adaptation: The already previously mentioned paper "Data-Centric Green Artificial Intelligence: A Survey" by Salehi & Schmeink (2024) presents the concept of knowledge transfer / sharing. This is a family of techniques that aim to leverage knowledge gained from one task or domain (source) to improve performance on another task or domain (target). Usually, transfer learning involves using a model trained on a source task with sufficient data and adapting it for a target task, often with limited data. This reduces the computational cost, as the bulk of training is already done.

Strategy	Peer-reviewed Papers	Grey Literature
R0 Refuse	0	1
R1 Reframe	7	1
R2 Reduce	61	3
R0 Re-use	0	2
R0 Release	0	1
R0 Revise	2	0

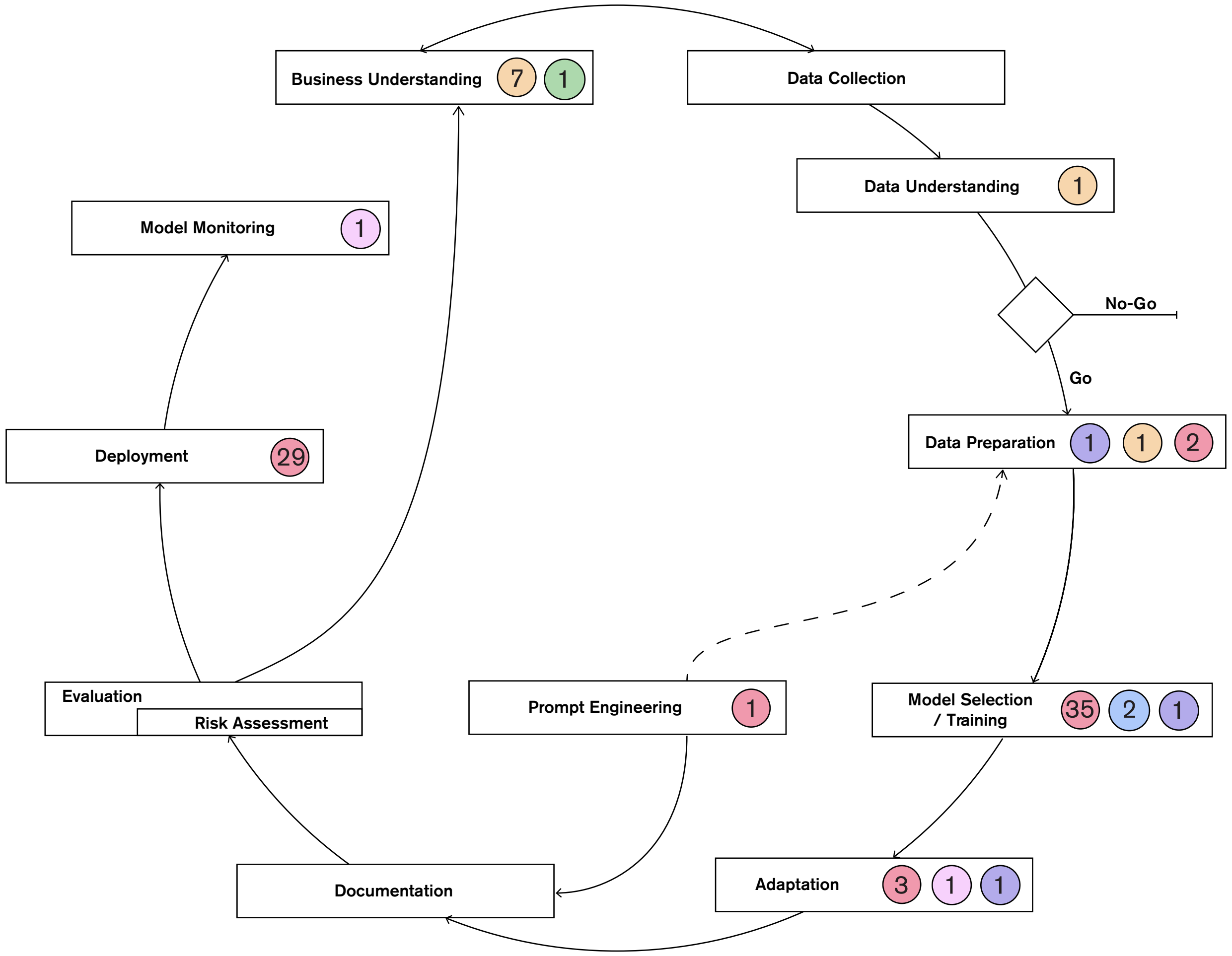
Table 4: Classification of literature into the 6Rs

Discussion

The scoping study revealed that approaches span all six R-strategy dimensions and their corresponding lifecycle stages. Moreover, every approach identified in the study that provided a direct environmental benefit could be mapped to one of these dimensions, suggesting that the framework is likely exhaustive. Therefore, the framework accurately represents circularity strategies in the context of GenAI applications.
It must be noted that an approach might affect multiple lifecycle stages and multiple strategy dimensions. When clustering approaches, the R strategies presented in this framework are therefore not necessarily exclusive to each other and overlaps exist. With more approaches arising and the GenAI lifecycle changing over time, new research might result in more exclusive classification of sustainability approaches. For now, I argue that this observation is not reason enough to change or abandon the dimensionality, as the goal of the framework is the allocation of concrete approaches to the lifecycle stages and the identification of research gaps across the different R strategies, rather than the labeling of approaches to a single strategy type.

"Reduce" approaches make up by far the majority of papers. This majority is so significant that approaches beyond are exceptions at this point in time. A hypothesis for this emphasizes on "reduce" approaches in the green AI literature is its connection to cost savings and increase of potential application. Reducing the computational demand of model training and inference means, that less energy and less powerful hardware needs to be available. While this seemingly win-win relationship that "reduce" approaches offer to sustainability and business needs, could be a driver behind this intense focus on this strategy type, it is not without risks. Decreasing costs and infrastructure requirement can lead to a scale up of use intensity and size, which can result in the already explained Jevons paradox (see chapter 3.3). While these approaches are valuable, they alone are not sufficient to build a holistic and safe sustainability strategy. The second most common type of approaches being found are "reframe" approaches. Similarly, to "reduce" approaches, these also increase the efficiency and thereby potentially reduce costs, which can explain their presence. Nonetheless their linkage to a concrete use-case makes them more resilient to negative effects such as the Jevons paradox, as a scale-up within the use case is not possible. All other approaches are drastically underdeveloped. A hypothesis for the absence of "refuse" approaches in this field is, that these approaches advocate for the abandonment of the technology, which likely goes against the interest of many experts and organization involved in the research and development of GenAI technology and applications

The strategies refuse, reuse, release and revise are all underdeveloped and offer relevant opportunities for future studies, presenting a highly relevant research gap. While these approaches might be more commonly found across grey literature, they are an integral part of green AI and therefore deserve more attention in peer-reviewed literature.



⊗ R0 **Refuse**

R1 **Rethink**

R2 **Reduce**

R3 **Re-use**

R4 **Release**

R5 **Revise**

Figure 9: Literature found across the strategies and lifecycle stages of the 6R framework for GenAI

The number in the circle indicates the amount of papers found in the scoping study, that could be allocated to the corresponding strategy and lifecycle stage.

Limitations

As the framework was formulated by the same researcher who conducted the scoping study, a confirmation bias was possibly present in the classification of the identified literature. This risk was mitigated by conducting expert interviews in parallel, which consisted of iterating on the R-strategies with outside experts.

Another constraint arises from the focus on peer-reviewed studies. Peer-reviewed journals /publishers are less quick to take up newly developed works as the review process takes time. With 36 papers out of 78 being published in 2024 or later, this field is just recently gaining traction. Therefore, new and unconventional approaches might not be captured. For this another literature review of also the grey literature would be beneficial, with more resources than those available for this work.
Lastly, the field growing so quickly in recent times, means, that the results presented here are likely to be outdated fairly quickly. Therefore it would be beneficial to update this study over time.

Conclusion

The scoping study supports the dimensionality of the framework, showing that the current arsenal of approaches is successfully enclosed and mapped out.

Beyond the validity of the framework, the study presents the status quo of currently available approaches, emphasizing on research gaps being present across the dimensions R0 Refuse, R3 Reuse, R4 Release and R5 Revise.
I advocate for a finetuning of the framework with more approaches being available in the future and the updating of the scoping study, to maintain an understanding of the research landscape.

4.4 Expert Interviews

Introduction

To interweave expertise beyond the literature into the framework, industry experts were consulted. With these expert interviews, the framework was tested, extended and further knowledge on the field was collected. The goal of this stage was to enrich the framework with further knowledge and context.
Each expert brought a unique perspective, contributing valuable insights. The interviews focused on identifying both direct and support strategies, assessing their impact, and understanding the barriers to their widespread adoption. This stage of research contributes to refining the framework by ensuring its applicability, completeness, and adaptability to current developments. The expert insights not only strengthened the validity the proposed sustainability strategies and their allocation to the GenAI lifecycle, but also allowed for a refined and more detailed framework.

Methods

For this, the experts were recruited based on their expertise in GenAI or AI and sustainability. All experts selected offered deep expertise in this topic and by that, valid insights. As the field is comparatively small, the recruitment process was challenging and only a handful of experts were identified due to the limited scope of this study.

Online interviews were conducted with a length between 30 minutes and one hour. The interviews were structured into a conversational part and a semi structured part. The conversational part aimed to address general developments in the field to further deepen insights into the current state of sustainable GenAI. The second part addressed the framework and was structured into the following topics: Firstly, the overall concept was explained without the framework being presented, the goal was to gather input without biasing the interview partners due to exposure to the framework; the second part was a presentation of the framework and an explanation of all strategy dimensions, after this, the interviewees were asked to provide an initial feedback; this initial feedback was followed by reviewing all strategies and checking there validity, exhaustiveness and their application to the lifecycle stages; then the interviewees were asked to identified the gaps of the framework; lastly the space was provided for closing remarks.

The interviews were analyzed using a thematic content analysis. For this, all relevant remarks were extracted from the interviews and clustered into the following themes: 1. Remarks that strengthen or weaken the validity of the framework, 2. Remarks that address concepts currently not addressed by the framework, 3. Remarks that don’t affect the framework directly, but are valuable insights.

The resulting remarks were then clustered into themes, with each theme being given a description and the frequency of its occurrence. Lastly this list of themes was divided into themes in the context of the framework that do require changes to the framework, themes in the context of the framework that don't require changes and themes that are not in the scope of the framework.

The framework was then adapted to the changes that arose from the list. The themes beyond those which resulted in direct changes to the framework are then discussed based on their meaning and impact to the topic.

Four experts were interviewed for this section of the project. All four partners offer contributions to the field from differing roles and organizations. This set of heterogenous perspectives allowed for the generation of diverse insights on the framework and its context. The four experts are the following.

Interview Partners

Dr. Noman Bashir is a researcher in sustainable computing and AI, serving as the Computing & Climate Impact Fellow at the MIT Climate & Sustainability Consortium and a postdoctoral researcher at MIT Computer Science & Artificial Intelligence Laboratory. His work challenges the common focus on energy efficiency in computing, emphasizing the need for a broader sustainability perspective. He develops systems, algorithms, and metrics that embed sustainability as a core objective in computer system design and operation. His contributions, including EcoVisor, CarbonContainers, CarbonScaler, and WattScope, have advanced sustainable computing practices, with one of his solutions on improving resource efficiency in data centers deployed across all Google data centers powering services like Search, Gmail, and YouTube.

Ioannis Kolaxis is a recognized expert in Green Software and sustainable computing, serving as a Director at Accenture Technology Sustainability Innovation. As the author of Green Software, an inventor with multiple patents, and a frequent speaker at international conferences, he actively drives innovation in environmentally responsible software development. He has a background at Atos, IBM, and Siemens.

Sophia Falk is a PhD researcher at the University of Bonn's Sustainable AI Lab, specializing in the intersection of AI ethics and environmental sustainability. Her work critically examines the claims around "Sustainable AI," advocating for clearer frameworks and responsible applications. She critically examines AI's environmental impact beyond carbon emissions, using the planetary boundary (PB) framework to assess AI's effects across different Earth systems throughout its entire hardware lifecycle.

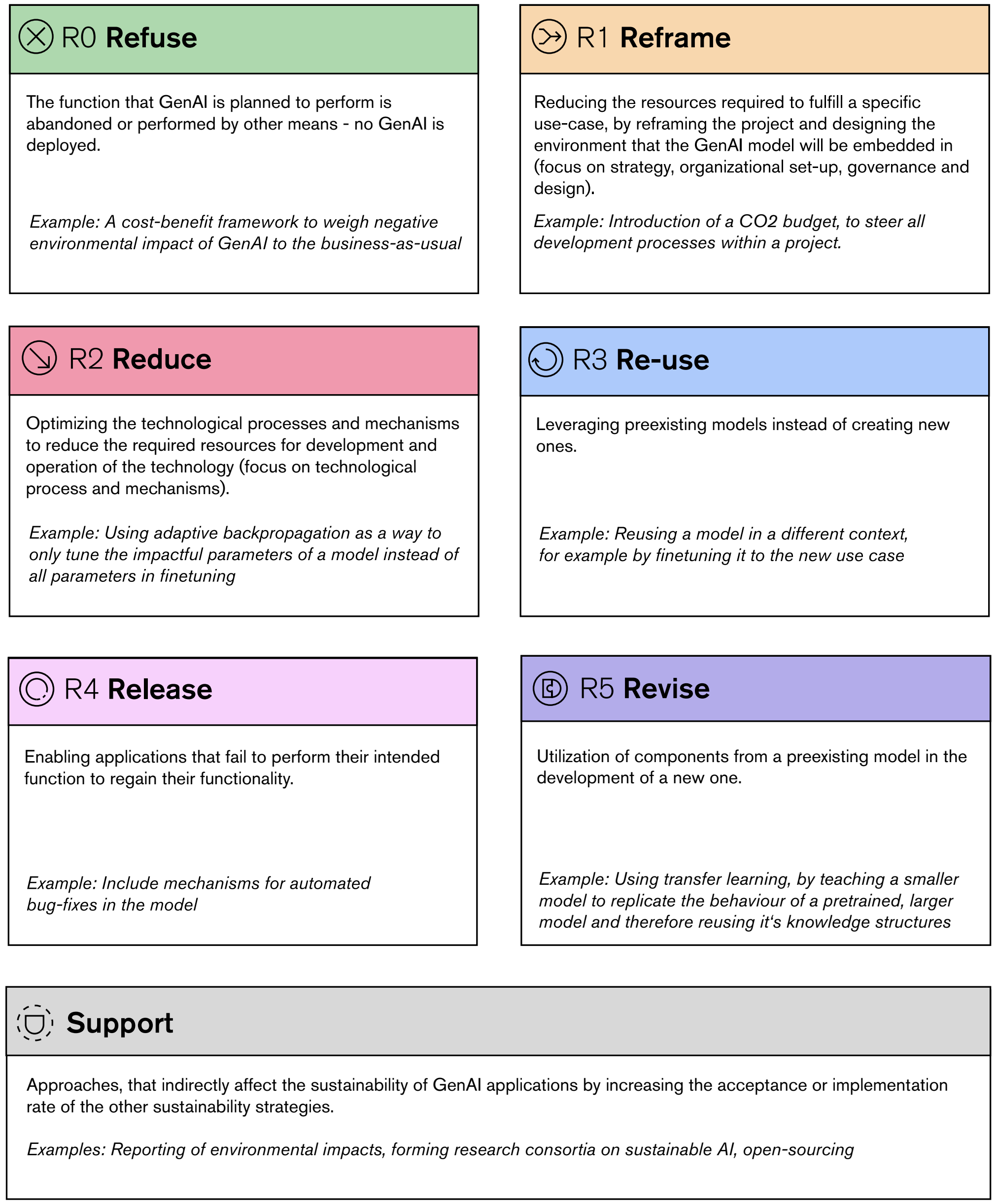
Wilco Burggraaf is a principal lead for green software engineering at HighTech Innovators, specializing in optimizing software performance with a focus on sustainability. His work involves developing efficient systems that reduce environmental impact while maintaining functionality. As a green software champion for the Green Software Foundation and a winner of Carbon Hack 24, he contributes to advancing sustainable software practices. He is also engaged in building a network of green software practitioners to promote collaboration and knowledge sharing.

Adjustments to the framework

Based on the themes arising during the interview, four changes are proposed to the framework. Two of these changes focus on the structure of the framework itself and the other two on the content.
Content-wise it was highlighted that strategies, beyond those that directly influence the sustainability of GenAI models are not captured within this framework. While the goal of the framework is the presentation of strategies with positive environmental impacts, some approaches and mechanisms can increase the adoption and implementation rate of the 6Rs and therefore exert an indirect positive influence. These support strategies can be for example efforts to increase awareness on the topic, the reporting of environmental impacts along the GenAI lifecycle (for example via tools such as CodeCarbon (CodeCarbon, n.d.)) or the formation of research consortia that further push the development and knowledge on sustainable approaches in the field. Due to the scope of this study, these indirect strategies have not been addressed in detail and further explored. Nonetheless, due to their meaningfulness and contribution to the adoption of the 6R, they will be taken up in the framework as a supporting role. As they are not further specified, they are not allocated to specific lifecycle stages.

Another theme highlighted in the interviews, was the relation of the GenAI lifecycle from a software perspective (as shown in the proposed framework) to the closely entangled hardware infrastructure used to run the GenAI lifecycle. While hardware related strategies (e.g. choosing the right datacenter to train on or timing the training of models with the availability of renewables in the energy grid) can be allocated to the different dimensions proposed, the hardware is not specifically part of the framework. As it is a substantial part of the overall GenAI lifecycle, even from a software perspective, it is added to the framework. It must be clear, that hardware systems influence the software performance in the GenAI lifecycle. This interaction is therefore visualized in the framework.

The most common remark regarding the structure of the framework was the importance of further finetuning the wording of definitions. The importance was highlighted to make every word count and increase the precision and distinction of the wordings. The second theme addressing the structure of the framework, was the request for adding examples of approaches directly on framework to increase clarity. Based on these remarks, the wordings have been adjusted and examples added.



R0 **Refuse**

The function that GenAI is planned to perform is abandoned or performed by other means - no GenAI is deployed.

Example: A cost-benefit framework to weigh negative environmental impact of GenAI to the business-as-usual

R1 **Reframe**

Reducing the resources required to fulfill a specific use-case, by reframing the project and designing the environment that the GenAI model will be embedded in (focus on strategy, organizational set-up, governance and design).

Example: Introduction of a CO2 budget, to steer all development processes within a project.

R2 **Reduce**

Optimizing the technological processes and mechanisms to reduce the required resources for development and operation of the technology (focus on technological process and mechanisms).

Example: Using adaptive backpropagation as a way to only tune the impactful parameters of a model instead of all parameters in finetuning

R3 **Re-use**

Leveraging preexisting models instead of creating new ones.

Example: Reusing a model in a different context, for example by finetuning it to the new use case

R4 **Release**

Enabling applications that fail to perform their intended function to regain their functionality.

Example: Include mechanisms for automated bug-fixes in the model

R5 **Revise**

Utilization of components from a preexisting model in the development of a new one.

Example: Using transfer learning, by teaching a smaller model to replicate the behaviour of a pretrained, larger model and therefore reusing it's knowledge structures

Support

Approaches, that indirectly affect the sustainability of GenAI applications by increasing the acceptance or implementation rate of the other sustainability strategies.

Examples: Reporting of environmental impacts, forming research consortia on sustainable AI, open-sourcing

Figure 10: The 6R Descriptions (refined)

Further remarks

Other themes directly connected to the framework arose, that deserve to be addressed in order to gain a better understanding of the field. They address managerial processes when implementing the 6R strategies in an organization; remarks addressing the mitigation of sustainability impacts of AI and remarks addressing the lifecycle process of the technology.

Managerial Factors

In the interviews it was stated, that implementing sustainability in this field comes with risks for companies (such as reducing the pace of developments). These risks must be clearly addressed, mitigated or accepted in favor of creating a sustainable future. Attention has to be paid to them, as they can hinder or even stop sustainability efforts. Another burden that needs to be addressed is the rising pressure if the use of GenAI delivers vast improvements and value to its designated use-case. This might shift the priorities to the detriment of sustainability KPIs. While these burdens are in the way, sustainability practices in GenAI also have advantages. If a company's vision includes sustainability targets or even specific sustainable IT goals, there is a need to align the day-to-day practices to these targets. If an organization uses GenAI, this can present a use case for the framework, as it allows the connection of the overall lifecycle stages to concrete sustainability practices.

Beyond the implementation of sustainability strategies, also their executing comes with potential pitfalls. A successful implementation requires actions of various stakeholders. It has been highlighted, that different stakeholders frequently use different types of information in their decision making process. For example, some might work with information being presented in reports (e.g. ESG reports), which is "mutilated" information - while others might use the information derived directly from raw data. It should be aimed to make decisions on the basis of a constant information type to improve the accuracy of decision making.

Another pitfall is the benefit allocation in the reuse and revise strategy. These strategies require two scenarios: One in which the model or specific components are offered for reuse and one in which a preexisting model and/or components are reused. While both sides are needed to achieve the sustainability improvement the question arises of which side gets credited to what extent for this improvement. This might become a problem when assigning the performance on GenAI implementation projects onto sustainability KPIs.

Lastly, the data preparation stage can provide complex challenges and is frequently underestimated. The process of cleaning data to ensure that only the data which actually contributes to decision making is left is important and also highly challenging. Therefore sufficient attention should be paid to this phase.

Sustainability Factors

It must be acknowledged, that some of the approaches that fall under the framework can come with negative sustainability impacts themselves. An example for such an impact is the increase of produced data amounts that can come from increased measuring and monitoring of the models "sustainability performance", such as its energy usage. All this data increases network traffic and storage demand, which increases resource consumption. These risks need to be considered, mitigated or weighted against the value contributed. Another issue is the use of proxies, a well-known problem in LCAs; where the use of proxies is well documented and still unsolved. As accurate data is frequently unavailable in this context, proxies are often used in the decision making process. These proxies can add inaccuracies and therefore negatively impact the decision making process. Attention must be paid to the use of proxies, their consistency and accuracy to mitigate the risk of false decision making.

In the interviews the concept of cost-benefit analysis was addressed multiple times. In theory, an action/process/decision is sustainable, if the positive impact created outweighs the negative impact. This benefit/cost comparison is not straightforward in practice, as various different types of impacts need to be compared and a translation between them found. It is therefore not directly applicable in practice, but the underlying narrative is important. It must be justifiable whether actions contribute enough positive impact to outweigh their negative impact. Note, that this concept is also addressed in an approach described earlier, classified under R0 Refuse by Bashir et al. (2024).

Remarks on the GenAI Lifecycle

Regarding the GenAI lifecycle, two remarks have been made. After the step prompt-engineering it was proposed to insert another step, in which an automated, binary-decision making step is introduced. In this decision making it is decided on whether a prompt returns acceptable results or not. While this process is frequently conducted in a trial-and-error manner, an automation, e.g. via another ML model, could streamline this process. So far, the step of prompt-engineering contained this process. As the goal the lifecycle is already a highly abstracted version, I argue to maintain this abstraction in order to keep it more concise. Secondly it was highlighted, that retraining frequently occurs once the model has been deployed. This would suggest the addition of a connection from "deployment" to "adoption". Similarly to the previous remark, it must be noted that the lifecycle illustration is an abstraction, highly simplifying the overall flow of a realistic lifecycle process. Therefore I argue against the inclusion of such a connection, in favor of simplification. The key objective of the lifecycle framework is the classification of the different lifecycle stages, which allows the allocation of strategies to the different stages. This functionality is not impacted, without the proposed changes being added.

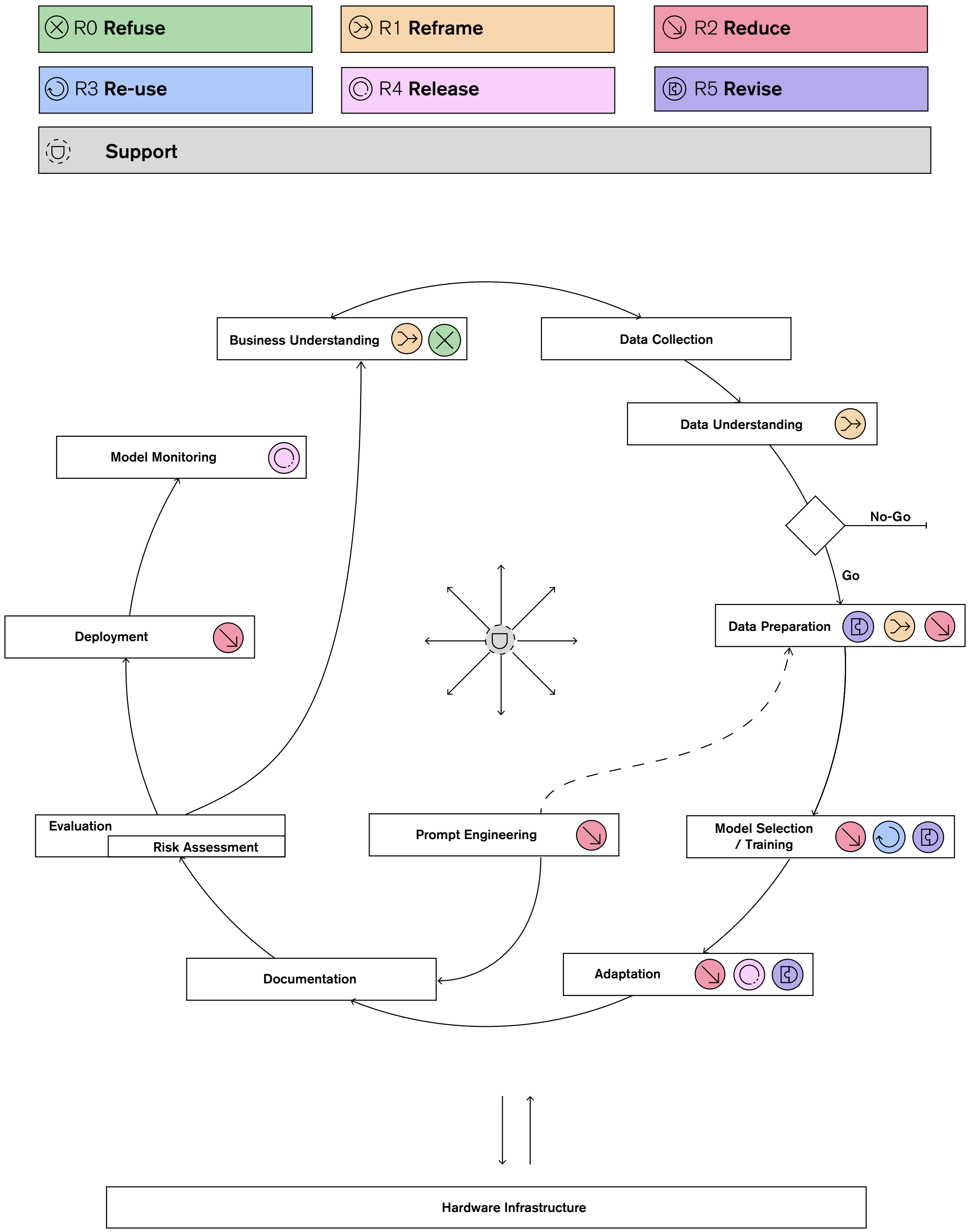


Figure 11: The 6R Framework of GenAI (refined)

Discussion

The proposed framework was refined through the feedback derived from the expert interviews. The diverse expertise of the interviewees ensured a broad range of perspectives on the framework's comprehensiveness, applicability and clarity. It became clear, that the framework fills an important gap of connecting GenAI research with the day-to-day practice in a structured manner.

Gaps and areas for improvement were identified by the experts which resulted in a refined framework. A key change was the adaptation of a support strategy dimension, containing those approaches that increase the adoption and implementation rate of the already identified R-strategies (e.g. via increased transparency and awareness). Beyond that, the hardware infrastructure was added to the GenAI lifecycle, to mitigate the risk of the software perspective taken by the framework leading to blindness considering its interplay with hardware. Further the definitions were improved and examples added to the definitions, for improved clarity.

Beyond that, contextual factors have been discussed and new knowledge on the managerial processes regarding the implementation of such sustainability approaches in organizations, the sustainability impacts of the presented frameworks and the overall GenAI lifecycle.

Overall, this chapter allowed to refine the framework in itself, but to get a more sophisticated understanding of the overall context.

Limitations

Due to the novelty of the field and the limited resources of this project, the recruitment process only produced four interview partners. Ideally, more experts should be consulted to rigorously test the exhaustiveness and validity of the framework. Besides that, it was decided against in-depth analysis of the interviews, as they served to gather feedback and not answer a specific research question.

Outlook

The role of support strategies and their definition remains superficial, since they have not been part of this research. As they contribute to reduced environmental impacts of GenAI, albeit indirect, they offer a relevant research gap and would present a valuable addition to this framework. Future work can present a taxonomy of such strategies, to understand them on a more granular level and see how the different types of support strategies interplay with the GenAI lifecycle phases.

Beyond that, the interviews displayed that the process of implementing such a strategy set is a critical area in order to successfully utilize such a framework in a corporate setting. I advocate for future research and sustainability efforts in companies to target this area, in order to prepare the ground for concrete, impactful sustainability approaches to be fruitful.

Lastly, there is a lack of structured measuring and reporting standards. This creates the risk of "comparing apples with pears" which can create fuzziness in the decision making processes. I therefore advocate for the development of a set of standards that can be used when comparing impacts, resource consumption and so forth.

Practical Blueprint

The theoretical framework proposed in the previous section offers a valuable taxonomy for identifying research gaps, broadening the perspective beyond technical approaches and efficiency strategies and selecting applicable strategies based on the lifecycle stage. While this is a valuable contribution in itself, it is an abstraction of the real-life processes which contribute and influence the environmental impacts created. Within these processes, environmental impacts occur. Thus, it is desirable to apply the proposed framework to this context. The setting of these processes is a complex context, influenced by many challenges and risks, resulting in messy processes that do not follow a linear lifecycle flow. In this environment the actors have various roles - with different motivations, backgrounds and goals. Beyond sustainability, the actions in the project further aim to fulfilll other objectives, such as profitability, feasibility, a competitive edge and performance benchmarks. The resulting environment is dynamic, complex and multidimensional, requiring an approach to be able to address changing trajectories and resource availabilities.

To understand what value the proposed framework can provide to practitioners and how an effective use of the framework can look like, this section will explore the framework in practice.

GenAI appears in the business environment in various forms. It can be deployed by individual employees who utilize it to increase their productivity, as part of the business tools and resources aimed to improve performance of certain tasks or as part of the product portfolio sold to clients. In enterprise contexts, GenAI capabilities are typically integrated into software applications to deliver targeted functionalities. From a business perspective, the GenAI lifecycle is therefore closely tied together with the development of GenAI-based applications. Therefore, the focus of this chapter lies on GenAI-based applications, rather than the isolated models. The aim is to develop an understanding, of how the framework can inform the environmental sustainability of the development and deployment of GenAI based applications.

While the blueprint is context-specific, this research also identifies generalizable patterns that may inform sustainability practices in other GenAI application contexts.

5.1 Field Research

Introduction

To ensure that the blueprint is both usable and actionable, it must be grounded in the specific context for which it is intended. For this purpose, a project within a multinational IT and consulting firm was selected as an exemplary case. Anchoring the blueprint in a concrete setting allows for the exploration of the dynamics, constraints, and stakeholder interactions that shape GenAI application development in practice. This chapter describes the case context and its key characteristics and derives implications to inform the design of a sustainability blueprint that is both context-sensitive and practically applicable.

Method

To gain this practical understanding of the reality of GenAI application projects, field research was conducted (see table 5). This was done by actively engaging in the practice and thereby extracting insights about this context, that can be used for developing the blueprint. Over the course of five months, various active engagements in the context were conducted. They can be divided into two types of field research.

The first part of the field research was the engagement in various activities in the realm of Green IT activities within the targeted company over five months. Exemplary activities are the coauthoring an industry whitepaper on Green IT, the organization and participation of an industry wide panel discussion, various trainings on Green IT and client presentations. The purpose of this type of field research was the identification of company/industry-specific characteristics and to gain an understanding of the current Green IT landscape in the industry. This part served to provide breadth to the field research

The second field research was a week-long participation in the sprint of the target project. Within this sprint, the project framing was sharpened, and a business case defined. Further a functioning demo was developed, and the results of the sprint were pitched to the country leadership and the European leadership responsible for this project type. The field research consisted of actively engaging in the project and noting down the observations. The insight types sought in this activity were practical requirements of GenAI based application development and the corresponding decision-making criteria. Further project-specific and AI-specific characteristics were extracted. This part served to provide depth to the field research.

Active participation in these activities were followed by note-taking of observations and the aggregation of these insights through thematic analysis. The resulting findings serve as the foundation for designing a targeted sustainability blueprint. The aim of this section is therefore both the description of the context as well as implications for the sustainability blueprint, which proposes sustainability measures that align with the project's operational realities while addressing leadership expectations and industry concerns about the responsible use of GenAI.

The unit of analysis of this research is the status quo of the context at the time of research (beginning 2025) and the development which occurs within the project and the involved stakeholders during the time of research. Therefore both past and future developments are not within the scope. Further company characteristics that cannot be observed within the described activities are excluded. Lastly client involvements are not subject of this research in order to not violate the sensitivity of client-related information.

	Insight Types	*Duration*	*People*	*Information Collection*
Participation in a sprint of the target project. Consisting of developing a business case, a demo of the application and a pitch to leadership	**Practical requirements, decision-making criteria, project-specific characteristics & AI-specific insights**	1 week + full time	project team & leadership	Active participation followed by note-taking and a thematic analysis
Engagement in GreenIT practice of the target company, through engaging in a whitepaper development, a panel discussion, various trainings and client pitches	**Company-specific characteristics, current state of GreenIT landscape**	5 months, part time	internal GreenIT experts & external industry experts	

Table 5: Field Research Overview

Company Context

The project takes place in the global, multinational professional services company, with a focus on information technology services and management consulting. The company is of significant size (over 750 thousand employees), it provides a large variety of (mostly IT focused) services to many private and public organizations globally. Currently, Green IT is an established capability within the company and first efforts arise around GreenAI. The company has a track record of implementing Green IT measures in public and private organizations (for example standardizations on quantifying carbon emissions of software or sustainability oriented software development). The company implements a significant amount of GenAI based applications in client organizations and therefore also has a significant stake in the environmental impacts created by GenAI based applications. While GreenAI is not a mature capability yet, current development suggest it will become one in the near future. On various occasions throughout the field research, measures to mitigate and reduce the environmental impacts of AI use have been highlighted as important through leadership actors of the company. This showcases that the importance and value of sustainability strategies for AI have been identified and are becoming increasingly addressed. Due the company being a for-profit organization, the offerings are mostly driven by customer demands. Therefore forces which shape the customer demands, such as consumer behavior or regulatory pressure have a significant impact on the company activities, including the uptake of GreenAI. Due to current, destabilizing geopolitical events, such the EU's omnibus proposal (European Comission | Directorate-General for Communication, 2025) which reduced regulatory pressure of companies' sustainability reporting, the future of customer demands in regard to sustainability of AI remains uncertain. Nonetheless, the environmental impacts deriving from AI-use have been recognized and GreenAI will likely become a central part of the sustainable IT offerings.

Project Context

The project is located in a European team focused on leveraging AI for sustainability, for example by automating processes around mandatory environmental, social and governance (ESG) reporting through AI. The projects of the team consist both of internal projects as well as client deliveries. Thus, an application developed can be both deployed within the organization itself but also sold to external organizations. The primary stakeholders involved are leadership, the core project team, asset owners and clients. The leadership sponsors the project, steers the overall trajectory and has the "last call" in the decision making. To secure funding for the project, leadership must be convinced and involved by the project team. The project team consists of both business and technology focused experts. They execute the project, manage the day-to-day operation and communicate with the other involved stakeholders. As the company already owns a vast selection of different assets (such as software applications) the use of such assets connects the asset owners to the projects. And lastly if the project, or a certain element of it get sold to a client, the client becomes a stakeholder. Through this, the client might sponsor parts of the project and in return be provided the application that is being developed.

The project is about the development of an application based on ESG data, which leverages GenAI to improve a range of ESG-related activities of companies. Such activities are for example ESG related compliance, for example the addressing and harmonization of various sustainability reporting frameworks and standards. Other activities the application addresses are for example the extraction of ESG insights, decision making support when formulating sustainability strategies or automated sustainability reporting.

This breadth of functionalities within one application is possible through the utilized technology. The underlying technology used is a multi-agent system. The central elements of a multi-agent systems are an orchestrator model and agents. The role of each agent is the performance of a specific task. Depending on the user input, the orchestrator dissects the required functionality into specific sub-tasks, which are then assigned to the corresponding agents. Depending on the architecture and request, a single agent might either perform the assigned task by itself or communicate with other agents to perform the task together. This results in highly targeted responses to requests, as agents are selected based on the nature and content of the request and each of the agents are focused on a certain task. This set-up allows for easy scalability, as new agents can be added to the system if the complexity or type of tasks changes over time. Further it is easily maintainable, as certain lacks in performance can be narrowed down and addressed in a targeted manner. Multi-agent collaboration allows LLMs to access external resources and collaborate on subtasks. This enhances the accuracy and efficiency of task execution (Chawla et al., 2024). Beyond that, the tailored response allows for the use of smaller models and therefore reduce energy demands and costs without trading off accuracy (Mangal et al., 2024).

In the context of the examined project, the agents can consist of pre-existing assets or be newly developed based on the needs of clients. The focus of the project is the development of the orchestrator and key agents which are required to fulfilll a starting section of the envisioned future functionality. Down the line more and more agents can be added towards creating a one-stop-shop for all ESG related activities.

For now, the project is seen as a long-term vision that gets constructed in subsections consisting of the connection of a selection of assets which then will get integrated with each other. Because of the project being long-term, only a small section of the project can be examined in the research. At the time of the research, the project is in an early stage. Currently, the focus lies on the strategic framing of the project, the identification and seizing of first opportunities and the development of single assets.

The process itself follows an agile methodology and is therefore highly iterative. This sequence of fast iterations means, that GenAI lifecycle stages are not followed in a linear order in which one passage of the lifecycle will yield a deployable application. Instead, stages are repeated multiple times. Beyond that, multiple lifecycle stages appear in parallel. Generally, the business understanding phase runs parallel to those phases occurring in the technical development of the application.

Characteristics & Implications

Various characteristics of the examined project need to be addressed in the sustainability blueprint. Based on the field research, the following characteristics emerged:

1. Project development occurs across locations and time: Through the patchwork-style nature of the project, different teams across different locations are responsible for elements of the project. This risks sustainability approaches remaining siloed within individual teams rather than being systematically adopted across the project. Additionally, workflows across teams may not be standardized, making it unclear how sustainability efforts can be consistently implemented. This raises the question on how sustainability efforts can be applied across time and location.

Implication for the blueprint:
As focusing on a single project phase and the involved workforce would not sufficiently address the project parts outside of this scope. Therefore, an overarching structure is required which ensures a systematic and standardized way of incorporating sustainability considerations into GenAI projects within the company. Such overarching mechanisms in a company are governance structures. The sustainability blueprint must therefore be an overarching governance blueprint.

2. Funding is a key constraint: The project gets developed in different stages with funding being uncertain for the next stage. The goal of each stage is to provide an application (or demo) which offers (or showcases) sufficient performance to be granted funding through various sources for the consequent stages. This results in resource constraints, an opportunity-driven and agile planning approach, and high unpredictability. The resources available for the development - and thus also those available for sustainability measures - are therefore fluctuating across different phases of the project.

Implication for the blueprint:
Early-stage sustainability measures must be cost-effective and low-overhead. This can facilitate broader acceptance and adoption. As discussed earlier, various sustainability approaches in IT are aligned with cost saving measures, mostly efficiency measures. Note, that these measures alone are likely insufficient due to the risk of promoting increased usage (Jevons Paradox), requiring additional measures and governance to prevent rebound effects (e.g. usage caps, carbon-aware scheduling). Once early stage sustainability measures are in place, and acceptance is rising, new measures can be introduced with rising intensity. This approach risks an overly strong focus on short-term solutions rather than long term visions. A phased sustainability approach should be developed to prioritize cost-effective short-term measures while ensuring alignment with long-term sustainability goals. To be able to effectively leverage the resources which are available, the sustainability blueprint needs to allow the flexibility to operate with whatever resources can be provided. A rigid blueprint that dictates a fixed set of sustainability measures is not suitable. Instead, the blueprint needs to map out conditions and rules which allow for sufficient flexibility to function in high resource volatility while ensuring that sufficient sustainability measures are in place to minimize environmental impacts.

3.Sustainable AI expertise is not a mature capability in the project team (yet): While the AI/ GenAI capabilities are quickly expanding and becoming a core focus of the organization, the sustainability of AI remains a fairly novel topic that - while being explored - remains mostly unaddressed. The overall workforce developing AI solutions in the organizations is not trained on sustainability of AI at this point in time. First initiatives are appearing both from bottom-up and top-down that aim to address this topic. If the trajectory of this development continues, it might become a mature capability in the future.

Implication for the blueprint:
With literacy of sustainability measures of AI being low and little to no resources being available, sustainability measures need to be identified, developed and implemented ad hoc. For this, the proposed conceptual framework introduced in previous chapter can be utilized. Through involving project teams in this process and thereby leveraging “learning-by-doing”, the literacy can be further improved over time, until GreenAI becomes a mature capability within the organization.

4.Existing assets are integrated into the system: Due to the agentic nature of the applications, some preexisting assets can be reused as agents. These assets have not been developed under sustainability guidelines but will be a part of the final application. At the current point in time, it is unclear what extent the final solution will be based on preexisting assets and how much will be newly developed.

Implication for the blueprint:
To prevent the integration of preexisting assets from creating uncontrolled impacts when being integrated into the new development, measures need to exist in the blueprint that identify and prevent environmental impacts and risks associated with these preexisting assets. It needs to be ensured, that whatever sustainability measures are implemented, these measures sufficiently address also the impacts of the integrated assets.

5. AI workloads occur mostly externally, e.g. at cloud providers: The computational work of the development and deployment of the application is being run externally (e.g. at cloud providers). Therefore quantifications and data on these computations must be collected from these third parties. Often, the provided insights are not rigorous and clear enough to be used in impact assessments (e.g. frequently the emissions are reported after carbon offsetting). While deployers of AI increasingly pressure cloud providers to increase transparency, widespread transparency so far is not the standard.

Implication for the blueprint:
The sustainability measures provided in the blueprint must function without impact measuring. Therefore, making the governance blueprint only applicable for projects over a certain impact threshold is not feasible. Instead, all GenAI projects must apply the blueprint and the implemented sustainability measures must address the impacts which are estimated. In the future, with more standardized quantification methods, a threshold could be introduced, but for now such a threshold would bear the risk of project teams pushing down estimates to prevent the governance blueprint to apply.

Discussion

The findings align closely with the previously analyzed literature. External stakeholders, internal stakeholders and leadership have emphasized on multiple occasions the need for sustainability of AI. This showcases that awareness of the sustainability issues deriving from the technology is already widespread in the industry. Nonetheless, little work has been done at the current point in time on the sustainability strategies for AI and no mature and sufficient capabilities are available. This means, that these capabilities will need to be built. A gap which offers opportunity for the previously presented framework for sustainability strategies of GenAI.

The uncertainty regarding the future of the project ends the temptation of long-term sustainability strategy roadmaps and waterfall approaches. Instead, an agile, experimentation and iteration based approach is promising, as it is able to respond to changing budget, time and personnel availabilities, can adjust based on the success or failure when experimenting with novel approaches and can be fitted to contextual factors, such as client demands. Such an approach will require the dedication, curiosity, risk appetite and creativity of the involved stakeholders, but can return innovative and effective approaches - potentially leading to new best practices and significant impacts. In case of lacking engagement, incentives need to be provided, for example through leadership.

5.2 A project-specific sustainability blueprint

Introduction

The goal of this chapter is the proposal of a blueprint, that demonstrates how the framework can be embedded to the organizational realities. By doing that, it serves as an exploration and demonstration on how sustainability can be integrated within the day-to-day operations of developing GenAI based applications. To achieve this, it builds on the organizational and technical realities. It addresses the fragmented, resource-constrained, and rapidly evolving nature of the project, proposing a phased, actor-driven, and iteration-based approach. The absence of mature sustainable AI capabilities is recognized; thus, the strategy introduces lightweight, progressively intensifying sustainability measures tailored to different workstreams. It emphasizes decentral coordination and experimentation-based development of approaches. By aligning with agile development cycles, the strategy aims at enabling sustainability integration without significantly disrupting delivery timelines. Due to the lack of best practices, the promoted approach demands a level of creativity, as the abstract mechanisms described in the proposed framework need to be translated to the project-specific context to uncover novel sustainability approaches which can be implemented.

Method

Building on the previous chapter, which explored the project context and derived design implications, this chapter translates those insights into a first version of the sustainability blueprint. To capture the necessary levels of granularity, a layered architectural approach was applied. The blueprint is grounded in the previously developed sustainability framework, which serves as a tool for identifying appropriate sustainability strategies across different stages of the GenAI application lifecycle.

The structure and content of the blueprint were informed by findings from the field research, which guided the identification of relevant layers and the required level of detail. Through iterative refinement, the blueprint was aligned with project realities while operationalizing the framework in a way that supports day-to-day development work. An initial validation was conducted through an informal feedback round with the portfolio owner overseeing the project, ensuring preliminary alignment with organizational practices and operational constraints.

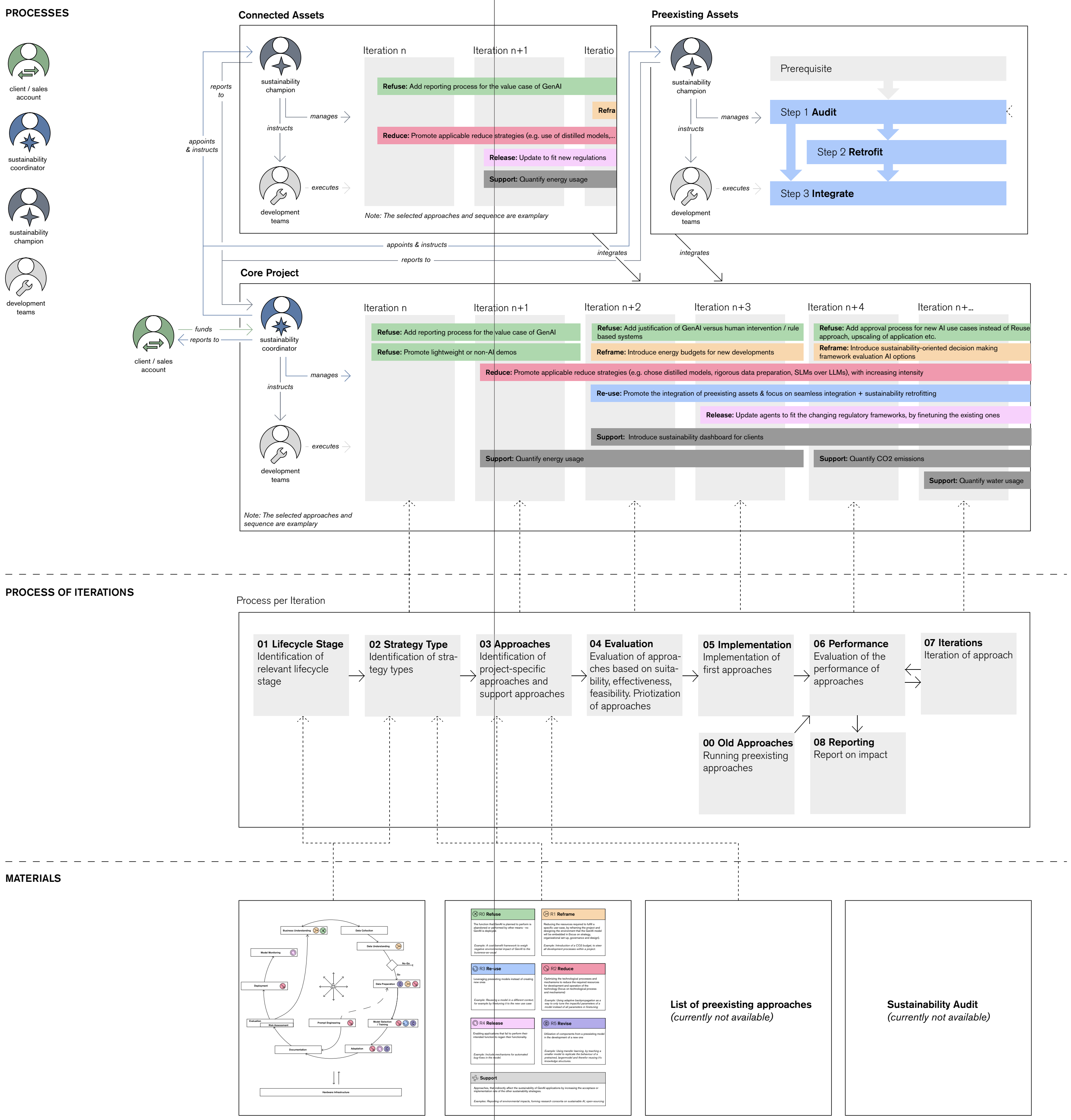

PROCESSES
client / sales account
sustainability coordinator
sustainability champion
development teams
Connected Assets
sustainability champion
manages
instructs
executes
development teams
Iteration n
Iteration n+1
Iteratio
Refuse: Add reporting process for the value case of GenAI
Refra
Reduce: Promote applicable reduce strategies (e.g. use of distilled models,...
Release: Update to fit new regulations
Support: Quantify energy usage
Note: The selected approaches and sequence are examplary
reports to
appoints & instructs
appoints & instructs
reports to
integrates
Preexisting Assets
sustainability champion
manages
instructs
executes
development teams
Prerequisite
Step 1 Audit
Step 2 Retrofit
Step 3 Integrate
integrates
Core Project
client / sales account
funds
reports to
sustainability coordinator
manages
instructs
executes
development teams
Iteration n
Iteration n+1
Iteration n+2
Iteration n+3
Iteration n+4
Iteration n+...
Refuse: Add reporting process for the value case of GenAI
Refuse: Promote lightweight or non-AI demos
Refuse: Add justification of GenAI versus human intervention / rule based systems
Reframe: Introduce energy budgets for new developments
Refuse: Add approval process for new AI use cases instead of Reuse approach, upscaling of application etc.
Reframe: Introduce sustainability-oriented decision making framework evaluation AI options
Reduce: Promote applicable reduce strategies (e.g. chose distilled models, rigorous data preparation, SLMs over LLMs), with increasing intensity
Re-use: Promote the integration of preexisting assets & focus on seamless integration + sustainability retrofitting
Release: Update agents to fit the changing regulatory frameworks, by finetuning the existing ones
Support: Introduce sustainability dashboard for clients
Support: Quantify energy usage
Support: Quantify CO2 emissions
Support: Quantify water usage
Note: The selected approaches and sequence are examplary
PROCESS OF ITERATIONS
Process per Iteration
01 Lifecycle Stage Identification of relevant lifecycle stage
02 Strategy Type Identification of strategy types
03 Approaches Identification of project-specific approaches and support approaches
04 Evaluation Evaluation of approaches based on suitability, effectiveness, feasibility. Priotization of approaches
05 Implementation Implementation of first approaches
06 Performance Evaluation of the performance of approaches
07 Iterations Iteration of approach
00 Old Approaches Running preexisting approaches
08 Reporting Report on impact
MATERIALS
R0 Refuse
R1 Reframe
R3 Re-use
R2 Reduce
R4 Release
R5 Revise
Support
List of preexisting approaches (currently not available)
Sustainability Audit (currently not available)

Figure 12
Blueprint (V1) - Overview

5.2.1 Blueprint Overview

The resulting blueprint addresses various levels of granularity (see Figure 12). This granularity is divided over three layers.

The first layer showcases the overall structure of a development project for a GenAI based application. This layer presents the different relevant project workstreams and their sub steps, such as project iterations. It is showcased how sustainability measures apply across these sub steps and phases overall. Further this layer presents the different roles required to execute the sustainability measures and their responsibilities.

The second layer outlines the process that occurs within each iteration defined in the first layer. In this process, actors apply abductive reasoning to make informed hypotheses about which sustainability approaches are appropriate for the specific iteration. These "best guess" decisions are based on the iteration's contextual characteristics and the guiding sustainability framework. The selected approaches are then implemented and evaluated. Those that prove effective are retained and refined in subsequent iterations, enabling a cumulative and adaptive integration of sustainability into the development process.

The third layer comprises the materials required to operationalize the process described in the second layer. The primary resource is the sustainability strategy framework, including the classification of strategy types and their alignment with the GenAI application lifecycle. As the project progresses and experience with specific sustainability approaches accumulates, these learnings are documented and added as supplementary materials to support future iterations. In addition, emerging standardizations - such as a sustainability audit for preexisting assets - are incorporated as they become available. The material base is designed to expand over time, reflecting growing organizational capabilities and resource availability.

Overall, the three layers showcase the overarching project structure and underlying processes and materials.

5.2.2 Deep dive - Layer 1: Structure

Workstreams

Core Projects: This element of the strategy describes the activities that are related to the central application that is being developed in the scope of the researched project. The development work in this field occurs in a sequence of various iterations. In each iteration the entire application is refined, which allows for the introduction of new sustainability approaches with each round.

Based on the findings, these approaches shall be fairly unobtrusive and steadily increase in their amount an intensity. For example, in the first iterations, two refuse approaches might be introduced: First, a reporting process regarding the value case of GenAI could be added, making sure that the value added through GenAI use is apparent. If that is not the case, no value case can be provided, and the technology cannot be used. Secondly, lightweight, or non AI prototypes can be used if this technology is not explicitly needed (e.g. when demonstrating a UI or workflow). This would remove the need for unnecessary AI models and therefore computational demands of prototyping. Down the line, other sustainability strategies can be added. Once the support strategy of measuring the energy use is in line, energy budgets for the iterations might be introduced as a reframe strategy. Beyond that Reduce strategies can be added, such as the use of distilled models, rigorous data preparation or the use of SLMs instead of LLMs. Throughout the iterations, the formation of various sustainability approaches, across the applicable sustainability dimensions gets increasingly comprehensive, rigorous and impactful (see Figure 13). The abovementioned approaches are exemplary.

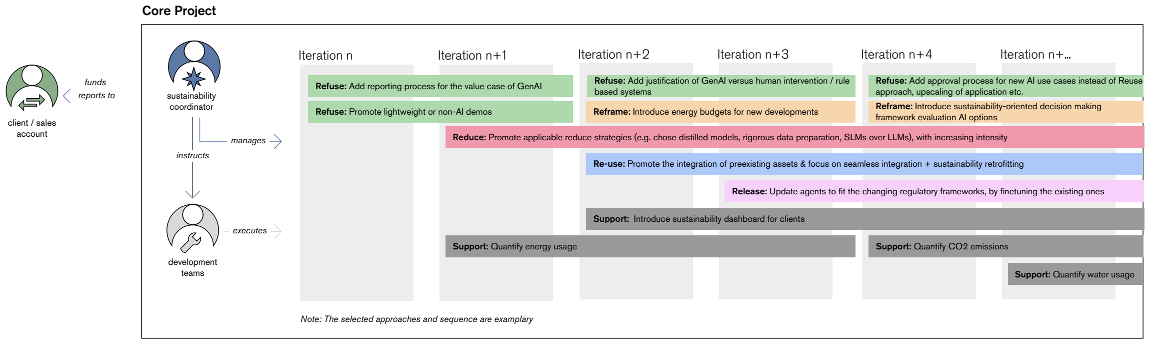
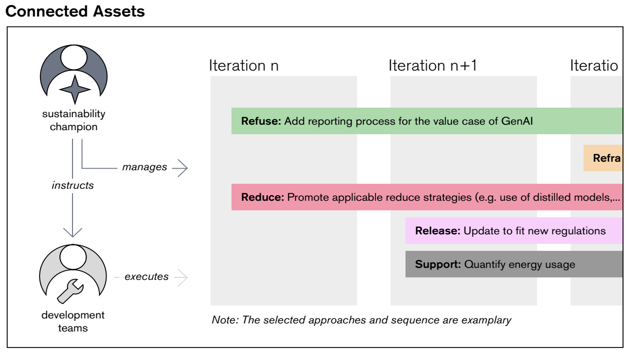

Core Project
client / sales account
funds
reports to
sustainability coordinator
manages
instructs
development teams
executes
Iteration n
Iteration n+1
Iteration n+2
Iteration n+3
Iteration n+4
Iteration n+...
Refuse: Add reporting process for the value case of GenAI
Refuse: Promote lightweight or non-AI demos
Refuse: Add justification of GenAI versus human intervention / rule based systems
Reframe: Introduce energy budgets for new developments
Refuse: Add approval process for new AI use cases instead of Reuse approach, upscaling of application etc.
Reframe: Introduce sustainability-oriented decision making framework evaluation AI options
Reduce: Promote applicable reduce strategies (e.g. chose distilled models, rigorous data preparation, SLMs over LLMs), with increasing intensity
Re-use: Promote the integration of preexisting assets & focus on seamless integration + sustainability retrofitting
Release: Update agents to fit the changing regulatory frameworks, by finetuning the existing ones
Support: Introduce sustainability dashboard for clients
Support: Quantify energy usage
Support: Quantify CO2 emissions
Support: Quantify water usage
Note: The selected approaches and sequence are examplary

Figure 13: Blueprint (V1) - Core Project

Connected Assets: Outside of the core project, assets get developed with the intention of being integrated with the core project. These function similar, as they also follow a sequence of iterations. As their integration into the core project is intended, their sustainability impact needs to be addressed in a similar manner. While the project team of the core project might have made first learnings, generated knowledge and skills on sustainability approaches, which can be shared with the connected assets, those developing the assets have not made these learnings themselves yet – therefore a gradual, phased approach is also required here (see Figure 14). Due to a reduced number of iterations compared to the core project, the sustainability approaches cannot be as exhaustive but still be increased over time. For example, they could start with implementing the same reporting process for the value case of GenAI as a refuse strategy or the reduce approaches which proved to be successful in the core project. Down the line, they can add more and more approaches, such as updating outdated assets (e.g. due to changing regulations) instead of developing a new version (a release approach). The selection of approaches is depending on the context and nature of the assets – the abovementioned approaches are exemplary.


Connected Assets
sustainability champion
manages
instructs
executes
development teams
Iteration n
Iteration n+1
Iteratio
Refuse: Add reporting process for the value case of GenAI
Refra
Reduce: Promote applicable reduce strategies (e.g. use of distilled models,...
Release: Update to fit new regulations
Support: Quantify energy usage

Note: The selected approaches and sequence are examplary

Figure 14: Blueprint (V1) - Connected Assets

Preexisting Assets: Integrating preexisting assets can be valuable from a cost-saving perspective but is also likely beneficial for the environmental impact due to the saved resources a new development would have required (it's a reuse strategy!). These assets have been developed without an integration into the core project being planned from the get-go, thus it is unclear to what extent, they were developed towards being sustainable. Before they can be integrated, various steps need to be performed on them to make sure that they don't harm the sustainability efforts of the core project (see Figure 15). First, a sustainability audit needs to be conducted to ensure they fulfilll the required sustainability needs. Second, depending on their maturity and the feasibility, they can be "retrofitted" to improve their sustainability, for example by removing certain components which are not required for their role in the project, to reduce storage and computation needs. If they are deemed suitable after the audit and retrofitting, they can be integrated into the main project.

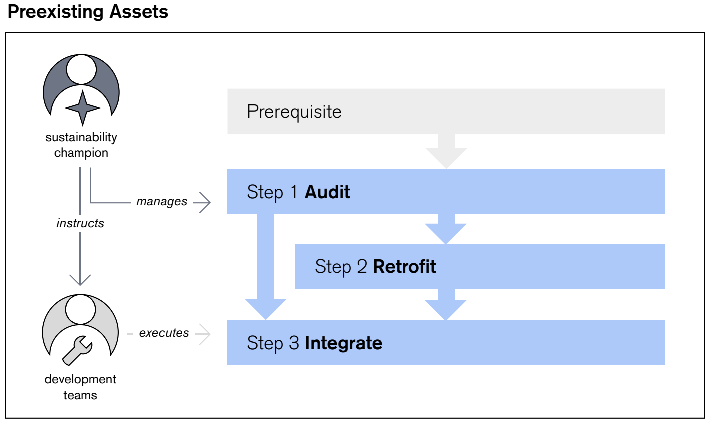


Figure 15: Blueprint (V1) - Preexisting Assets

Actors

Various actors are needed to ensure seamless coordination of sustainability efforts in the project (see Figure 16). While few to no experts are available most tasks will have to be carried out by the preexisting team members. I propose the following roles:

Sustainability Coordinator: This actor serves as the central manager of all sustainability efforts related to the project. As this effort is significant, I propose the appointment of a designated sustainability coordinator (e.g. a company internal sustainable IT expert). This coordinator reports the sustainability performance to the client accounts or sales teams which can use this information to a) argue for a higher price due to the projects sustainability performance or b) use it as a competitive advantage to increase the chances that the project gets sold. The central activity of the sustainability coordinator is to manage the sustainability approaches of the core project, by a) planning them and b) instructing the development teams on how to execute the planning. Further, they appoint sustainability champions in the connected assets and preexisting assets and instruct those on how to execute the overarching strategy.

Sustainability Champions: These actors are team members of the connected assets and preexisting assets. They receive the additional role of managing the sustainability approaches of their corresponding asset. Like the sustainability coordinator, this consists of planning the sustainability approaches and instructing the development teams to execute the planning. In the case of preexisting assets, the coordinators conduct the audits, plan approaches for the retrofit and give the prerequisite free for integration.

Development Teams: The development teams integrate the chosen sustainability approaches into their development practice. They execute the planning of the sustainability champions / coordinators.

Client Account: The client account or sales leads bring the added sustainability benefits to the clients. This added value can be used to increase the service offering and by this might allow charging a higher price, but it also serves as a point of differentiation to the competitors which can increase the chances of closing the deal. The value added here is forwarded to the sustainability efforts in the form of funding, which increases the overall sustainability performance and allows the company to build up a mature sustainability of AI practice.

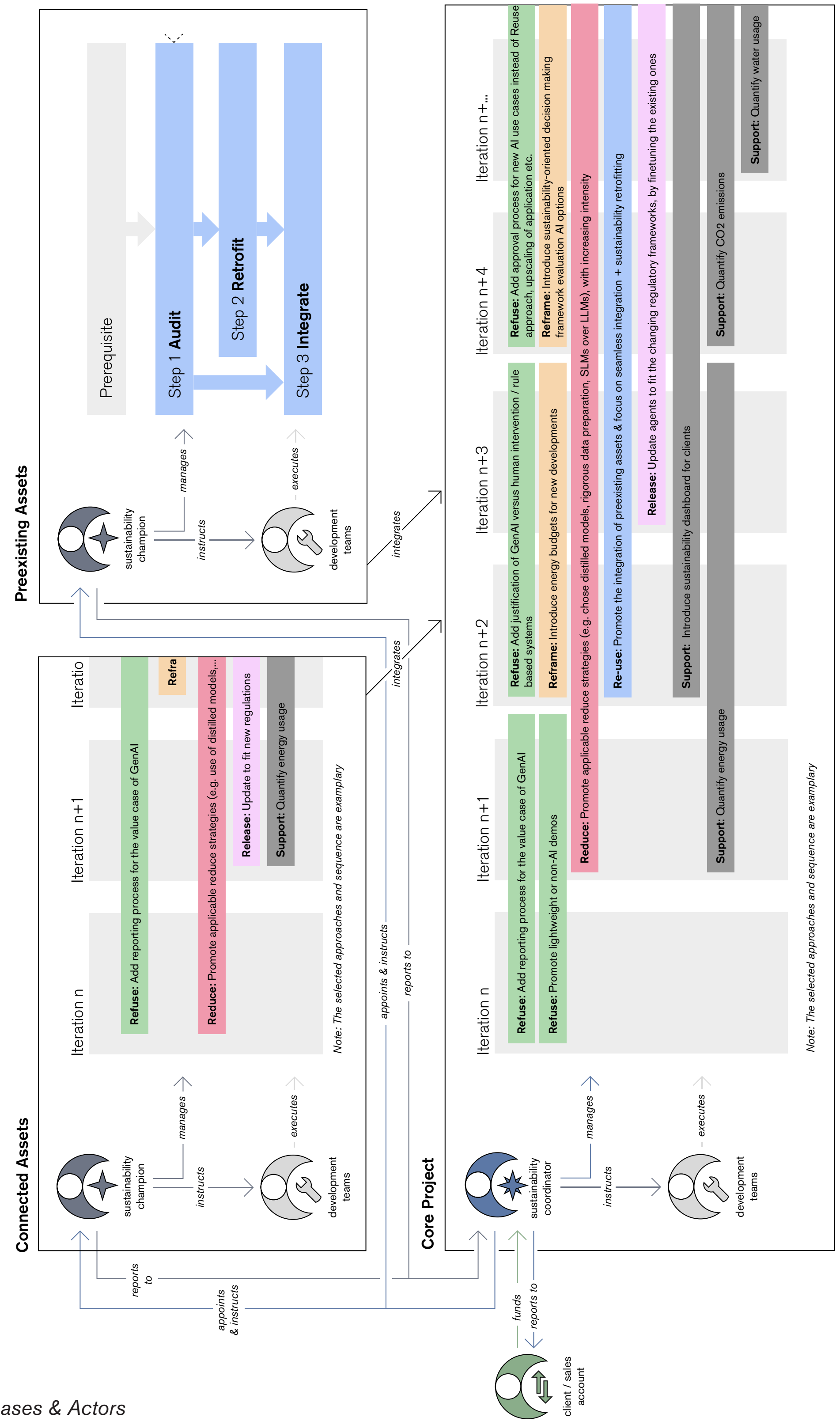

Preexisting Assets
Prerequisite
Step 1 Audit
Step 2 Retrofit
Step 3 Integrate
sustainability champion
manages
instructs
executes
development teams
integrates
Connected Assets
Iteration n
Iteration n+1
Iteratio
Refuse: Add reporting process for the value case of GenAI
Refra
Reduce: Promote applicable reduce strategies (e.g. use of distilled models...
Release: Update to fit new regulations
Support: Quantify energy usage
Note: The selected approaches and sequence are examplary
sustainability champion
manages
instructs
executes
development teams
reports to
appoints & instructs
integrates
Core Project
Iteration n
Iteration n+1
Iteration n+2
Iteration n+3
Iteration n+4
Iteration n+...
Refuse: Add reporting process for the value case of GenAI
Refuse: Promote lightweight or non-AI demos
Reduce: Promote applicable reduce strategies (e.g. chose distilled models, rigorous data preparation, SLMs over LLMs), with increasing intensity
Support: Quantify energy usage
Refuse: Add justification of GenAI versus human intervention / rule based systems
Reframe: Introduce energy budgets for new developments
Re-use: Promote the integration of preexisting assets & focus on seamless integration + sustainability retrofitting
Support: Introduce sustainability dashboard for clients
Release: Update agents to fit the changing regulatory frameworks, by finetuning the existing ones
Refuse: Add approval process for new AI use cases instead of Reuse approach, upscaling of application etc.
Reframe: Introduce sustainability-oriented decision making framework evaluation AI options
Support: Quantify CO2 emissions
Support: Quantify water usage
Note: The selected approaches and sequence are examplary
sustainability coordinator
manages
instructs
executes
development teams
appoints & instructs
reports to
funds
reports to
client / sales account

Figure 16
Blueprint (V1) - Phases & Actors

5.2.3 Deep dive - Layer 2: Process within Iterations

Within each iteration loop in the core project and the connected assets, a process needs to occur, to effectively identify, implement and improve the sustainability approaches suitable. For this, I propose eight steps (see Figure 17):

01 Lifecycle Stage Identification: In not all iterations all phases across a GenAI lifecycle occur or are emphasized. Therefore it is the starting point to identify the lifecycle stages which are relevant to the iteration.

02 Strategy Types: With the help of the proposed framework, the strategy types which are applicable to the relevant lifecycle stages can be identified – as likely only a selection of strategies is applicable for each iteration.

03 Concrete Approaches: The concrete actions occurring in the iteration can be considered when identifying the sustainability approaches to be implemented under the strategy. For example, if a demo will be developed to showcase clients' interactions with the application though a recording, a refuse strategy might be applicable through avoiding the use of LLMs in the backend and instead hard coding the queries that are used to demonstrating the functionality of the application, removing the need for resource-intense AI use while remaining the same functionality. These approaches are highly context and action dependent. Those that prove to be successful can be collected and serve as a toolbox for future use. This process requires a level of creativity from the teams, as the abstract mechanisms presented by the framework, have to be translated and made actionable for the specific context. While this might proof to be challenging, it can yield highly innovative and impactful approaches. The need for creativity gets reduced with increasing experience, as teams might be able to fall back to proven approaches once experienced, but this will present a trade-off regarding the innovativeness. Therefore, I advocate for incentivizing creativity within the various project teams.

04 Evaluation: The identified approaches need to be evaluated based on their feasibility and effectiveness. The most promising ones are then prioritized and selected to be further pursued. This step is conducted by the sustainability coordinator / champion in discussion with the development teams.

05 Implementation: The identified selection of approaches is implemented in the project.

00 Old Approaches: Approaches that have been implemented in previous iterations and proven successful and are carried over into the next iteration.

06 Performance evaluation: It is checked, to what extend the implemented

approaches are fulfillling their intended function and whether unintended side effects occur.

07 Iterations: The approaches are then optimized, if room for improvement is identified and feasible.

08 Reporting: The impacts from both the newly implemented approaches, but also the preexisting once that have been carried over from previous iterations are reported. This consists of a description of the approaches, if possible, a quantification (or estimation) of the impact and if not quantifiable, a description of the positive approaches.

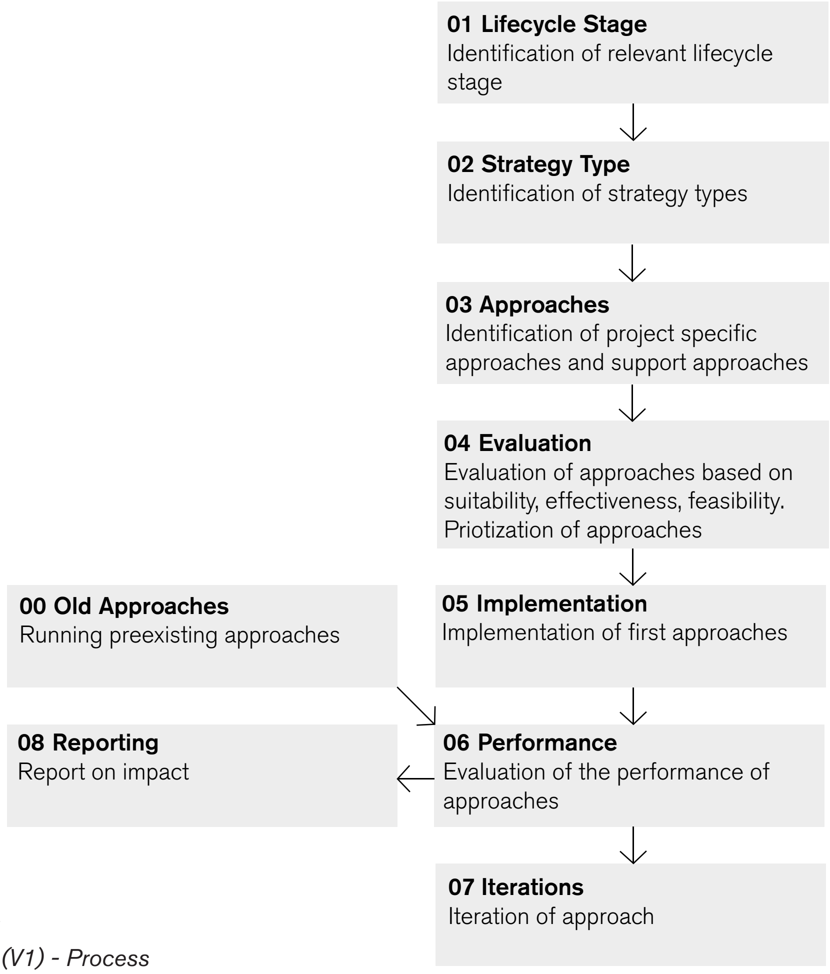

01 Lifecycle Stage
Identification of relevant lifecycle stage
02 Strategy Type
Identification of strategy types
03 Approaches
Identification of project specific approaches and support approaches
04 Evaluation
Evaluation of approaches based on suitability, effectiveness, feasibility. Priotization of approaches
00 Old Approaches
Running preexisting approaches
05 Implementation
Implementation of first approaches
08 Reporting
Report on impact
06 Performance
Evaluation of the performance of approaches
07 Iterations
Iteration of approach

Figure 17
Blueprint (V1) - Process

5.2.4 Deep dive - Layer 3: Materials

To inform the activities conducted by the actors, various materials need to be provided:

Framework: Sustainability Strategies across the GenAI Lifecycle: The previously presented framework and its different elements serve as the key material underlying the different processes. It is used as the basis to identify and formulate project specific sustainability approaches.

Overview of preexisting approaches: This document can be developed over time, once approaches have been conducted and their performance evaluated. It will contain historic approaches and their performance, to provide a knowledge base that can inform future iterations or projects.

Standardized sustainability audit for preexisting assets: To ensure the preexisting assets don't damage the sustainability efforts in the core project, a sustainability audit on preexisting assets can serve as a safeguard. This audit is not existing at this point in time. For the beginning it might consist of an unstandardized examination by a sustainability coordinator / champion but further on it shall become a standardized and rigorous process.

Discussion

The proposal showcases how sustainability can be introduced into the development of GenAI based applications. It distinguishes between core project, connected, and preexisting assets, and by this allows for differentiated interventions - while maintaining coherence across different teams and workstreams. The use of iterative processes enables sustainability measures to evolve alongside the project. Due to the current absence of proven best practices, this blueprint leverages abductive reasoning to make "best predictions" for suitable sustainability approaches. Through iterations, these predictions then get tested and refined or abandoned. This ensures a contextual fit and responsiveness to change (e.g. driven by resource availabilities or client needs). The strategy encourages teams to translate general strategy types into project-specific sustainability approaches, supported through incentives from leadership - this combines bottom-up and top-down dynamics.

However, this design also presents risks: Without experienced sustainability practitioners or willingness from the various actors, early implementations may lack rigor, and the approaches developed may vary in quality and sustainability impact. Furthermore, the dependence on lightweight, low-barrier measures in early phases may delay the adoption of more impactful but resource-intensive interventions. Motivation and pressure need to exist, to go beyond the lowhanging fruit and experiment with more challenging but impactful sustainability approaches.

The integration of sustainability as a differentiator in value propositions is promising yet requires credible evidence of impact. Quantifications of impact can be highly difficult, as a significant part of the environmental impacts occurs upstream in the supply chain (e.g. in cloud providers), which frequently don't provide sufficient reporting on their impacts. Lacking data can be partially replaced by estimations and if this is not feasible, be replaced with qualitative reporting on the sustainability measures in place. With increasing occurrence of new sustainability approaches in science and pressure towards transparency from user side, we can be hopeful that quantification techniques become increasingly accurate. Until then we shall proceed with our best efforts, even if quantification might not be optimal.

5.3 Stakeholder Interviews

Introduction

To understand the validity of the developed sustainability blueprint for the project, interviews with key stakeholders were carried out. The goal was to explore the desirability, feasibility and validity of the approach. For this purpose, semi-structured interviews were conducted.

Given the various perspectives involved in the development process, it was essential to gather diverse perspectives across strategic, managerial, and technical roles. The selection of interview partners included representatives from leadership, portfolio owners, asset owners, project management, and technological development. The selected interview partners were all from the immediate context of the examined project. Hereby the goal was not to derive at a consensus but rather to capture diverse perspectives. This allows to understand viability under the current conditions from diverse views, further it allows to understand where additional support or adaption is required. The result is a grounded, practice-oriented reflection on the proposed strategy in the context of the project and organization.

Method

Five semi-structured interviews were conducted. The selected participants represented a cross-section of the responsibilities present in the project, they consisted of: A representative of leadership, the portfolio owner of the core project and assets, two product owners of assets, and a technology expert. Each interview lasted between 30 and 45 minutes and was conducted individually.

The interviews were guided through a semi structured process with open-ended questions. The process was tailored to gather the interviewee-specific perspective on the project. The general structure of the interview consisted of an introduction into the project, without showing or explaining any materials. This was followed by the first part of the interview, in which the role of "environmental sustainability of GenAI" in the organization and project work was discussed. Here the current maturity of sustainability practices was explored, as well as the current and predicted role the topic will occupy in the organization. After the general part, the general framework as well as the proposal for the project specific strategy were presented by the interviewer and space for questions and clarifications as provided.

Once the interview partners understood the presented materials, the second part of the interview was conducted. Space was provided for voicing overall remarks and comments which was then followed by questions aimed to explore the feasibility and viability, desirability, scalability, and usability of the strategy. Lastly space was provided for closing remarks.

The interviews were recorded with permission to ensure accuracy in the post-processing phase. The transcripts were then coded and grouped into the arising themes: Context, Value, Barriers, Focus, and Setup. Within the themes, identical insights were aggregated and the different types of insights listed. These insights will then be used to inform the next step.

It must be noted that the results represent the personal perspectives of the interviewees and do not represent the voice of the company.

5.3.1 Overall Sentiment

Across the interviews, participants consistently acknowledged the need to address the environmental impacts of GenAI, framing sustainability as an organizational responsibility.

While the proposed blueprint was viewed as theoretically sound, doubts were raised about its practical implementation at the project level. These concerns were primarily linked to the absence of a compelling business case, driven by limited client demand and the current lack of regulatory pressure. Interviewees emphasized that without a clear and direct value proposition for project teams, there would be little incentive to allocate additional resources toward its adoption. Successful implementation was therefore seen to depend on explicitly articulating its added value and institutionalizing its use, for example by linking it to project budgets or embedding it in companywide policy.

Table 6 shows an overview of which topics were mentioned in how many interviews (out of five). Note that this serves to showcase how present which topic was across the interviews and is not an indicator for the relevance of each topic. As general topics are known across the different roles, they score higher, while more specific topics might be only known to some roles, therefore scoring lower. Further these topics are generalizations, the detailed insights are discussed in the following sections.

Theme	*Topic*	*Interview coverage (n=5)*
Added value	Cost savings from reduced resource consumption	5
	Competitive differentiation through standardized sustainability blueprint	3
	Enhanced environmental performance of AI applications (e.g. as moral duty or market positioning)	2
	Anticipation of future regulatory requirements	2
	Increased degree of innovation through higher context specificty	1
Barriers	Low adoption willingness due to limited external pressure	3
	Sustainability concerns deprioritized by competing business drivers (e.g., speed-to-market)	3
	Absence of a clear business case for sustainability adoption	2
	Insufficient GreenAI literacy within workforce	2
	Difficulty in quantifying environmental impacts due to lacking transparency in supply chains	1
Area of Focus	Current prioritization of internal projects due to limited external demand	4
Context	Clients are partly concerned about sustainability performance	2
	Measures for social sustainability of AI are partly preexisting	2
	If there is demand, adoption will follow	2
Blueprint Setup	Clear role division is needed	4
	Clear governance is essential to ensure the blueprint is implemented effectively.	4
	The blueprint has to be actionable	3
	Added value has to be explicit	3
	Early-stage flexibility paired with post-rollout sustainability guardrails	2
	Quantification of impacts is needed (either measured or estimated)	2

Table 6: Interview Topics

5.3.2 Contextual Insights

Context: The regulatory landscape already requires companies to be transparent (e.g. the CSRD), this can extend to the application of GenAI - for this a framework will be required - in order to mitigate the associated environmental risks. This goes in hand with many companies already pursuing sustainability targets which such a framework can support. Along with these goals, some organizations already demand ethical practices from suppliers and partners and are aware about the negative environmental impacts of AI.

Some foundations towards environmentally sustainable GenAI are already in place: While mostly focusing on social sustainability, audits for AI already partly exist - as a response to the EU AI Act. Similar to this, efficiency mechanisms are already used, such as the preference of SMLs over LLMs, motivated by the saving of costs.

Barriers

While the overall context seems promising, various barriers exist which need to be overcome. At the current point in time, sustainability concerns are overshadowed by other drivers, such go-to-market speed, cost-effectiveness and performance gains. Core drivers behind this lack of urgency are the absence of regulatory pressure and business cases powered by the sustainability of AI. Therefore, companies display a low willingness to risk extra costs arising through sustainability measures. Due to the budget constraints other factors (such as functionality), dominate the competition for resources, further heightening skepticism of taking up the added responsibility and workload of sustainability in GenAI projects by workers, such as product owners.

A practical barrier presents the quantification of impacts, as GenAI workloads are mainly conducted through external cloud providers which do not disclose precise impact information. Quantifications are important to make impacts assessments more accurate, improve sustainability measure and heighten the importance of the issue on a corporate and regulatory side.

Lastly, sustainability measures might produce negative side effects from a business perspective, such as creating costs through added development times or losing business by ruling out GenAI use for certain use cases. This can hinder the willingness of companies to pursue sustainability measures.

Focus

From a consulting perspective, sustainability of AI is at the current point in time difficult to sell, as the majority of client organizations do not prioritize this topic yet. Nonetheless, the topic is promising and relevant for the future, therefore there is an internal interest in extending the associated capabilities. A focus on internal projects is therefore the priority. Nonetheless, incorporating sustainability into GenAI offers to clients can offer a competitive edge, especially for large enterprises which are affected by regulations (e.g. the Corporate Sustainable Rep and CSDDD) and those having sustainability targets.

Value Case

Sustainability measures aim to reduce the resource use of applications. Less resources use and higher resource efficiency mean in return cost reductions. Rising costs for using large multipurpose models increase the significance of these cost reductions. Depending on the scale of the application, these reductions can be significant.

Beyond that, corporations pursuing sustainability targets and company values around sustainability benefit from the use of such a framework as it strengthens the credibility of their market positioning. Further some, especially large enterprises, can see it as a moral obligation to incorporate such risk mitigation approaches for environmental impacts of AI use. Improving the sustainability impact of an organization can be seen as both taking responsibility for the own actions and impacts but also be beneficial for public perception.

Another benefit that comes with the uptake of the framework for businesses is the differentiation potential. With the offerings of AI applications being vast, few are transparent about the approaches they take to improve the sustainability performance of their applications.

Next to the business performance, benefits are also on compliance side. While the future of compliance is uncertain, the negative effects resulting from unsustainable practices on humanity are increasing – making stricter regulations in regard to sustainability risks a likely scenario. But also, current regulations, such as the Corporate Social Reporting Directive (CSRD) puts pressure on large enterprises. Companies falling under the CSRD are obliged to conduct a double materiality analysis in which among other risks sustainability impacts are reported upon, for which they have to provide measures to mitigate such risks. With increasing AI adoption, the vast resource demands associated with this technology will likely start to appear in many organizations' double materiality analysis. The adoption of a blueprint, such as the proposed one can serve both as means to increase transparency but also as a risk mitigation effort.

The last value highlighted is the increase of internal technological innovation and ownership. In most cases, sustainability efforts in the development of GenAI-based applications will push towards the replacement of general purpose models for smaller tasks specific models and architectures. The specificity of a model/ application does not only safe resources but also allows to be developed in many cases in-house. This decreases dependance on the vendors of the large multi-purpose models, which can be beneficial in for example price increases. A very significant advantage of this is the potential to increase the protection of the developed application and models through patents, as the developed solutions offer a higher degree of innovation - in contrast to the frequently seen products that are only wrappers to large preexisting GenAI models. While the development of task specific applications might come with higher up-front investments, the operating costs are lowered through reduced resource consumption and entry-barriers are raised for competitors.

5.3.3 Insights on the set-up of the blueprint

Early on in a project, a mostly unrestricted space needs to exist, in which room for experimentation and try-outs is provided. This is necessary to not suffocate potentially valuable ideas. On the other end of the project, once rolled out and getting scaled, the grip is lost to introduce sustainability measures. To ensure roll-out and scaling is happening without unintended environmental side-effects, fixed sustainability guardrails need to be in place that get followed. Therefore, the identification of approaches and the step-by-step increasing implementation of approaches - as proposed in the previously discussed blueprint - needs to occur in between the described to phases, in a third phase. Additionally, the sufficiency of the defined guardrails must be assessed before the roll-out, to ensure that the environmental risks are adequately mitigated.

To ensure that the blueprint is consistently used within the examined context, it must be made mandatory, following a top-down approach. This can be achieved via an audit before the roll-out, in which the sustainability impact and related risks are examined (e.g. via using existing frameworks such as the software carbon intensity (SCI)) and suitable risk mitigation measures, in the form of sustainability guardrails, must be presented. If not sufficient, no go-decision is provided. To integrate the blueprint into the internal compliance structure, leadership support is required. Beyond making it compulsory, the budget allocation needs to be interwoven. To not punish teams that perform well from a sustainability perspective (and by this reduce costs of resources and computation) the saved budget shall be kept within the team for them to reinvest (e.g. to finance the extra development costs that might arise). Additionally, the blueprint should establish clear thresholds for when and where it is applied, focusing on meaningful use cases.

To ensure actionability role clarity is critical. A likely division of roles is that the product owner is responsible for managing the different approaches, identifying risks and proposing sustainability guardrails for roll-off. To assist, a Center of Excellence provides guidance and evaluates risks and sustainability guardrails. The workforce involved in the development of the product executes the approaches under guidance of the product owner. All of them must be provided with specific actions and ideally tracked with KPIs to track the performance. Complexity should be minimized to lower entry barriers and encourage participation.

Regarding the overall impact of the blueprint, quantification is important: For this, develop sound quantification methods to credibly estimate financial and environmental impacts. This allows to convince stakeholders of the value added by the blueprint. Further it increases transparency to reduce regulatory hurdles and can be used to communicate it to clients and externals.

Discussion

The conducted interviews provided valuable insights into the relevance, applicability, and potential impact of the proposed sustainability blueprint for GenAI projects. The diversity of perspectives, spanning leadership, portfolio ownership, product management, and technical expertise, enriched the evaluation by highlighting both alignment and tension points across organizational levels.

While sustainability concerns were overall regarded as valid and important, it is yet far away from a priority. Other challenges, like speed-to-market are taking the central stage, which not only take away attention from sustainable practices but often even conflict. With little pressure coming from external actors such as clients and regulatory entities, motivation has to be created internally. While this can also mean fostering grass-root initiatives, it mostly means top-down directions. The mandatory nature of such a blueprint and its interference with running processes likely leads to negative sentiment in the project teams and therefore decreases motivation. In order to minimize the arising negative sentiment and interference in preexisting processes, the blueprint must provide teams with a sufficient degree of freedom and be minimally invasive. Further the created value must be clear to the involved stakeholders (e.g. by saving budgets). To close potential loopholes, the blueprint must be simple, unambiguous and consequent.

The interviews were highly helpful, as they enriched the blueprint with human perspectives, highlighting concerns and opportunities. The derived insights will be used to refine the blueprint in the next step.

While being beneficial, the chosen setup has several limitations. The sample size of five is highly limited, therefore the findings can only be seen as explorations of the immediate context and not as scientifically based truth. Further all interview partners are located in the immediate and neighboring context of the examined project, while this allows to create a rich perspective around the project in question, it limits the generalizability of the findings. The arising findings present personal perspectives and are not the voice of the company in which the interviews were conducted. Moreover, the interviews occurred at an early stage of blueprint development, meaning that feedback primarily reflects anticipated rather than experienced implementation challenges. Actual adoption, integration into business processes, and long-term impacts remain untested at this stage.

5.4 Revision of the Blueprint

Introduction

This chapter presents a revised environmental sustainability blueprint for GenAI based projects. The revision was conducted in response to the results of stakeholder interviews, build on the previously presented version (see chapter 5.2). The aim of this revision is to further sharpen the blueprint to represent the organizational realities, barriers and enables. While the goal is not to develop a ready-to-implement-blueprint, this chapter aims to showcase, how the developed framework can be applied in practice. Despite the explorative goals, the resulting blueprint aims to be actionable and provide a starting point, for companies to derive their own governance models for the environmental sustainability of AI. For this, it examines what is needed to implement sufficient sustainability measures within a real organization and context.

Method

The revised sustainability blueprint was developed from adapting to initially proposed blueprint to the empirical findings derived from the stakeholder interviews. The initial blueprint served as the conceptual foundation. Against its structure and content, the interview findings, clustered into the themes “context”, “values”, “barriers”, “focus” and “set-up”, were systematically mapped. Where gaps, tensions, or opportunities were identified, and targeted revisions were made. This included the restructuring of roles, refinement of the process phases and integration of governance checkpoints and addition of the value case.

The approach aims to ensure contextual fit, improved usability for the targeted organization and increased likelihood of adoption. Further, this methodology ensured that the revised blueprint is not only theoretically robust but also grounded in organizational reality.

5.4.1 Changes to the previous blueprint

Various changes are made to the blueprint, informed by the conducted expert interviews. The central changes made are the following:

The roles got changed and extended, from more individual roles to functional roles. What was previously the sustainability champion, sustainability coordinator and development teams, is now fused into a single group, the operational role. This is because they are all involved in the execution and field work. While project owners might maintain the same responsibilities as the previously described sustainability champions and sustainability coordinators, they are all involved in the execution and therefore grouped in the scope of this governance oriented blueprint. The knowledge and resources are now provided by a different role, the advisory role, in this case a center of excellence, to reduce the workload of project owners and increase sustainability of GenAI literacy in the overall organization, not isolated in projects. Lastly, a steering and decision making role is added, external of the operational role, this is necessary, to reduce bias and ensure a more rigorous execution.

Another change is the division of the workflow into different key phases with varying levels of governance and sustainability measures. The interviews showed that more granularity was required, especially in the case such a blueprint becomes mandatory. The different phases allow for various levels of freedom and ensure to maintain the capability to innovate, an explorative phase to develop project specific sustainability approaches, a central decision making stage to ensure that the blueprint is executed as ordered and a strictly planned roll-off and scale-up to ensure mitigation of impacts throughout the project life. The previous workflow is now part of the proof of concept stage.

The addition of new phases also required the detailing of actions to be performed by the roles across the stage. The actions listed in the process of iterations in the old blueprint is incorporated in the actions from the operating role in the proof of concept stage. For other roles and phases, new actions were added.

The old blueprint differentiated between three types of workflows (core project, preexisting asset and connected asset). This division has been removed to increase simplicity and doubling. The different sustainability measures which were presented in the three workstreams are now part of the central project process. For example, the uptake of problematic, preexisting assets would require the operational teams to increase the intensity of sustainability approaches in order to pass the newly introduced Go-/No-Go stage, therefore they would need to introduce new measures to the asset (previously called retrofit) or abandon the asset.

Additionally, values have been added, to make it clear to all those involved, on why the uptake of the blueprint is beneficial, and how the blueprint can be leveraged to decrease sustainability impact while creating business value.

Lastly, wordings have been changed to fit the industry language, such as the rename of iteration to sprint.

5.4.1 The revised blueprint

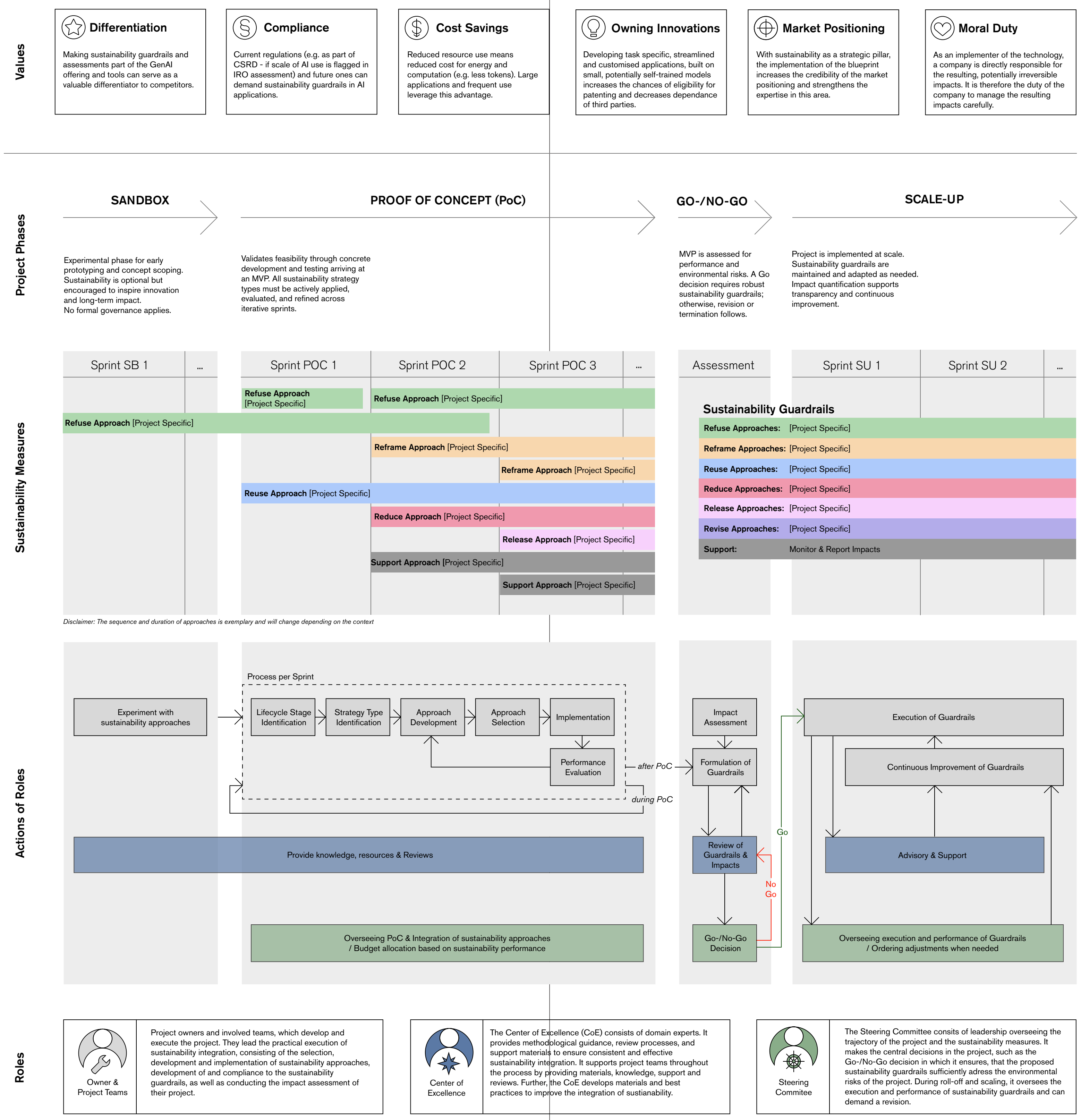
The revised blueprint consists of various layers (see Figure 18). To highlight the value the blueprint adds, the outcomes are arranged on the top, while the more concrete elements of it are further down. In the case of implementation, the blueprint should be read from bottom up. The foundation at the bottom consists of the role division which needs to be made. The next higher levels are the concrete actions the roles must execute, followed by the resulting sustainability measures in the four phases. The highest level of hierarchy consists of the value added through the execution of the sustainability measures and the adoption of the blueprint.

The core structure is oriented on the key phases of a project, starting from the early stages on the left of the blueprint, to the scale-up of a project on the right of the blueprint.

Phases

The central element of the blueprint are the four identified project phases and the corresponding sustainability measures (see Figure 19). The stages are a sandbox phase, a proof of concept phase, a go-/no-go phase and a scale-up phase.

Sandbox:
The Sandbox describes the first set of iterations and sprints in a project. This phase provides space for experimentation, prototyping and try-outs. These processes are highly important for driving innovation, getting a better understanding of the concept and early trial and error to scope the project. To not hinder these processes, and as only little resources are consumed in this phase, no sustainability measures are mandatory. Still, it is encouraged to consider how sustainability can be integrated, as early adoption of such measures is beneficial for the overall impact of the project. Further, sustainability perspectives can also drive innovation, for example by reducing dependence on large, general-purpose models (such as GPTs) and instead promoting more specific, fit-for-purpose, custom architectures and smaller models. The extent of which sustainability measures are integrated in the sandbox phase is up to the project team and depending on the context. No governance structure regulates it. Where the Sandbox phase ends, and the Proof of Concept (PoC) phase begins is up to the project team. The only condition is that by the end of the PoC phase, all required conditions are met. Depending on the context (e.g. the resources available or the timeline) this can allow for a more extended sandbox or for an early start of the PoC phase.



Values

Differentiation
Making sustainability guardrails and assessments part of the GenAI offering and tools can serve as a valuable differentiator to competitors.

Compliance
Current regulations (e.g. as part of CSRD - if scale of AI use is flagged in IRO assessment) and future ones can demand sustainability guardrails in AI applications.

Cost Savings
Reduced resource use means reduced cost for energy and computation (e.g. less tokens). Large applications and frequent use leverage this advantage.

Owning Innovations
Developing task specific, streamlined and customised applications, built on small, potentially self-trained models increases the chances of eligibility for patenting and decreases dependance of third parties.

Market Positioning
With sustainability as a strategic pillar, the implementation of the blueprint increases the credibility of the market positioning and strengthens the expertise in this area.

Moral Duty
As an implementer of the technology, a company is directly responsible for the resulting, potentially irreversible impacts. It is therefore the duty of the company to manage the resulting impacts carefully.

Project Phases

SANDBOX
Experimental phase for early prototyping and concept scoping. Sustainability is optional but encouraged to inspire innovation and long-term impact.
No formal governance applies.

PROOF OF CONCEPT (PoC)
Validates feasibility through concrete development and testing arriving at an MVP. All sustainability strategy types must be actively applied, evaluated, and refined across iterative sprints.

GO-/NO-GO
MVP is assessed for performance and environmental risks. A Go decision requires robust sustainability guardrails; otherwise, revision or termination follows.

SCALE-UP
Project is implemented at scale. Sustainability guardrails are maintained and adapted as needed. Impact quantification supports transparency and continuous improvement.

Sustainability Measures

Refuse Approach [Project Specific]
Refuse Approach [Project Specific]
Refuse Approach [Project Specific]
Reframe Approach [Project Specific]
Reframe Approach [Project Specific]
Reuse Approach [Project Specific]
Reduce Approach [Project Specific]
Release Approach [Project Specific]
Support Approach [Project Specific]
Support Approach [Project Specific]

Sustainability Guardrails

Refuse Approaches:	[Project Specific]
Reframe Approaches:	[Project Specific]
Reuse Approaches:	[Project Specific]
Reduce Approaches:	[Project Specific]
Release Approaches:	[Project Specific]
Revise Approaches:	[Project Specific]
Support:	Monitor & Report Impacts

Disclaimer: The sequence and duration of approaches is exemplary and will change depending on the context

Actions of Roles

Experiment with sustainability approaches
Process per Sprint
Lifecycle Stage Identification
Strategy Type Identification
Approach Development
Approach Selection
Implementation
Performance Evaluation
after PoC
during PoC
Provide knowledge, resources & Reviews
Overseeing PoC & Integration of sustainability approaches / Budget allocation based on sustainability performance
Impact Assessment
Formulation of Guardrails
Review of Guardrails & Impacts
Go-/No-Go Decision
Go
No Go
Execution of Guardrails
Continuous Improvement of Guardrails
Advisory & Support
Overseeing execution and performance of Guardrails / Ordering adjustments when needed

Roles

Owner & Project Teams
Project owners and involved teams, which develop and execute the project. They lead the practical execution of sustainability integration, consisting of the selection, development and implementation of sustainability approaches, development of and compliance to the sustainability guardrails, as well as conducting the impact assessment of their project.

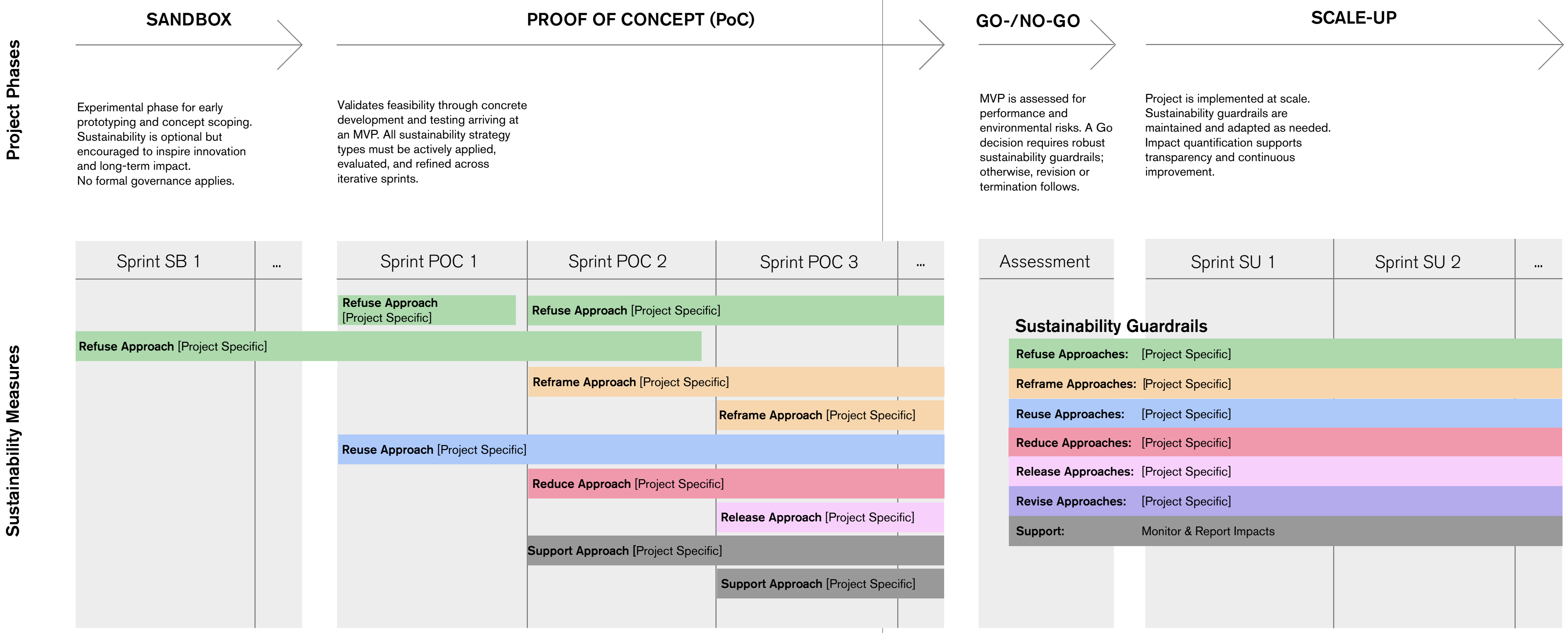
Center of Excellence
The Center of Excellence (CoE) consists of domain experts. It provides methodological guidance, review processes, and support materials to ensure consistent and effective sustainability integration. It supports project teams throughout the process by providing materials, knowledge, support and reviews. Further, the CoE develops materials and best practices to improve the integration of sustianability.

Steering Commitee
The Steering Committee consits of leadership overseeing the trajectory of the project and the sustainability measures. It makes the central decisions in the project, such as the Go-/No-Go decision in which it ensures, that the proposed sustainability guardrails sufficiently adress the environmental risks of the project. During roll-off and scaling, it oversees the execution and performance of sustainability guardrails and can demand a revision.

Figure 18: Revised Blueprint

Proof of Concept (PoC):
In the Proof of Concept phase the feasibility and viability of the application concept gets validated. The need is analyzed, the concept defined, and solutions developed. The solution is prototyped and the prototype tested. The PoC phase ends with a minimum viable product (MVP). In this phase various sprints are conducted getting increasingly concrete. Within each sprint, different lifecycle phases are in focus, depending on the stage of the PoC (e.g. the “business understanding” phase will be in focus in early stages, such as when defining the needs the project aims to fulfilll and phases such as “model selection” will be more prominent during prototyping). Throughout the PoC, all seven types of sustainability strategies must be considered (refuse, reframe, reuse, reduce, release, revise and support). Within each sprint of the PoC, the lifecycle phases in focus are identified, and - based on the previously presented framework - the corresponding strategy types are selected. The mechanisms of the suitable strategies are then applied to the action space of the sprint, and through this, actionable and specific sustainability approaches are formulated. This can be a highly creative process, which might lie outside of the usual operating space of some; It is therefore desirable to pay special attention to encouraging those involved in engaging in this activity. The identified approaches are then evaluated based on their suitability and effectiveness and a selection is made. This selection depends on the resources available and the overall context. The selected approaches are then implemented. Once installed, they can be kept in place throughout the following sprints of the PoC, replaced with new approaches that might yield better impacts or abandoned, in the case they are deemed unusable. Throughout the sprints the different sustainability strategy types will be applied and at the end of the PoC, a selection of suitable and effective approaches has been identified and tested. This selection of project specific sustainability approaches across the different sustainability strategy types is a requirement that teams need to fulfilll.

Go-No-Go Point:
Before widespread implementation and the scaling of the application, the MVP gets assessed across various dimensions. Depending on the MVPs performance, a Go- or a No-Go-decision is made, meaning that satisfactory


Project Phases
Sustainability Measures
SANDBOX
PROOF OF CONCEPT (PoC)
GO-/NO-GO
SCALE-UP
Experimental phase for early prototyping and concept scoping. Sustainability is optional but encouraged to inspire innovation and long-term impact. No formal governance applies.
Validates feasibility through concrete development and testing arriving at an MVP. All sustainability strategy types must be actively applied, evaluated, and refined across iterative sprints.
MVP is assessed for performance and environmental risks. A Go decision requires robust sustainability guardrails; otherwise, revision or termination follows.
Project is implemented at scale. Sustainability guardrails are maintained and adapted as needed. Impact quantification supports transparency and continuous improvement.
Sprint SB 1
...
Sprint POC 1
Sprint POC 2
Sprint POC 3
...
Assessment
Sprint SU 1
Sprint SU 2
...
Refuse Approach [Project Specific]
Refuse Approach [Project Specific]
Refuse Approach [Project Specific]
Reframe Approach [Project Specific]
Reframe Approach [Project Specific]
Reuse Approach [Project Specific]
Reduce Approach [Project Specific]
Release Approach [Project Specific]
Support Approach [Project Specific]
Support Approach [Project Specific]
Sustainability Guardrails
Refuse Approaches: [Project Specific]
Reframe Approaches: [Project Specific]
Reuse Approaches: [Project Specific]
Reduce Approaches: [Project Specific]
Release Approaches: [Project Specific]
Revise Approaches: [Project Specific]
Support: Monitor & Report Impacts

Disclaimer: The sequence and duration of approaches is exemplary and will change depending on the context

Figure 19: Revised Blueprint - Project Phases & Measures

performance of the MVP allows for the project to get scaled up, and unsatisfactory performance will mean that further iterations are required before scaling or the project being abandoned. As part of the sustainability blueprint, the project gets analyzed for the risks it poses to the environment (such as high carbon emissions through extensive energy use, contributing to climate change). For the identified risks, appropriate risk mitigation measures must be presented – in this case, sustainability guardrails for scaling the project. The guardrails consist of the selection of identified and tested sustainability approaches (across the different strategy types such as refuse, reframe etc.) from the PoC. If the guardrails are deemed sufficient, the project can be continued from a sustainability standpoint. If the guardrails are deemed insufficient, they must be revised, or the project is abandoned.

Scale-Up:
After the Go-decision has been given, the project can be implemented and scaled up. Throughout this process, the previously formulated sustainability guardrails are installed maintained. Along the process of scaling the application, unforeseen circumstances can occur, for example updates of frameworks on which the application is built upon, such as regulatory frameworks. In the case of unforeseen developments, the existing sustainability guardrails might not be sufficient, and new sustainability approaches might need to be added. Such as releasing a new version, able to access the new frameworks, instead of risking the loss of functionality and thereby creating the need for a completely new application, which would once again consume additional resources. Quantification of impacts as part of the sustainability guardrails along the scaling process can indicate the overall sustainability performance of the project, which can then be used to a) report on the impacts in order to create transparency or compliance and b) be used to inform future sustainability measures for projects.

Roles

To ensure that the sustainability measures are executed along the various phases, dedicated roles are appointed (see table 7). The minimum of roles involved are three types. An operational role, executing the project and sustainability approaches, an advisory role, providing knowledge and support to empower the operational role to manage the sustainability approaches and lastly a leadership role, involved in the central decision making and steering. The division is required, as the groundwork is being conducted by a workforce that lacks literacy of sustainability measures for the project at the current time. This literacy is provided through the advisory role in which the domain knowledge is connected. The quality and rigor of output produced by the operational role and informed by an advisory role cannot be ensured within the two, to ensure impartiality. A third role is needed to make the go-decisions and provide oversight and steering. The three roles are divided as followed:

Owners and project teams (operational): They develop and execute the project and sustainability approaches. In their current role, they already serve as project owners and teams in which they develop and deploy GenAI-based projects. Under the proposed governance blueprint, they additionally develop and implement sustainability measures “in action”. This connection of “business-as-usual” project work in combination with the sustainability measures, ensures increasing awareness and literacy on the topic, which allows building a widespread and mature “sustainability of AI” capability within the organization. They will likely display hesitation to adopt to the blueprint, as it will require them to perform extra tasks, therefore sustainability measures need to be obligatory and incentivized. At the current point in time, this role does not

Owner & Project Teams	Project owners and involved teams, which develop and execute the project. They lead the practical execution of sustainability integration, consisting of the selection, development and implementation of sustainability approaches, development of and compliance to the sustainability guardrails, as well as conducting the impact assessment of their project.
Center of Excellence	The Center of Excellence (CoE) consists of domain experts. It provides methodological guidance, review processes, and support materials to ensure consistent and effective sustainability integration. It supports project teams throughout the process by providing materials, knowledge, support and reviews. Further, the CoE develops materials and best practices to improve the integration of sustianability.
Steering Commitee	The Steering Committee consits of leadership overseeing the trajectory of the project and the sustainability measures. It makes the central decisions in the project, such as the Go-/No-Go decision in which it ensures, that the proposed sustainability guardrails sufficiently adress the environmental risks of the project. During roll-off and scaling, it oversees the execution and performance of sustainability guardrails and can demand a revision.

Table 7: Revised Blueprint - Roles

possess sufficient sustainability of AI literacy. Therefore, they need to gradually construct this literacy over time by being pressured through governance and empowered through a sufficient resources and knowledge.

Center of Excellence (advisory): The Center of Excellence (CoE) is an entity consisting of domain experts. Its role is both the creation of new material, tools and knowledge on the topic of AI for sustainability and the empowerment of other internal actors to act in alignment to that topic. They aim to fill the knowledge deficits of the operational teams by both advising them and reviewing their work. The CoE is not involved in the active execution of any tasks. In this case, they serve as owners of the sustainability framework on which the blueprint is based. They continuously improve their own knowledge, skills and resources and strive to uncover and establish best practices within their domain. While they may provide suggestions for decision-making, the actual steering and decision making is made in the third group.

Steering committee (leadership): Composing of leadership, this entity serves to provide oversight and decision making. Through the steering committee, the sustainability blueprint is enforced throughout the project. The steering committee gets involved from the PoC until the scale-up of the project. Throughout the process it makes the key decisions (such as the Go-No/Go decision), steers the overall trajectory and manages the budgets. Further it has the ability to intervene if the sustainability measures do not yield sufficient impact and effectiveness.

For maximal impact, a fourth role should be introduced – auditors. Due to the early stage of this capability and prevailing absence of regulations and reporting requirements, this role is not specified in the blueprint, to reduce complexity. Ideally internal and / or external auditors would be needed to ensure robust reporting end rigorous execution of sustainability measures and guardrails. To do that, they would conduct various audits along the project. The results of their audits would then be used to a) ensure a rigorous application of the blueprint, b) improvement to the blueprint and specific measures over time and c) report the created impacts for compliance and communication purposes. As this extra line of protection adds complexity beyond the scope of this project and the current state of maturity, it is disregarded from here onwards.

Actions of Roles

Along the different phases, the roles are required to perform different actions (see Figure 20), it is in their responsibility to execute these actions, and their performance can be measured with KPIs. In the case of the project owner and project teams, the ultimate responsibility lies with the project owner.

Sandbox:
The sandbox serves as an experimental space for the operational teams. Here the blueprint does not dictate any actions from the owners and project teams. Nonetheless they are encouraged to experiment with sustainability approaches. To make this possible, the CoE provides knowledge and resources (such as the framework) and is there to review any sustainability work being done. This assistance is not mandatory for the owners / project teams. To ensure the freedom to experiment and as no mandatory actions or standards are required, the steering committee is not involved here.

Proof of Concept:
The PoC consist of various sprints with increasingly more concrete results, until an MVP is achieved. Within each sprint, the owners and project teams will identify, implement and evaluate one or more sustainability approaches. For this, they undergo a process consisting of the following actions. First, the relevant lifecycle phases (from the framework) for the upcoming sprint should be identified. Depending on the stage, this can be a single one (e.g. early on there will likely be a focus on the business understanding phase) or various (e.g. during prototyping many phases are addressed, such as the phases around data collection and preparation and model selection, training and adaptation as well as deployment). Depending on the selected phases, the applicable strategy types can then be derived from the framework (see chapter 4). The underlying mechanisms of the applicable strategy types are then applied to the actions planned in the upcoming sprint. While this requires a degree of creativity, the owner and project team can thereby identify and develop suitable sustainability approaches for the upcoming sprint. The approaches are then selected based on the available resources and predicted effectiveness. Those that get selected are then implemented and their performance is evaluated. In case the evaluation presents room for improvement, the approaches get iterated. This process is repeated with the next sprint, until the PoC is completed. Throughout the PoC, all levels of sustainability strategies are therefore considered. To ensure, that the owners and teams are empowered to execute this process, the CoE will exist in the same manner as in the sandbox stage. In the PoC, the steering committee oversees the development of the project and the integration of sustainability measures. Its task is to ensure, that the owner and project teams execute the blueprint and implement sufficient approaches. As a mean to ensure this, they can manage the budgets of the teams. If a team showcases a satisfactory sustainability performance, they are more likely to achieve the overall values that comes with the blueprint and are therefore entitled to further budget for development, while underperforming projects will not yield these results and therefore face restricted budgets.

Go-/No-Go:
The owner and project teams perform an impact assessment, to estimate the environmental impacts their project has on both a positive and a negative side. Beyond that, they will create a formation of sustainability approaches which will serve as sustainability guardrails for the roll-off and scale-up of the project. The extent of these guardrails depends on the level of impact (projects with large, estimated impacts will require a rigorous and extensive guardrails compared to a project with smaller estimated impacts). The CoE reviews both impact assessment and the formulated sustainability guardrails. Iterations are conducted by the owner and project teams until agreement is reached. The guardrails and impact assessment are then presented to the steering committee, which will derive at a Go-/No-Go decision. In case of a No-Go, the guardrails are reviewed again by the CoE and iterated by the owner and project teams. If a Go decision is given, the project is ready for roll-off (from a sustainability perspective).

Scale-up:
During roll-off and scale-up, the owner and project teams execute the sustainability guardrails. During this, the CoE provides support and advisory, to ensure this process happens to plan. The steering committee oversees the execution and performance of the guardrails. If needed (e.g. due to unexpected changes in the context), the steering committee can order adjustments to the guardrails. Based on this, the project owner and project team are improving the guardrails.

Created Value

As derived from the interviews, such a blueprint can add value across various dimensions to the organization (see Figure 21). These values are needed to convince the involved stakeholders about the relevance of the blueprint. While the improved sustainability is the central value from the perspective

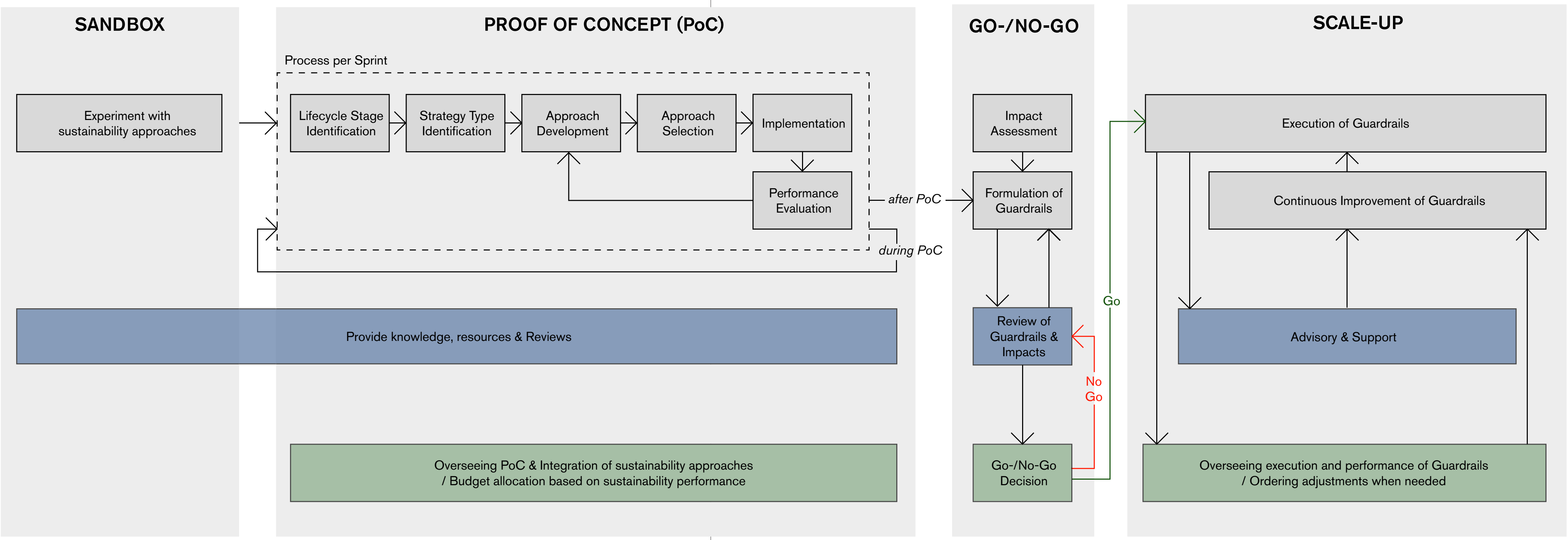
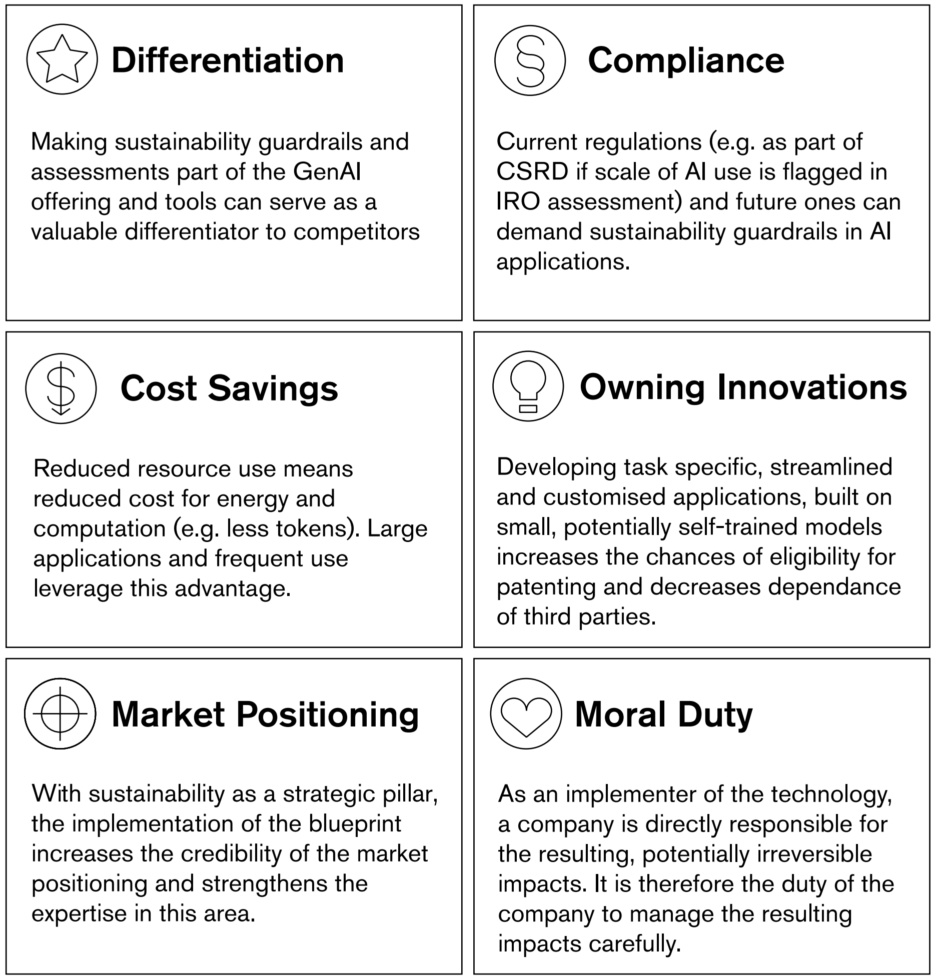

SANDBOX
PROOF OF CONCEPT (PoC)
GO-/NO-GO
SCALE-UP
Process per Sprint
Experiment with sustainability approaches
Lifecycle Stage Identification
Strategy Type Identification
Approach Development
Approach Selection
Implementation
Performance Evaluation
after PoC
during PoC
Impact Assessment
Formulation of Guardrails
Review of Guardrails & Impacts
Go-/No-Go Decision
Go
No Go
Execution of Guardrails
Continuous Improvement of Guardrails
Advisory & Support
Provide knowledge, resources & Reviews
Overseeing PoC & Integration of sustainability approaches / Budget allocation based on sustainability performance
Overseeing execution and performance of Guardrails / Ordering adjustments when needed

Figure 20: Revised Blueprint - Actions

of this project, benefits from a business perspective drastically increase the likelihood of adoption and to justify the additional expenses, resulting from the uptake of the blueprint.



Differentiation

Making sustainability guardrails and assessments part of the GenAI offering and tools can serve as a valuable differentiator to competitors

Compliance

Current regulations (e.g. as part of CSRD if scale of AI use is flagged in IRO assessment) and future ones can demand sustainability guardrails in AI applications.

Cost Savings

Reduced resource use means reduced cost for energy and computation (e.g. less tokens). Large applications and frequent use leverage this advantage.

Owning Innovations

Developing task specific, streamlined and customised applications, built on small, potentially self-trained models increases the chances of eligibility for patenting and decreases dependance of third parties.

Market Positioning

With sustainability as a strategic pillar, the implementation of the blueprint increases the credibility of the market positioning and strengthens the expertise in this area.

Moral Duty

As an implementer of the technology, a company is directly responsible for the resulting, potentially irreversible impacts. It is therefore the duty of the company to manage the resulting impacts carefully.

Figure 21: Revised Blueprint - Actions

Discussion

The proposed blueprint was developed in the context of a large professional services firm. It is therefore built onto the structures of such an enterprise which constraints its applicability to other contexts. Nonetheless, measures have been taken to ensure the generalizability to a certain extent, by highly abstracting the project and roles. It is therefore able to be populated by various configurations of individual roles, depending on their presence in the organization (e.g. some companies might have a dedicated model risk management, AI Governance teams or a Center of Excellence on sustainable IT overall). Beyond the presence of individual roles, some organization might already have preexisting structures for AI governance (e.g. for social sustainability), depending on the maturity of this instance, the environmental sustainability can be directly incorporated into these preexisting structures. In that case, the present blueprint will likely not be applicable to a full extent.

The current framing of the blueprint focuses on enforcing project teams to incorporate environmental sustainability considerations. Depending on the culture and type of organizations, providing positive incentives for the teams might provide a smoother entry or are valuable additions to the proposed blueprint.

Over time, specific sustainability approaches might prove to be effective, decreasing the need for the creative process needed in the PoC. It is important to keep the pressure up towards consistently identifying new approaches and experimenting, to not freeze in a suboptimal status quo. Nonetheless proven approaches are valuable, decrease workload, development costs and risks, and therefore should be integrated over time.

As we will likely see AI applications moving away from being built on single, large multipurpose GenAI models, towards more intricate and streamlined architectures, it is important to highlight, that the proposed blueprint does not just apply to the traditional GenAI models, but also to other AI/ML models utilized in the applications. The scope of GenAI was chosen, as they present the largest, most energy hungry models at the current point in time, with the largest need to be addressed. Only specific sustainability approaches derived from the blueprint might be exclusive to model types, but the overall structure caters for the full variety.

The blueprint demonstrated how the framework can serve as a foundation to develop practical sustainability measures.

Limitations

While a benefit of the blueprint is the saving of costs, due to decreased resource use, the deployment of the blueprint might come with a high effort. Knowledge and materials need to be created, and additional development time spent. This can pose a problem in the case of low willingness to invest, lack of resources or small scale applications. The larger the applications in scope, the larger the cost savings of resources and thus the smaller the issue over time.

Further, this research was created by an individual researcher who previously developed the framework mentioned above. This creates a bias towards to suitability of the framework. Ideally triangulation would be needed and other frameworks compared. Additionally, the blueprint is based in one limited field research and five expert interviews. The constrained extent of the research therefore does not allow to frame this a scientific truth, rather it should be seen as an explorative study.

Future directions

For future work, I advise to start making the blueprint highly specific for different organizations and contexts and adopt it. Further I propose to introduce standardized quantification mechanisms of environmental impacts across all AI applications, to increase awareness and transparency. To execute this, organizations should collectively put pressure on their vendors to release precise information on energy consumption and other impacts (e. g. water use), and until this is sufficient to make measured quantifications, I advise to develop and introduce standardized estimation methods as a transitional solution.

Beyond that, it is highly valuable to research future directions for regulations and policies, to ensure consequent adoption of AI considerations across all organizations and suppliers. While this cannot be fully ensured from a research perspective, it increases the likelihood and feasibilities for such policies and regulations to be introduced across regional or global markets.

Lastly, I advocate for identifying means to consistently embed environmental considerations upfront in the deployment of GenAI and AI instead of them appearing as an afterthought or being bypassed fully. While this body of research already aims at doing this, it can be only seen as a starting point.

Closing Remarks

This thesis contributes to the emerging discourse on sustainable AI by introducing a structured framework for environmental sustainability in the development of GenAI based applications. By adapting principles from circular economy - such as the R-strategies - the framework extends sustainability thinking beyond energy efficiency, toward a more holistic and overarching approach, built on an exhaustive collection of strategy types, beyond efficiency measures.

6.1 Reflection on Research Questions

All five research questions have been answered across the various phases of this research:

RQ1: *Which lifecycle stages does a GenAI model experience?*

RQ1 has been answered by presenting an adapted CRISP-DM lifecycle model in chapter 4.1, which divides the GenAI lifecycle into the stages: Business understanding, Data Collection, Data Understanding, Data Preparation, Model Selection and Training, Adaptation (Fine tuning), Prompt Engineering, Documentation, Evaluation and Risk Assessment, Deployment and Monitoring. Within the chapter, the characteristics and activities of the stages were presented, which were used to inform further work.

RQ2: *Which sustainability strategy types can be applied to the GenAI lifecycle?*

In total, seven strategies have been identified throughout chapter 4, which can be applied across the GenAI lifecycle. The strategies have been named: Refuse, Reframe, Reuse, Reduce, Release, Revise and Support. The 6R strategies all have a directly positive environmental impact, while the support strategy increases the adoption rate of the R strategies and therefore has an indirect impact. Within the framework produced by chapter 4, these strategies have been described, and examples have been provided.

RQ3: *In which lifecycle stages can which sustainability strategies be applied?*

Across chapter 4, the lifecycle stages have been connected to the sustainability strategies, by comparing the activities within each phase to the mechanisms which underly the strategy types. If the mechanism can be applied to an activity, the strategy type of the mechanism is applicable to the corresponding lifecycle stage.

RQ4: *How mature is the scientific research landscape around the sustainability strategy types for GenAI?*

A scoping study was conducted in chapter 4.3 which identified and allocated sustainability approaches present in peer-reviewed articles. This study showcased, that vast differences are present in regard to the number of existing approaches per strategy and lifecycle stage. While some strategy types, such as Reduce-strategies, contain a high number of approaches, others, such as

Refuse-strategies are barely explored. This showcases significant research gaps at the current point in time.

RQ5: *How can the knowledge collected from RQ1 to RQ4 be applied to industry practice?*

Lastly, chapter 5 showcased, how the framework can support the formulation of practical sustainability measures, in the form of a blueprint. The developed governance blueprint utilized the framework to identify and establish suitable sustainability approaches per project, to ensure that the created sustainability impacts are sufficiently mitigated. It utilizes various roles to effectively execute the sustainability measures, increase literacy on the topic in the workforce, as well as establish control mechanisms to ensure impactful execution.

6.2 Contribution

This thesis makes both conceptual and practical contributions to the emerging field of sustainable Generative AI.

The first key contribution is the development of a conceptual framework that maps environmental sustainability strategy types across the GenAI lifecycle. This framework enables a systematic classification of existing sustainability approaches, reveals underrepresented lifecycle stages and strategy types, and serves as a foundational mapping of the field for future research. In doing so, it provides theoretical clarity in a domain where sustainability efforts are often technically siloed and disconnected.

The second contribution is a scoping study that showcased how mature the current scientific literature is across the different identified strategy types. Through this, various gaps in literature have been identified, which can serve to inform future research.

The third contribution is a governance-oriented sustainability blueprint, developed through field research in a large international professional services firm. This blueprint operationalizes the conceptual framework, demonstrating its applicability in real-world organizational contexts. It offers a concrete, actionable model for integrating environmental sustainability into the development processes of GenAI-based applications. The blueprint is iteratively refined through stakeholder feedback, ensuring its practical relevance and adaptability.

Together, the conceptual framework, scoping study and the practical blueprint establish a bridge between academic insight and industry need. The research not only advances theoretical understanding of sustainability in GenAI but also provides a prototype for how such understanding can be implemented in practice, thereby contributing both to academia and to responsible innovation in applied settings.

6.3 Industry Adoption

An important indicator of a framework's practical value lies in its applicability beyond the academic setting. The non-profit consortium SustainableIT.org - a prominent alliance of global IT executives and sustainability leaders - is in the process of developing a lifecycle governance model for Responsible AI. Their goal is to broaden the traditional focus of AI governance, which typically emphasizes data integrity, bias mitigation, and algorithmic transparency, to also include environmental sustainability dimensions.

SustainableIT.org became aware of the sustainability framework presented in this thesis. After engaging with the model and its core propositions, the consortium began integrating selected principles into their governance model targeted at enterprise IT leaders.

In recognition of this influence, the principle researcher of SustainableIT.org Rick Pastore and Wiebren van der Zee, member of the SustainableIT.org's European Advisory board provided the following written statement:

"Mr. Jung has developed a concept for governing the proliferation of AI in the enterprise that incorporates environmental sustainability factors that do not typically appear in AI governing processes. In fact, most AI governance models stipulate data integrity – accuracy, bias, copyrights, privacy, transparency, etc. – but ignore the power and materials consumption, emissions, and water usage impacts of AI. Since January of 2024, SustainableIT.org's IT executive members have been struggling to develop a general lifecycle governance model for responsible AI that encompasses environmental and social sustainability. So, we were delighted to come across Mr. Jung's thesis research and model presentation. It has helped inform our broad AI governance model for IT leaders, encompassing such critical requirements as upstream and downstream data management, model selection, business case fit-for-purpose, and infrastructure suitability. It is gratifying to see graduating technology leaders like Mr. Jung focusing the sustainable future of business technology infrastructure."

This endorsement serves as a validation of the framework's impact and a demonstration of it's usability. It also affirms the broader relevance of incorporating environmental sustainability into AI governance structures, which historically have remained focused on ethical, legal, and social implications, but often neglect the environmental impacts of AI.

The adoption of key elements by SustainableIT.org reinforces one of the central claims of this thesis: That environmental sustainability shall be treated as a first-class concern in the governance of GenAI systems. Furthermore, it showcases the framework's potential to guide IT decision-makers seeking to align innovation with environmental responsibility

6.4 Limitations

Despite its contributions, this study is subject to several limitations. Due to the scope of the research, the practical blueprint was not implemented and tested. Having seasoned IT leaders recognize the contribution of the framework for practice (as demonstrated in the previous chapter) increases confidence in its value for the industry, but so far it is unclear, how the best-practice for the implementation in companies might look like. The practical blueprint serves as a explorative study, supported through stakeholder engagement, but the reality of implementing such a governance blueprint will likely yield far more hurdles.

A limitation of this work, is it being developed in „isolation". In practice, various workflows, company structures and standardized processes are preexisting, into which such a blueprint might be integrated. While the study only aimed at providing an overarching understanding, integrating the presented framework into these processes can increase the actionability of the framework.

Another limitation of the framework is the combination of all strategies that support the adoption of the R-strategies into the class of support strategies. Within this category, different types exist, such as quantification approaches, incentives to drive research into the topic or awareness-creation. Creating a clear distinction between the different classes of strategies under the label "support-strategies" can increase the actionability.

A practical sustainability approach might be addressing various strategy types at once, making the allocation of such an approach to an individual strategy type difficult. This issue was solved by tagging such approaches with all applicable strategy types. Nonetheless, this lack of exclusiveness might result in difficulties. This issue should be somewhat minimal, as the primary function of the framework is to provide a mapping of the system to derive strategies, and not the labeling of approaches in retrospective.

Lastly, the abstract nature of the framework requires may lead different subjective interpretations, depending on the user and the context. Adding further granularity could increase specificity and reduce the danger of misinterpretations. So far, misunderstandings could not be observed within the involved group of stakeholders and it must be noted that different interpretations may lead to beneficial outcomes, as it may result in more diverse approaches being developed.

6.5 Future Work

As GenAI models evolve at scale and in use, so do their environmental impacts. While the field of environmental sustainability of AI is picking up pace at this point in time, various new fields for future work emerge.

Future research can be done on providing a conceptual framework of support strategies. As the current context does not provide favorable conditions for the implementation of sustainability measures, these strategies are promising.

Besides that, research on integrating the framework into preexisting workflows and standardized processes can increase its actionability and practicability.

Next, the framework can be extended beyond environmental concerns, towards social and economic sustainability, to cover the full range within a single model. Such efforts are already underway (see chapter 6.3) This can increase the diversity of impacts and its fields of application.

To increase interest and pressure in favor of driving the uptake of sustainability measures for GenAI, more and increasinglt accurate impact quantifications need to be conducted (beyond CO2 emissions). Therefore quantifications of the impacts from hardware (e.g. GPUs) and models themselves would prove helpful.

Lastly, to increase the update of such measures, further research should be conducted on policy frameworks, that make sustainability measures for AI applications mandatory, to ensure a coordinated and systematic approach across organizations and stakeholders.

6.6 Personal Reflection

I became interested in the intersection of sustainability and AI out of curiosity - how this evolving and increasingly present technology could be utilised to solve sustainability challenges. I read about its use in increasing resource efficiency by analysing and optimising systems, its potential to drive sustainability innovations, and its ability to help us build resilience to the effects of climate change, for example through better weather predictions. While AI's potential to support humanity in addressing major challenges is exciting, its environmental downsides quickly became apparent

The astonishing amount of work this technology can perform comes with a significant resource price tag: extreme amounts of energy for its training and operation, vast amounts of water for the cooling of its hardware, and the rapid construction of more and more data centres, built with critical resources.
And all of this, on an already overexploited and increasingly unstable planet.
It became clear to me that the real question was not what AI can do, but how we develop and deploy it. As designers in this time, we are already used to designing products or systems in more sustainable manners, so why not AI applications?

Emerging in the field, I realised early in the project how deep technological determinism runs through the environment surrounding AI and its application. Ever-larger models continue to be developed, pushing us closer to environmental limits. Just last week, Sam Altman, the CEO of OpenAI - the company behind ChatGPT - proclaimed that "the cost of AI will converge to the cost of energy (…) the abundance of it will be limited by the abundance of energy." Resources will become the limiting factor, and instead of aiming to utilise our available resources more wisely and purposefully, the tech community seems to strive towards scaling up energy production and resource extraction.

While some voices in the sustainable AI field proclaim that AI is inherently unsustainable (I agree!), there seems no way to stop its rapid expansion. Therefore, this project became about pushing towards a shift - from the predominant technological determinism towards a more social constructivism. We simply cannot afford to forget that technology must serve us and our environment, to not risk deteriorating our quality of lives and homes. As a designer, this project has underlined my responsibility: to actively and consciously consider how we want to create the products, systems, and applications around us and to design them based on desirable values. To me, that means questioning a blind push towards supposed progress - as real progress rarely comes one-dimensional.

This project allowed me to explore both the academic landscape and the industry around AI.

Emerging in the academic landscape surrounding the sustainability of the technology was highly interesting and proved reason for hope. Across all the various strategic dimensions I propose, I could find approaches being developed - from software to hardware to how and when we deploy it. I came across GPUs being taken apart in labs to derive precise LCAs, cost-benefit frameworks to consider the trade-off between environmental impacts versus created value, and much more. The research landscape is quickly expanding, and with it come increasingly mature sustainability approaches.

In the industry landscape, I observed that the involved actors were generally well-meaning, aware of the sustainability challenges, and the need to reshape the way we handle AI. Against this individual good-will stands another form of determinism - a capitalistic one. The need to survive and drive forward means that companies will act in ways that increase their profitability and strength in the market. This means that factors such as maximising performance, pursuing go-to-market speed, and saving costs and resources on the way leave little space for sustainability concerns.

While criticism of capitalism is appropriate at this point, I don't expect this system will change fast enough to outpace the growing environmental strains it puts on our environment. Therefore, I prefer to be pragmatic. To change the way we deploy AI in the industry, we must make it desirable from a business perspective. This means creating a business case - for example by increasing pressure from the consumer side or by creating regulation and policy. Only then can the good-will of the industry actors be transformed into effective action. After this project I only see one realistic way forward: regulations must demand transparency and sustainability safeguards from all those who deploy the technology.

Returning to my earlier comment, which may have seemed surprising: Yes, I do believe that AI is inherently unsustainable. We have had many impactful technological revolutions in the past, and we know where this led us. Yes, AI offers unprecedented capabilities - but its resource demands are equally unprecedented. Yet I still want to continue working on it, as I am certain that it is here to stay. I notice it in my own behavior. Throughout the project, I've learned about all its downsides, and still I am using it daily. Why? A combination of curiosity and the undeniable utility it offers - it simplifies, accelerates, and enhances much of my work. Its potential is so vast that we should embrace its existence while making sure we shape it in a way that serves us and our environment.

I have explored the tension between technological ambition and ecological responsibility, aiming to offer a structured response grounded in both theory and practical reality. Through engaging in company practice, I gained insights into a large AI deployer, built expertise in sustainability approaches, and explored the real-world applications of GenAI in corporate settings. Through this, I fulfilled my intended learning goals and found a direction for future work I'm very passionate about. This thesis reflects my position as a researcher and designer committed to shaping technological progress with purpose: to serve both people and planet. To turn goodwill into meaningful change, we must design better alternatives, raise awareness, and implement enforceable sustainability policies. Only then can the future of AI be aligned with ecological responsibility.

Acknowledgement

I would like express my deepest gratitude for those who provided guidance, resources, feedback, inpiration, knowledge and motivation throughout the development of my thesis.

I'd like to thank my university supervisors for their interest in my work, their academic guidance, expertise and valuable feedback:

Prof. Dr. Gerd Kortuem, chair
Prof. Dr. Conny Bakker, mentor

Further I am grateful to experience practice at the company Accenture.

I thank my company mentors for sharing experiences, providing me with insights in organizational realities and supporting my project through feedback and knowledge.

Sophie Vos,	Accenture
Dr. Angelika Schmid,	Accenture

I am also grateful to the experts who contributed their time and perspectives through interviews:

Dr. Noman Bashir,	MIT Climate & Sustainability Consortium | Computer Science & Artificial Intelligence Laboratory
Sophia Falk,	Sustainable AI Lab - University of Bonn
Ioanix Kolaxis,	Accenture, Technology Research
Wilco Burggraaf,	Green Software Foundation & Hightech Innovators

And all interviewees who remain anonymous - your input was essential to shaping this work.

Special thanks go to Wiebren van der Zee (Sustainable IT Expert in CIO office of ABN AMRO Bank) and Rick Pastore (Principal Researcher at SustainableIT.org and Senior Director of the Hackett Group) for their external support, constructive feedback, and genuine interest in this research.

This thesis was made possible through the engagement of practitioners and researchers, aswell as friends and family, especially Jenný Mikaelsdóttir and David Soche, who supported this work through proof reading, feedback and support and Amelie Straubmüller who designed the title visual.

References

A

Akhtar, Z. Bin. (2024). Unveiling the evolution of generative AI (GAI): a comprehensive and investigative analysis toward LLM models (2021–2024) and beyond. Journal of Electrical Systems and Information Technology, 11(1), 22. https://doi.org/10.1186/s43067-024-00145-1

Arsanjani, A. (2023, March 21). The Generative AI Life-cycle. Https://Dr-Arsanjani.Medium.Com/the-Generative-Ai-Life-Cycle-Fb2271a70349.

B

Badiger, M., & Mathew, J. A. (2022). Retrospective Review of Activation Functions in Artificial Neural Networks. In V. Bindhu, J. M. R. S. Tavares, & K.-L. Du (Eds.), Proceedings of Third International Conference on Communication, Computing and Electronics Systems (pp. 905–919). Springer Singapore.

Bandi, A., Adapa, P. V. S. R., & Kuchi, Y. E. V. P. K. (2023). The Power of Generative AI: A Review of Requirements, Models, Input–Output Formats, Evaluation Metrics, and Challenges. In Future Internet (Vol. 15, Issue 8). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/fi15080260

Banh, L., & Strobel, G. (2023). Generative artificial intelligence. Electronic Markets, 33, 63. https://doi.org/10.1007/s12525-023-00680-1

Bashir, N., Donti, P., Cuff, J., Sroka, S., Ilic, M., Sze, V., Delimitrou, C., & Olivetti, E. (2024). The Climate and Sustainability Implications of Generative AI. https://mit-genai.pubpub.org/pub/8ulgrckc

Berthelot, A., Caron, E., Jay, M., & Lefèvre, L. (2024). Estimating the environmental impact of Generative-AI services using an LCA-based methodology. Procedia CIRP, 122, 707–712. https://doi.org/https://doi.org/10.1016/j.procir.2024.01.098

Boston Consulting Group. (2024). From Potential to Profit with GenAI.

Brevini, B. (2020). Black boxes, not green: Mythologizing artificial intelligence and omitting the environment. Big Data & Society, 7(2), 2053951720935141. https://doi.org/10.1177/2053951720935141

Brown Hamilton, T. (2022, June 1). In a small Dutch town, a fight with Meta over a massive data center. Https://Www.Washingtonpost.Com/Climate-Environment/2022/05/28/Meta-Data-Center-Zeewolde-Netherlands/.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. http://arxiv.org/abs/2005.14165

Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? Workforce implications: Profound change is coming, but roles for humans remain. Science, 358(6370), 1530–1534. https://doi.org/10.1126/SCIENCE.AAP8062

C

CBS. (2023). How many litres of water do we use per day? - The Netherlands in numbers. https://longreads.cbs.nl/the-netherlands-in-numbers-2023/how-many-litres-of-water-do-we-use-per-day/

CBS. (2024). How big is the Dutch economy? - The Netherlands in Numbers 2024. https://longreads.cbs.nl/the-netherlands-in-numbers-2024/how-big-is-the-dutch-economy/

CE, & MVO Nederland. (2015). The potential for high value reuse in a circular economy. https://susdi.org/doc/CE/high%20value%20reuse%20 27102a5465b3589c6b52f8e43ba9fd72.pdf

Chawla, C., Chatterjee, S., Gadadinni, S. S., Verma, P., & Banerjee, S. (2024). Agentic AI: The building blocks of sophisticated AI business applications. Journal of AI, Robotics & Workplace Automation, 3(3), 196. https://doi.org/10.69554/XEHZ1946

Chui, M., Hall, B., Singla, A., Sukharevsky, A., & Yee, L. (2023). The state of AI in 2023: Generative AI's breakout year. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year

Chui, M., Hazan, E., Roberts, R., Singla, A., Smaje, K., Sukharevsky, A., Yee, L., & Zemmel, R. (2023). The economic potential of generative AI: The next productivity frontier. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier

City of Altoona Iowa. (2021). Choosing a location. Https://Www.Altoona-Iowa.Com/Business/Economic_development/Choosing_a_location.Php.

Cloudscene. (2024, March 14). Leading countries by number of data centers as of March 2024 [Graph]. In Statista. Https://Www-Statista-Com.Tudelft.Idm.Oclc.Org/Statistics/1228433/Data-Centers-Worldwide-by-Country/.

CodeCarbon. (n.d.). Landing Page. Https://Codecarbon.Io/.

Council for the Environment and Infrastructure Netherlands. (2015). Circular Economy: From wish to practice. www.rli.nl

Crawford, K. (2024). Generative AI's environmental costs are soaring - and mostly secret. https://www.nature.com/articles/d41586-024-00478-x

Crawford, K., & Joler, V. (2018). Anatomy of an AI system: The Amazon Echo as an anatomical map of human labor, data and planetary resources. https://anatomyof.ai/

D

Davis, J. C., Jajal, P., Jiang, W., Schorlemmer, T. R., Synovic, N., & Thiruvathukal, G. K. (2024). Reusing Deep Learning Models: Challenges and Directions in Software Engineering. http://arxiv.org/abs/2404.16688

Dell Technologies. (n.d.). Network architecture for the next computer clusters | Dell Technologies Fabrics and GenAI: The New World of Artificial Intelligence. Https://Infohub.Delltechnologies.Com/En-Us/l/Dell-Technologies-Fabrics-and-Genai-the-New-World-of-Artificial-Intelligence/Network-Architecture-for-the-next-Computer-Clusters/.

Dencik, J., Goehring, B., & Marshall, A. (2023). Managing the emerging role of generative AI in next-generation business. Strategy & Leadership, 51(6), 30–36. https://doi.org/10.1108/SL-08-2023-0079

Devlin, J., Chang, M.-W., Lee, K., Google, K. T., & Language, A. I. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://github.com/tensorflow/tensor2tensor

E

Ellen MacArthur Foundation. (2013). Towards the circular economy Vol. 1: an economic and business rationale for an accelerated transition. https://www.ellenmacarthurfoundation.org/towards-the-circular-economy-vol-1-an-economic-and-business-rationale-for-an

European Comission | Directorate-General for Communication. (2025, February 25). Commission proposes to cut red tape and simplify business environment. Https://Commission.Europa.Eu/News/Commission-Proposes-Cut-Red-Tape-and-Simplify-Business-Environment-2025-02-26_en.

F

Falk, S., & van Wynsberghe, A. (2023). Challenging AI for Sustainability: what ought it mean? AI and Ethics. https://doi.org/10.1007/s43681-023-00323-3

Falk, S., van Wynsberghe, A., & Biber-Freudenberger, L. (2024). The attribution problem of a seemingly intangible industry. Environmental Challenges, 16, 101003. https://doi.org/10.1016/J.ENVC.2024.101003

G

Gatla, R. K., Gatla, A., Sridhar, P., Kumar, D. G., & Rao, D. S. N. M. (2024). Advancements in Generative AI: Exploring Fundamentals and Evolution. 2024 International Conference on Electronics, Computing, Communication and Control Technology (ICECCC), 1–5. https://doi.org/10.1109/ICECCC61767.2024.10594003

Geissdoerfer, M., Savaget, P., Bocken, N., & Hultink, E. J. (2017). The Circular Economy – A new sustainability paradigm? Journal of Cleaner Production, 143, 757–768. https://doi.org/https://doi.org/10.1016/j.jclepro.2016.12.048

George, J. (2024). Data to AI: Building a solid data foundation for your generative AI applications in the cloud. https://doi.org/10.30574/ijsra.2024.11.1.1056

Goldman Sachs. (2023, April 5). Generative AI could raise global GDP by 7%. Https://Www.Goldmansachs.Com/Insights/Articles/Generative-Ai-Could-Raise-Global-Gdp-by-7-Percent.Html.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press. http://www.deeplearningbook.org

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Networks. http://arxiv.org/abs/1406.2661

Google Earth Pro V 7.3.6.9796. (24.08.2023). The Citadel Campus, Nevada, USA. 39°30‘48.89“N,119°28‘53.19“E, alt 1,5 km. 2023, http://www.earth.google.com

H

Haakman, M., Cruz, L., Huijgens, H., & van Deursen, A. (2021). AI lifecycle models need to be revised. Empirical Software Engineering, 26(5), 95. https://doi.org/10.1007/s10664-021-09993-1

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. http://arxiv.org/abs/2006.11239

Hoveling, T., Svindland Nijdam, A., Monincx, M., Faludi, J., & Bakker, C. (2024). Circular economy for medical devices: Barriers, opportunities and best practices from a design perspective. Resources, Conservation and Recycling, 208, 107719. https://doi.org/https://doi.org/10.1016/j.resconrec.2024.107719

Huang, K., Yin, H., Huang, H., & Gao, W. (2023). Towards Green AI in Fine-tuning Large Language Models via Adaptive Backpropagation. ICLR 2024. http://arxiv.org/abs/2309.13192

Humlum, A., & Vestergaard, E. (2024). The Adoption of ChatGPT. https://doi.org/http://dx.doi.org/10.2139/ssrn.4827166

Husein, M., & Chung, I.-Y. (2019). Day-Ahead Solar Irradiance Forecasting for Microgrids Using a Long Short-Term Memory Recurrent Neural Network: A Deep Learning Approach. Energies, 12, 1856. https://doi.org/10.3390/en12101856

I

International Energy Agency, Cozzi, L., Chen, O., & Kim, H. (2023). The world's top 1% of emitters produce over 1000 times more CO2 than the bottom 1%. https://www.iea.org/commentaries/the-world-s-top-1-of-emitters-produce-over-1000-times-more-co2-than-the-bottom-1

Islam, M. J., Pan, R., Nguyen, G., & Rajan, H. (2020). Repairing deep neural networks: fix patterns and challenges. Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering, 1135–1146. https://doi.org/10.1145/3377811.3380378

J

Jordan, M. I., & Mitchell T.M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 253–255. https://doi.org/10.1126/science.aac4520

K

Kaack, L. H., Donti, P. L., Strubell, E., Kamiya, G., Creutzig, F., & Rolnick, D. (2022). Aligning artificial intelligence with climate change mitigation. Nature Climate Change, 12(6), 518–527. https://doi.org/10.1038/s41558-022-01377-7

Kanbach, D. K., Heiduk, L., Blueher, G., Schreiter, M., & Lahmann, A. (2024a). The GenAI is out of the bottle: generative artificial intelligence from a business model innovation perspective. In Review of Managerial Science (Vol. 18, Issue 4, pp. 1189–1220). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/s11846-023-00696-z

Kanbach, D. K., Heiduk, L., Blueher, G., Schreiter, M., & Lahmann, A. (2024b). The GenAI is out of the bottle: generative artificial intelligence from a business model innovation perspective. In Review of Managerial Science (Vol. 18, Issue 4, pp. 1189–1220). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/s11846-023-00696-z

Kaswan, K. S., Dhatterwal, J. S., Malik, K., & Baliyan, A. (2023). Generative AI: A Review on Models and Applications. 2023 International Conference on Communication, Security and Artificial Intelligence (ICCSAI), 699–704. https://doi.org/10.1109/ICCSAI59793.2023.10421601

Kingma, D. P., & Welling, M. (2013). Auto-Encoding Variational Bayes. http://arxiv.org/abs/1312.6114

L

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. In Nature (Vol. 521, Issue 7553, pp. 436–444). Nature Publishing Group. https://doi.org/10.1038/nature14539

Li, M., Liu, Y., Liu, X., Sun, Q., You, X., Yang, H., Luan, Z., Gan, L., Yang, G., & Qian, D. (2020). The Deep Learning Compiler: A Comprehensive Survey. https://doi.org/10.1109/TPDS.2020.3030548

Li, P., Yang, J., Islam, M. A., & Ren, S. (2023). Making AI Less 'Thirsty': Uncovering and Addressing the Secret Water Footprint of AI Models. http://arxiv.org/abs/2304.03271

Ligozat, A. L., Lefevre, J., Bugeau, A., & Combaz, J. (2022). Unraveling the Hidden Environmental Impacts of AI Solutions for Environment Life Cycle Assessment of AI Solutions. Sustainability (Switzerland), 14(9). https://doi.org/10.3390/su14095172

Luccioni, A. S., & Hernandez-Garcia, A. (2023). Counting Carbon: A Survey of Factors Influencing the Emissions of Machine Learning.

M

Mangal, P., Mak, C., Kanakis, T., Donovan, T., Braines, D., & Pyzer-Knapp, E. (2024). Coalitions of Large Language Models Increase the Robustness of AI Agents.

Meemken, B., & Poth, A. (2024). Eco-Friendly AI: A Guide to Energy-Efficient Techniques Across the AI Life-Cycle. In M. Yilmaz, P. Clarke, A. Riel, R. Messnarz, C. Greiner, & T. Peisl (Eds.), Systems, Software and Services Process Improvement (pp. 36–50). Springer Nature Switzerland.

Microsoft Coorperation. (2024, June 25). The Generative AI Application Lifecycle. Https://Learn.Microsoft.Com/de-de/Shows/Generative-Ai-for-Beginners/the-Generative-Ai-Application-Lifecycle-Generative-Ai-for-Beginners?WT.Mc_id=academic-105485-Koreyst.

Mytton, D. (2021). Data centre water consumption. Npj Clean Water, 4(1), 11. https://doi.org/10.1038/s41545-021-00101-w

O

OECD. (2022). Measuring the environmental impacts of artificial intelligence compute and applications: The AI footprint (Vol. 341). OECD Publishing.

OpenAI. (n.d.). Tokenizer. Https://Platform.Openai.Com/Tokenizer.

P

Paulillo, A., & Sanyé-Mengual, E. (2024). Approaches to incorporate Planetary Boundaries in Life Cycle Assessment: A critical review. In Resources, Environment and Sustainability (Vol. 17). Elsevier B.V. https://doi.org/10.1016/j.resenv.2024.100169

Potting, J., Hekkert, M., Worrell, E., & Hanemaaijer, A. (2017). Circular Economy: Measuring innovation in the product chain.

R

Radford, A., Narasimhan, K., Sutskever, I., & Salimans, T. (2018). Improving Language Understanding by Generative Pre-Training. https://gluebenchmark.com/leaderboard

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. https://github.com/codelucas/newspaper

Ranta, V., Aarikka-Stenroos, L., & Mäkinen, S. J. (2018). Creating value in the circular economy: A structured multiple-case analysis of business models. Journal of Cleaner Production, 201, 988–1000. https://doi.org/10.1016/j.jclepro.2018.08.072

Raper, R., Boeddinghaus, J., Coeckelbergh, M., Gross, W., Campigotto, P., & Lincoln, C. N. (2022). Sustainability Budgets: A Practical Management and Governance Method for Achieving Goal 13 of the Sustainable Development Goals for AI Development. SUSTAINABILITY, 14(7). https://doi.org/10.3390/su14074019

Reuters. (2024, January 16). OpenAI CEO Altman says at Davos future AI depends on energy breakthrough. Https://Www.Reuters.Com/Technology/Openai-Ceo-Altman-Says-Davos-Future-Ai-Depends-Energy-Breakthrough-2024-01-16/.

Richardson, K., Steffen, W., Lucht, W., Bendtsen, J., Cornell, S. E., Donges, J. F., Drüke, M., Fetzer, I., Bala, G., Von Bloh, W., Feulner, G., Fiedler, S., Gerten, D., Gleeson, T., Hofmann, M., Huiskamp, W., Kummu, M., Mohan, C., Nogués-Bravo, D., … Rockström, J. (2023). Earth beyond six of nine planetary boundaries. https://www.science.org

Robbins, S., & van Wynsberghe, A. (2022). Our New Artificial Intelligence Infrastructure: Becoming Locked into an Unsustainable Future. Sustainability (Switzerland), 14(8). https://doi.org/10.3390/su14084829

Rockström, J., Steffen, W., Noone, K., Persson, Å., Chapin, F. S., Lambin, E. F., Lenton, T. M., Scheffer, M., Folke, C., Schellnhuber, H. J., Nykvist, B., de Wit, C. A., Hughes, T., van der Leeuw, S., Rodhe, H., Sörlin, S., Snyder, P. K., Costanza, R., Svedin, U., … Foley, J. A. (2009). A safe operating space for humanity. Nature, 461(7263), 472–475. https://doi.org/10.1038/461472a

Roussilhe, G., Pirson, T., Xhonneux, M., & Bol, D. (2024). From silicon shield to carbon lock-in? The environmental footprint of electronic components manufacturing in Taiwan (2015–2020). Journal of Industrial Ecology, 28(5), 1212–1226. https://doi.org/https://doi.org/10.1111/jiec.13487

Rubei, R., Moussaid, A., di Sipio, C., & di Ruscio, D. (2025). Prompt engineering and its implications on the energy consumption of Large Language Models. http://arxiv.org/abs/2501.05899

S

Salehi, S., & Schmeink, A. (2024). Data-Centric Green Artificial Intelligence: A Survey. IEEE Transactions on Artificial Intelligence, 5(5), 1973–1989. https://doi.org/10.1109/TAI.2023.3315272

Saltz, J. (2024a, November 18). CRISP-DM is Still the Most Popular Framework for Executing Data Science Projects. Https://Www.Datascience-Pm.Com/Crisp-Dm-Still-Most-Popular/.

Saltz, J. (2024b, November 19). The GenAI Life Cycle. Https://Www.Datascience-Pm.Com/the-Genai-Life-Cycle/.

Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. Communications of the ACM, 63(12), 54–63. https://doi.org/10.1145/3381831

Semrush. (2024, February 28). Worldwide visits to ChatGPT web page (chat.openai.com) from April 2022 to January 2024 (in millions) [Graph]. Https://Www.Statista.Com/Statistics/1463713/Chatgpt-Chat-Openai-Com-Monthly-Visits/.

Sengar, S. S., Hasan, A. Bin, Kumar, S., & Carroll, F. (2024). Generative Artificial Intelligence: A Systematic Review and Applications. http://arxiv.org/abs/2405.11029

Shearer C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing, 5(4).

Shumskaia, E. I. (2022). Artificial Intelligence—Reducing the Carbon Footprint? In E. B. Zavyalova & E. G. Popkova (Eds.), Industry 4.0: Fighting Climate Change in the Economy of the Future (pp. 359–365). Springer International Publishing. https://doi.org/10.1007/978-3-030-79496-5_33

Stanford University. (2024, April 15). Emission of CO_2 equivalent by artificial intelligence (AI) models in 2024 (in metric tons) [Graph]. Https://Www.Statista.Com/Statistics/1465353/Total-Co2-Emission-of-Ai-Models/.

Statista. (2025, March 25). Generative artificial intelligence (AI) market size worldwide from 2021 to 2031 (in billion U.S. dollars) [Graph]. Https://Www.

Statista.Com/Forecasts/1449838/Generative-Ai-Market-Size-Worldwide.

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. http://arxiv.org/abs/1906.02243

Structure Research. (2023). 2023 Global Hyperscale Datacenter: Research Report & Analysis. https://www.structureresearch.net/product/2023-hyperscale-self-build-data-centre-report/

Stryker, C., & Bergmann, D. (2024, June 12). What is a Variational Autoencoder? . Https://Www.Ibm.Com/Think/Topics/Variational-Autoencoder.

Synergy Research Group. (2023, October 17). Hyperscale Data Center Capacity to Almost Triple in Next Six Years, Driven by AI. Https://Www.Srgresearch.Com/Articles/Hyperscale-Data-Center-Capacity-to-Almost-Triple-in-next-Six-Years-Driven-by-Ai.

Synergy Research Group. (2024, April 17). Number of hyperscale data centers worldwide from 2015 to 2023 [Graph]. In Statista. Https://Www-Statista-Com.Tudelft.Idm.Oclc.Org/Statistics/633826/Worldwide-Hyperscale-Data-Center-Numbers/.

T

Thakur, D., Guzzo, A., Fortino, G., & Piccialli, F. (2024). Green Federated Learning: A new era of Green Aware AI. http://arxiv.org/abs/2409.12626

Tian, M., Lu, J., Gao, H., Wang, H., Yu, J., & Shi, C. (2022). A Lightweight Spiking GAN Model for Memristor-centric Silicon Circuit with On-chip Reinforcement Adversarial Learning - 2022 IEEE International Symposium on Circuits and Systems (ISCAS). 2022 IEEE International Symposium on Circuits and Systems (ISCAS), 3388–3392. https://doi.org/10.1109/ISCAS48785.2022.9937639

U

U.S. Department of Energy. (n.d.). Data Centers and Servers. Https://Www.Energy.Gov/Eere/Buildings/Data-Centers-and-Servers.
van Wynsberghe, A. (2021). Sustainable AI: AI for sustainability and the sustainability of AI. AI and Ethics, 1(3), 213–218. https://doi.org/10.1007/s43681-021-00043-6

V

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, \L{}ukasz, & Polosukhin, I. (2017). Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems, 6000–6010.

Verdecchia, R., Sallou, J., & Cruz, L. (2023). A Systematic Review of Green AI. http://arxiv.org/abs/2301.11047

W

Wang, S., Shao, Z., & Lui, J. C. S. (2024). Next-Word Prediction: A Perspective of Energy-Aware Distributed Inference. IEEE Transactions on Mobile Computing, 23(5), 5695–5708. https://doi.org/10.1109/TMC.2023.3310536

Wang, Y.-C., Xue, J., Wei, C., & Kuo, C.-C. J. (2023). An Overview on Generative AI at Scale with Edge-Cloud Computing. https://doi.org/10.36227/techrxiv.23272271

Wu, C.-J., 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., Melnikov, A., Candido, S., Brooks, D., Chauhan, G., Lee, B., Lee, H.-H. S., … Hazelwood, K. (2021). Sustainable AI: Environmental Implications, Challenges and Opportunities. http://arxiv.org/abs/2111.00364

Y

Yarally, T., Cruz, L., Feitosa, D., Sallou, J., & van Deursen, A. (2023). Uncovering Energy-Efficient Practices in Deep Learning Training: Preliminary Steps Towards Green AI. http://arxiv.org/abs/2303.13972

Z

Zhang, C., Zhang, F., Chen, K., Chen, M., He, B., & Du, X. (2023). EdgeNN: Efficient Neural Network Inference for CPU-GPU Integrated Edge Devices. 2023 IEEE 39th International Conference on Data Engineering (ICDE), 1193–1207. https://doi.org/10.1109/ICDE55515.2023.00096

Appendix

Appendix A: Generative Artificial Intelligence

Introduction

The sustainability impacts of Generative Artificial Intelligence (GenAI) are a result not just of the technology itself but also its context, of how, where, why and when we use it. The actors involved in this context are of diverse backgrounds, beyond just artificial intelligence experts and researchers. Therefore, I highlight the need of addressing a broad audience, one that is not necessarily familiar with the technology already. For these reasons, I advocate for an inclusive presentation of research, one that gives space not just to the specifics but also to the foundations. Therefore, I provide this appendix, for readers unfamiliar with the the topic of generative artificial intelligence.

The first parts “Artificial Intelligence & Machine Learning”, “Types of Machine Learning” and “Deep Learning” will provide a brief overview and understanding of the classification of concepts in the field. The aim of these chapters is to provide a more nuanced understanding and clarify the conceptual differences between Artificial Intelligence, Machine Learning and Deep Learning. The chapter “Artificial Neural Networks” explains how models are able to “learn” and extract increasingly complex features from data, enabling them to produce the desired output, in order to create an understanding of the underlying paradigm.

Next, the chapter „Generative Artificial Intelligence“ delves into the core technology examined in this thesis. It covers the most relevant types of GenAI models as of December 2024, exploring their development, functionality, and scale. This sets the stage for the following section, „The GPT-3 Architecture,“ which details the design of this landmark model. The exploration provides insight into how such models generate data and the scale of computations required for their training and operation.

Lastly the chapter “the physical infrastructure of Generative AI” provides an overview of the hardware that is required to operate GenAI. This chapter aims to create the awareness, that abstract concepts such as Generative AI or cloud computing, which at times might appear as a non-physical instance, are indeed supported by a vast geological and materialized backbone.

Overall, this appendix aims to equip all readers with a basic understanding of GenAI technology, its key concepts, operational mechanisms, and the scale of computations behind its operation.

Methodology

The search strategy utilized for the Appendix A aimed at selecting the foundational works in the presented fields. On Google Scholar the following search string was applied: „generative AI“ OR „GenAI“ OR „generative artificial intelligence" OR "Machine Learning" OR "Deep Learning". From the results those papers were selected that offered a general understanding of the technology.

When examining a specific model – such as GPT-3 - the corresponding foundational papers were selected directly. These were found via snowballing papers derived from the Google Scholar search.
In addition, papers were searched that provided an overview of the hardware infrastructure of GenAI. On Google Scholar the following search string was applied: „Infrastructure" OR „Hardware" OR „Requirements" AND "GenAI" OR "Generative AI" OR "Generative Artificial Intelligence". Papers that provided information on the hardware requirements were selected.

In order to provide insights into the trends and statistics surrounding the physical infrastructure of GenAI, a Google Search was conducted with the search string: „Infrastructure" OR „Data Center" OR „Hyperscale" OR "Network" AND "GenAI" OR "Generative AI" OR "Generative Artificial Intelligence" OR "Artificial Intelligence" OR "AI".

As the aim of this chapter is not an exhaustive review of the research landscape but rather the explanation of different concepts, the literature search was ended once all the concepts of interest have been addressed.

Appx.A.1 Artificial Intelligence & Machine Learning

The term 'artificial intelligence' (AI) describes the overarching category of computational models performing tasks, that require a form of intelligence, such as pattern recognition or experiential learning (Banh & Strobel, 2023).

A subcategory of AI is machine learning (ML). ML models can detect rulesets and rationales from historic data and apply these learned mechanisms to newly provided data (Banh & Strobel, 2023). These rulesets do not need to be explicitly programed, as they are automatically learned by the exposure to data. In many cases, modern ML algorithms have allowed the creation of computer systems more capable and accurate than manually programmed ones (Brynjolfsson & Mitchell, 2017). Jordan and Mitchell define machine learning problems as "the problem of improving some measure of performance when executing some task through some type of training experience." (Jordan & Mitchell T.M., 2015, p. 255). For example, in the case of spam detection, the learning could aim at assigning the labels "spam" and "no spam" to emails. The training experience could be a collection of mails with the corresponding - in retrospect assigned - labels "spam" and "no spam" and the measure of performance for example the accuracy of the assigned classification.

Appx.A.2 Types of Machine Learning

Within the field of machine learning, different categories have emerged. The most common categories are supervised learning, unsupervised learning and reinforcement learning. In supervised learning, algorithms are trained on labeled data (such as in the spam detection example), learning to assign outputs to inputs (Jordan & Mitchell T.M., 2015). In case of the example, the assigned classification is the output (y) and the corresponding mail is the input (x).

Unsupervised learning is the process of models learning to detect hidden patterns in the data themselves, for example as used in clustering algorithms (Banh & Strobel, 2023). An exemplary problem for unsupervised learning is clustering - the partition of unlabeled data into groups based on specific assumptions about each cluster, e.g. the semantic meaning of words.

Lastly reinforcement learning algorithms train optimal decision making by interacting with an environment, collecting rewards depending on the quality of its output, with the goal of maximizing the rewards collected (Banh & Strobel, 2023). The data on which reinforcement models are trained doesn't indicate a false or correct output to the related input, but instead assigns rewards based on the quality of output over a sequence of inputs (Jordan & Mitchell, 2015). Such an approach is useful when the problem requires a sequence of decision making, such as learning to perform well at a game.

Appx.A.3 Deep Learning

Deep learning emerged as a more advanced subset of ML. While there is not clear line at which a ML model can be described as a deep learning (DL) model, DL models differentiate themselves by their ability to automatically extract features from data and sort these features into hierarchical representations of the data (I. Goodfellow et al., 2016). Features are properties and characteristics of the data, some of which are utilized to create the desired output (e.g. in the example of an image classification model features can be colors, edges or corners). The model learns from the resulting multiple levels of composition and is therefore able to work with high-dimensional data and solve complex problems. It does that by building complex concepts out of multiple simpler concepts and is thus is able to derive a deep representation of features (I.Goodfellow et al., 2016). In the example of an image classification model tasked to identify bird flocks this feature hierarchy could be edges and contours as low-level features, elements such as beaks or wings as next bigger feature level and whole objects, such as birds as the follow-up level (see Figure 2).

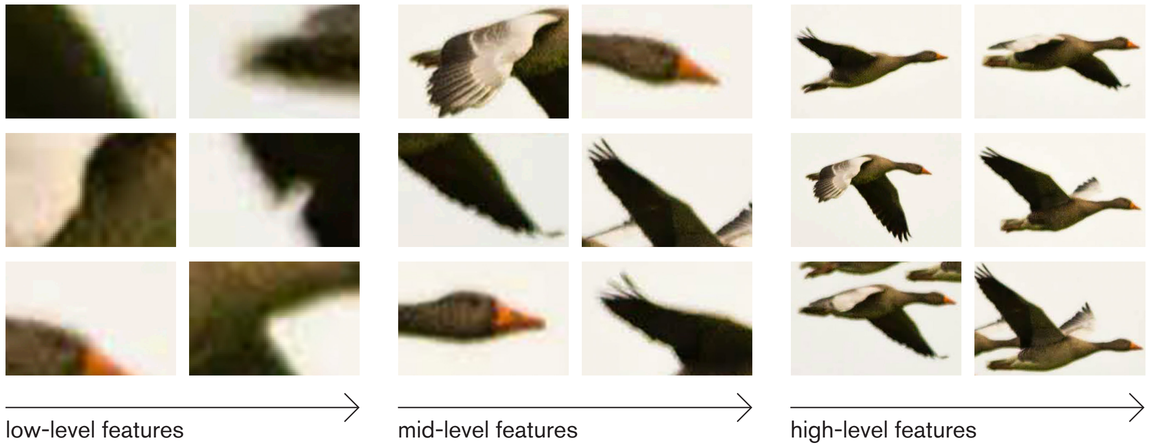


Figure 22: Example of feature hierarchy - based on Photograph by Fahad AlAni

Appx.A.4 Artificial Neural Networks

IIn deep learning, the models utilize artificial neural networks to detect and model hidden patterns in large datasets. These artificial neural networks consist of multiple layers with each being made up with a number of nodes and connections between these nodes (I. Goodfellow et al., 2016). There are different types of neural networks with the most common one being feed-forward neural networks. Here the value of one node is being forwarded along the connections to all nodes in the next layer and adjusted with weights (w) and biases (b) to make up the value of the next node. Both values and bias are trainable: First their value is initialized at random and updated throughout the training process to minimize prediction errors (LeCun et al., 2015). Within feed-forward neural networks there are three different layer types (see Figure 23): The input layer, the hidden layers and the output layer. The initial data is represented in the nodes of the input layer. The hidden layers are all layers between input and output layer. Here, each node takes the weighted (wi) sum of its inputs (xi) from the previous layer, adds a bias term (b), and passes the result through an activation function (f) (see Figure 24) (LeCun et al., 2015). The weights determine the influence of each input on the output, the bias allows the function to more accurately fit patterns that are not centered around the origin and the activation function introduces non-linearity which allows the network to learn about complex, non-linear patterns.

This process allows the network to learn and extract increasingly complex features from the data as it moves through each successive hidden layer. Lastly the output layer represents the desired output, for example a probability value in classification tasks (I. Goodfellow et al., 2016).

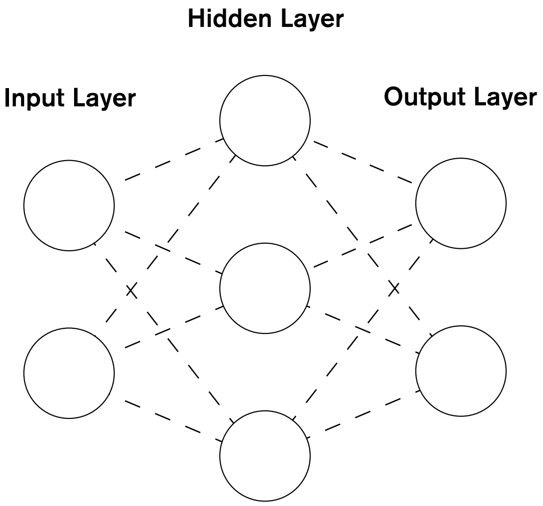


Figure 23: Simplified Neural Network Architecture Based on Husein & Chung, 2019

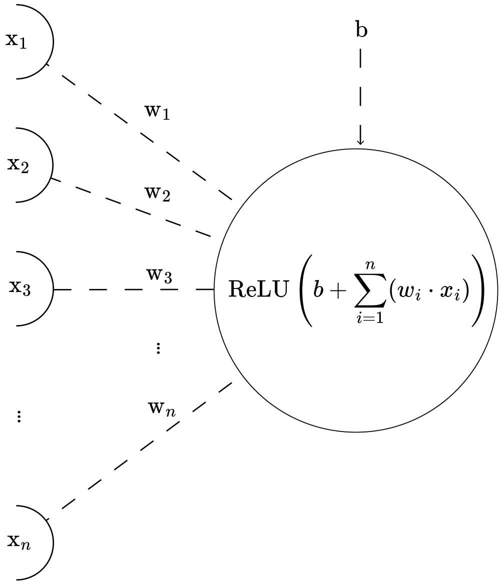

x_1
x_2
x_3
x_n
w_1
w_2
w_3
w_n
b
$\mathrm{ReLU}\left(b + \sum_{i=1}^{n}(w_i \cdot x_i)\right)$

Figure 24: Mechanism of a feed forward layer Based on Badiger & Mathew, 2022

Appx.A.5 Generative Artificial Intelligence

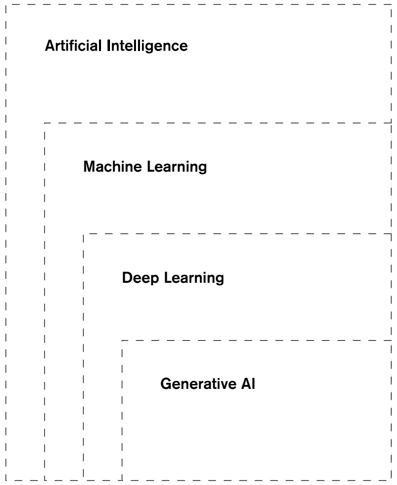


Figure 25: Classification of AI Concepts Based on Goodfellow et al., 2016

Generative Artificial Intelligence (GenAI) describes the generation of miscellaneous data types by a machine learning model, usually a deep learning model (see Figure 25) (Kaswan et al., 2023; Sengar et al., 2024). The diverse content created by GenAI include text, code, images, audio, video and simulations (Akhtar, 2024). Different model types are available with the more commonly found ones being transformer-based models, generative adversarial networks, variational autoencoders and diffusion models (Sengar et al., 2024)

Transformer based models: The input in form of sequential data (e.g. a text) is embedded into a sequence of tokens (e.g. a text). Each token is encoded as a vector, which is then processed through a mechanism known as self-attention. This mechanism enables the model to determine the relative importance of all previously processed tokens in relation to the current token. This allows the model to consider the context of each data item of the sequence (e.g. word) globally, rather than locally. That makes transformer based models excel at tasks that require long-range dependencies (e.g. text translations) (Sengar et al., 2024; Vaswani et al., 2017).

Generative adversarial networks (GANs): GANs consist of two neural networks, a “generator” and a “discriminator”. While the generator creates “fake data” from random noise (e.g. images), the discriminator tries to label the generated output and “real data” examples as fake and real examples. While the generator becomes better at “fooling” the discriminator, the discriminator becomes better at correctly labeling the given data. This allows GANs to create highly realistic outputs that are ideal for tasks requiring high fidelity (e.g. when generating realistic photographs, as seen in Figure 26) (I. J. Goodfellow et al., 2014; Sengar et al., 2024).

Figure 26: Image generated by GAN-based model for the prompt:
„A photograph of a geese flock flying in front of a monotone, light-brown sky“

Variational Autoencoders (VAE): Variational Autoencoders encode the input data into a compressed, lower-dimensional space, called “latent space”. After that, the input is reconstructed from the latent space. Variational Autoencoders assume a probability distribution in the latent space, allowing the generation of new data, by sampling from this distribution. They generate new data in the form of variations of their input.

This makes them suitable for tasks that require the generation of coherent but diverse outputs (e.g. slightly different variations of an object) (Kingma & Welling, 2013; Sengar et al., 2024; Stryker & Bergmann, 2024)
Diffusion Models: Diffusion models work by adding noise to the input data in small, incremental steps until the data is completely transformed into random noise. This process of adding noise makes the original data (e.g., an image) gradually lose its recognizable features. Once the data is reduced to pure noise, the generation process begins. During generation, the model reverses the noise addition process by progressively „denoising“ a random input. Through each step, the model refines the noisy data, gradually reconstructing it into an output that resembles the training data. This iterative process of removing noise allows diffusion models to produce high-quality outputs, particularly when it comes to tasks that require fine details, such as image or video generation (Ho et al., 2020; Sengar et al., 2024).

Since the launch of ChatGPT - a transformer based large language model – by the company OpenAI in 2022, GenAI has experienced a large-scale uptake for both private and commercial purposes (Humlum & Vestergaard, 2024). The widespread adoption and the high potential transformative power of the technology has attracted a lot of attention on GenAI. With the rising demand, also the complexity and scale of GenAI models has drastically increased (Gatla et al., 2024). In the example of OpenAI’s GPT models (GPT ⊠ generative pre-trained transformer model), the parameter count over the models displays the growth in scale. Parameters are internal variables that are learned during training, which govern how the model processes and generates data. GPT-1, released in 2018 contained 117 million parameters and GPT-4, released in 2023 contains 1,76 trillion parameters, displaying an over 1500-times increase in five years only (Gatla et al., 2024).

Appx.A.6 The GPT-3 Architecture

To get a better understanding about the scale and mechanisms of these models it is important to closer examine one. While the various types and models that fall under GenAI have varying architectures and mechanisms as explained above, the goal of this chapter is not the creation of novel insights, but rather the understanding about how an AI model can generate outputs, such as text and what the underlying infrastructure is. Therefore we will only explore one GenAI model: GPT-3 by OpenAI. GPT-3 is a large language model (LLM) launched in 2020 (Brown et al., 2020). The model is a transformer model, a deep learning system with 175 billion of trainable parameters (Brown et al., 2020).

The following process description is based on the foundational paper

“Language Models are Few-Shot Learners” by Brown et al. (2020), which introduced GPT-3:

Step 1

Textual Input: Text is used as the input to for GPT-3. Text is a form of sequential data. To further process this data, the text is split into tokens. Tokens are part of the text sequence (see Figure 27). These tokens are then forwarded into to model after each other. In GPT-3, the maximum number of tokens that the model can process at once (called the context window) is 2048 (Brown et al., 2020), this counts both for the input as well as for the output text. Overall, GPT has a vocabulary size of 50257 words (Brown et al., 2020).

A flock of geese flies in front of a monotonous, light-brown sky

Figure 27: Tokenization of GPT-3
Based on OpenAI (n.d.)

Step 2

Embedding & Positional Encoding: Each token is transformed into a vector. This vector serves as the numerical representation of each token. In GPT-3 each token is converted into a vector with 12288 dimensions (Brown et al., 2020). The space to which the vector refers, contains semantic information; Therefore, the vector representation assigns the token with semantic meaning. Additionally, the vector stores information about the position of each token in the input sequence. This matters as varying token sequences can have different meanings (see Figure 28).

A flock of geese flies in front of a monotonous, light-brown sky

A flock of light-brown geese flies in front of a monotonous sky

Figure 28: Why positional encoding matters

Step 3

Self-Attention: The vector is fed into a series of “attention heads”. In each attention head, a mechanism, called “self-attention” occurs, in which the importance of each token, relative to all other tokens is determined. For every token, it is calculated, how much attention should be paid to all other tokens in the sequence. The output of each attention head is then synthesized into the resulting output of all 96 attention heads (Brown et al., 2020). This allows GPT-3 to consider context beyond the neighboring tokens, which allows for the capturing of long range dependencies and relationships in the input sequence.

Step 4

Feed-Forward Neural Network (FFNN): The output of the Multi Head Attention Layer is passed into a Feed Forward Neural Network. In this step, the vector is processed and transformed further, based on weights and biased which were learned during the training of the model.

Step 5

Transformer Blocks (96x): Together with normalization layers (layers that normalize the result output by the previous layer for higher consistency and stabilization), the self-attention and FFNN make up a transformer block. Overall, GPT-3 uses 96 transformer blocks in a sequence (Brown et al., 2020). The output of each transformer block serves as the input for the next one. This process allows the model to improve its contextual understanding of the input sequence step by step.

Step 6

Decoding: After all the transformer blocks have been passed, the output of them is a 2048 x 12288 matrix. This matrix captures a vector with 12288 dimensions for each of the 2048 output positions. This matrix is transformed into a matrix with a vector of 12288 dimensions over its vocabulary size of 50257 (Radford et al., 2019). In this matrix, all tokes in the vocabulary are assigned a vector which contains a raw, unscaled prediction. A softmax function is used which normalizes each of these into a value of >0 and <1 (see Figure 30). This last value serves as the probability of each token. Depending on the sampling strategy (e.g. choosing the token with the highest probability), the output token gets chosen.

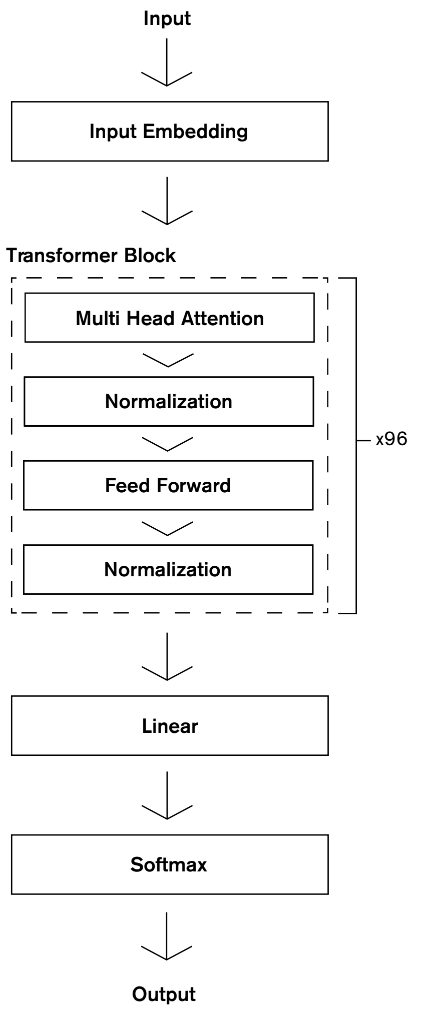


Figure 29: Transformer-model architecture
Adapted from on Vaswani et al., 2017

Step 7 Iterative Generation: After the token has been chosen, it is appended to the input sequence and the model repeats the process. This is repeated until a stop condition is met, such as the end of the 2048 token context space or a token signalizing the end of a sequence.

Step 8 Text Output: Finally, the resulting token sequence is transformed into text, which serves as the overall model output.

Adding together all vectors and matrices over the entire GPT-3 architecture, we derive at a parameter count of around 157 billion. While this showcases the significant scale of the model, GPT-4 is believed to contain around 1,76 trillion parameters (over 11 times the size of GPT-3) and GPT-5 is expected to be even larger. This hints at the scale of computational power required to develop and run such models.

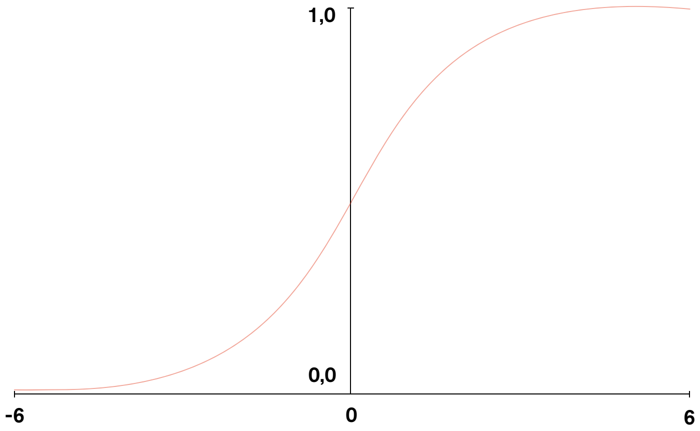


Figure 30: Softmax function

Appx.A.7 The physical infrastructure of GenAI

The high amount of computational power and data required for training and running GenAI models demands a vast hardware infrastructure. This hardware infrastructure must provide computational power, storage solutions for data and the networks between the system components.

Computational power is provided by graphics processing units (GPUs), tensor processing units (TPUs) and sometimes edge devices (Bandi et al., 2023; Zhang et al., 2023).

The large amounts of data that are used by GenAI applications in training, fine-tuning and inference need to be stored. Companies can store their data in internal, company owned data storage facilities. If more flexibility is required, data can be stored through cloud storage, such as via services like Google Cloud, AWS or Azure (George, 2024).

The computing and storage infrastructure for GenAI primarily resides in data centers. In recent years, both the number and size of data centers have grown significantly. As of 2024, there are over 8,000 data centers worldwide, including 5,381 in the United States and 291 in the Netherlands (Cloudscene, 2024). The rise of large-scale workloads is also driving rapid growth in hyperscale data centers — massive facilities, built to handle enormous computing demands. In 2023 there existed 992 of them, which is more than a double compared to five years prior (Synergy Research Group, 2024). Further, hyperscale datacenter capacity is expected to almost triple in the next six years from 2023, driven by AI (Synergy Research Group, 2023).

Running these datacenters requires high amounts of energy and water. Data centers consume such significant resources that they rank among the most energy-intensive building types, surpassing most commercial and industrial facilities in energy use (U.S. Department of Energy, n.d.).

Edge-cloud computing is crucial for three main reasons (Wang et al., 2023): First, the transmission latency in Generative AI (GenAI) applications is substantial due to the vast volumes of data these systems generate. Second, GenAI is currently largely consumer-centric, making localized computational infrastructure more practical than centralized systems. For example, performing finetuning and inference on a company-owned edge device ensures that sensitive company data used in these steps remains within the organization and is not transmitted externally, improving privacy. Lastly, the immense resource requirements of running GenAI models make centralized infrastructure unsustainable and cost-inefficient, particularly due to the high data transmission demands.

Besides storage and computing infrastructure, the hardware consists of networks. The distributed computing in GenAI requires lossless, high-performance and scalable networks which connect all the difference computing resources (such as GPUs in data centers and edge devices) (Dell Technologies, n.d.).

Overall, the scaling of GenAI requires the expansion of hardware infrastructure. This infrastructure consists of computational infrastructure, storage infrastructure and network infrastructure. Vast amounts of resources are demanded for the production and running of these systems. For instance, running GPT-3 is estimated to consume electricity worth $600,000 per day (Wang et al., 2023) a Figure that underscores the operational costs of large AI models and even excludes the high expenses of initial training. At the World's Ecomonic Forum in Davos in January 2024, Sam Altman the CEO of OpenAI, warned, that the next wave of GenAI systems will vastly exceed the expected energy consumption making the AI industry head for an energy crisis (Crawford, 2024).

Discussion

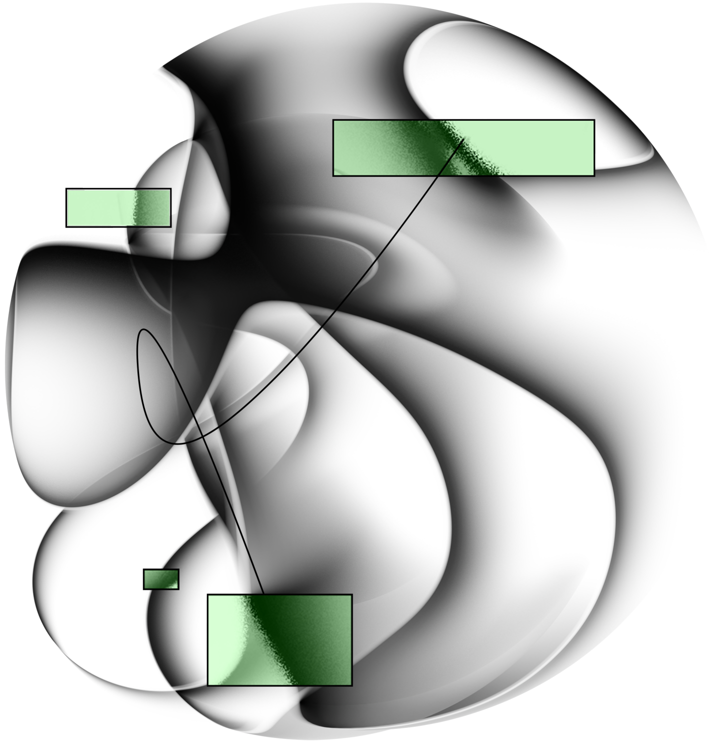
The exploration of GenAI presented in this chapter underscores its complexity and far-reaching implications. GenAI is not a monolithic technology but rather an ecosystem shaped by technical architectures, operational processes, and infrastructural requirements. This multi-layered nature calls for a holistic perspective.

GenAI must be considered as a physical and geological process based on vast amounts of hardware infrastructure. Its operation depends on extensive data centers equipped with specialized, resource intensive hardware such as GPUs and TPUs, whose operation consumes large amounts of energy.
From a technical standpoint, the chapter presented that GenAI models, particularly large language models like GPT-3, rely on intricate architectures with billions of parameters. The size and complexity of the models are experiencing an exponential growth with even larger models being expected in the future. This architectural depth enables GenAI‘s impressive capabilities and potential, but also amplifies its computational demands.

Appendix B: Executive Briefing

Raphael Jung

May 2025



Green GenAI in practice

Executive Briefing

This document is targeted to tech-leaders and executives in firms which develop and deploy GenAI based applications. It provides insights into environmental sustainability strategies for GenAI in the enterprise and is based on an MSc thesis at the TU Delft.

GenAI adoption is driving value creation in enterprises

Generative AI is rapidly transforming the enterprise landscape. **Adoption is accelerating across industries,** reshaping operations, customer engagement and innovation. GenAI enables competitive advantage through personalisation, scale, automation, and cost reduction. The business case is compelling:

7%

increase of the global GDP via GenAI is expected[1]

2,6-4,6

trillion USD are projected to be added annually to the global economy through GenAI [2]

85%

of executives plan to increase their investments in GenAI [3]

… but it comes with significant environmental impacts.

Generative AI requires a significant volume of natural resources. The computational workloads require large amounts of energy to run, the data centres consume water for cooling and the hardware requires critical raw materials. **But this risk goes largely unseen.**

502 t

CO2e were emitted by the training of GPT3 [4]

5,237

Million liter water would have been used if GPT 3 would have been trained in the Netherlands [5]

6 of 9

of the earth systems are negatively impacted by GenAI use [6]

0,5 %

of companies deploying AI try mitigating it's environmental risks [7]

TU Delft

[1] Goldman Sachs. (2023, April 5). *Generative AI could raise global GDP by 7%*. Https://Www.Goldmansachs.Com/Insights/Articles/Generative-Ai-Could-Raise-Global-Gdp-by-7-Percent.Html.

[2] Chui, M., Hazan, E., Roberts, R., Singla, A., Smaje, K., Sukharevsky, A., Yee, L., & Zemmel, R. (2023). *The economic potential of generative AI: The next productivity frontier*. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier

[3] Boston Consulting Group. (2024). From Potential to Profit with GenAI.

[4] Stanford University. (2024, April 15). Emission of CO_2 equivalent by artificial intelligence (AI) models in 2024 (in metric tons) [Graph]. Https://Www.Statista.Com/Statistics/1465353/Total-Co2-Emission-of-Ai-Models/.

[5] Li, P., Yang, J., Islam, M. A., & Ren, S. (2023). Making AI Less 'Thirsty': Uncovering and Addressing the Secret Water Footprint of AI Models. http://arxiv.org/abs/2304.03271

[6] Falk, S., van Wynsberghe, A., & Biber-Freudenberger, L. (2024). The attribution problem of a seemingly intangible industry. *Environmental Challenges*, *16*, 101003. https://doi.org/10.1016/J.ENVC.2024.101003

[7] Chui, M., Hall, B., Singla, A., Sukharevsky, A., & Yee, L. (2023). *The state of AI in 2023: Generative AI's breakout year*. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year

Value Case

Introducing effective sustainability measures and guardrails into your GenAI development projects brings **various benefits to your enterprise.**

Differentiation

Making sustainability guardrails and assessments part of the GenAI offering and tools, can serve as a valuable differentiator to competitors. It increases valued transparency and accountability.

Compliance

Current regulations (e.g. as part of CSRD - if extent of AI use is flagged in double-materiality assessment) and future regulations can demand sustainability guardrails in AI applications. Introducing such measures from early on can reduce the burden coming with new regulations and provide a competitive advantage.

Cost Savings

Reduced resource use means reduced costs for energy and computations (e.g. less tokens used). This becomes especially impactful with large applications and high numbers of users.

Owning Innovation

Developing task specific, streamlined and customised applications, built on small, potentially self-trained models increases the chances of eligibility for patenting and decreases dependance of third parties.

Market Positioning

With sustainability frequently being a strategic pillar of large enterprises, the implementation of Green GenAI measures increases the credibility of the market positioning and strengthens the expertise in this area

Moral Duty

As an implementer of the technology, a company is directly responsible for the resulting, potentially irreversible impacts. It is therefore the duty of the company to manage the resulting impacts carefully

R-Strategies

To this end, **seven strategy types addressing the environmental impacts of GenAI across its development and deployment were identified** through research and validated by leading experts.

Currently R2 Reduce strategies take the centre stage, due to their alignment with cost saving measures, while other strategy types are vastly underused. This **increased efficiency in using a resource can lead to higher overall consumption** of that resource - not less - because efficiency makes its use cheaper and more attractive.

Therefore it is essential to build formations from multiple strategy types.

R0 **Refuse**

The function that GenAI is planned to perform is abandoned or performed by other means - no GenAI is deployed.

Example: A cost-benefit framework to weigh negative environmental impact of GenAI to the business-as-usual

R1 **Reframe**

Reducing the resources required to fulfill a specific use-case, by reframing the project and designing the environment that the GenAI model will be embedded in (focus on strategy, organizational set-up, governance and design).

Example: Introduction of a CO2 budget, to steer all development processes within a project.

R2 **Reduce**

Optimizing the technological processes and mechanisms to reduce the required resources for development and operation of the technology (focus on technological process and mechanisms).

Example: Using adaptive backpropagation as a way to only tune the impactful parameters of a model instead of all parameters in finetuning

R3 **Re-use**

Leveraging preexisting models instead of creating new ones.

Example: Reusing a model in a different context, for example by finetuning it to the new use case

R4 **Release**

Enabling applications that fail to perform their intended function to regain their functionality.

Example: Include mechanisms for automated bug-fixes in the model.

R5 **Revise**

Utilization of components from a preexisting model in the development of a new one.

Example: Using transfer learning, by teaching a smaller model to replicate the behaviour of a pretrained, largermodel and therefor reusing it‘s knowledge structures.

Support

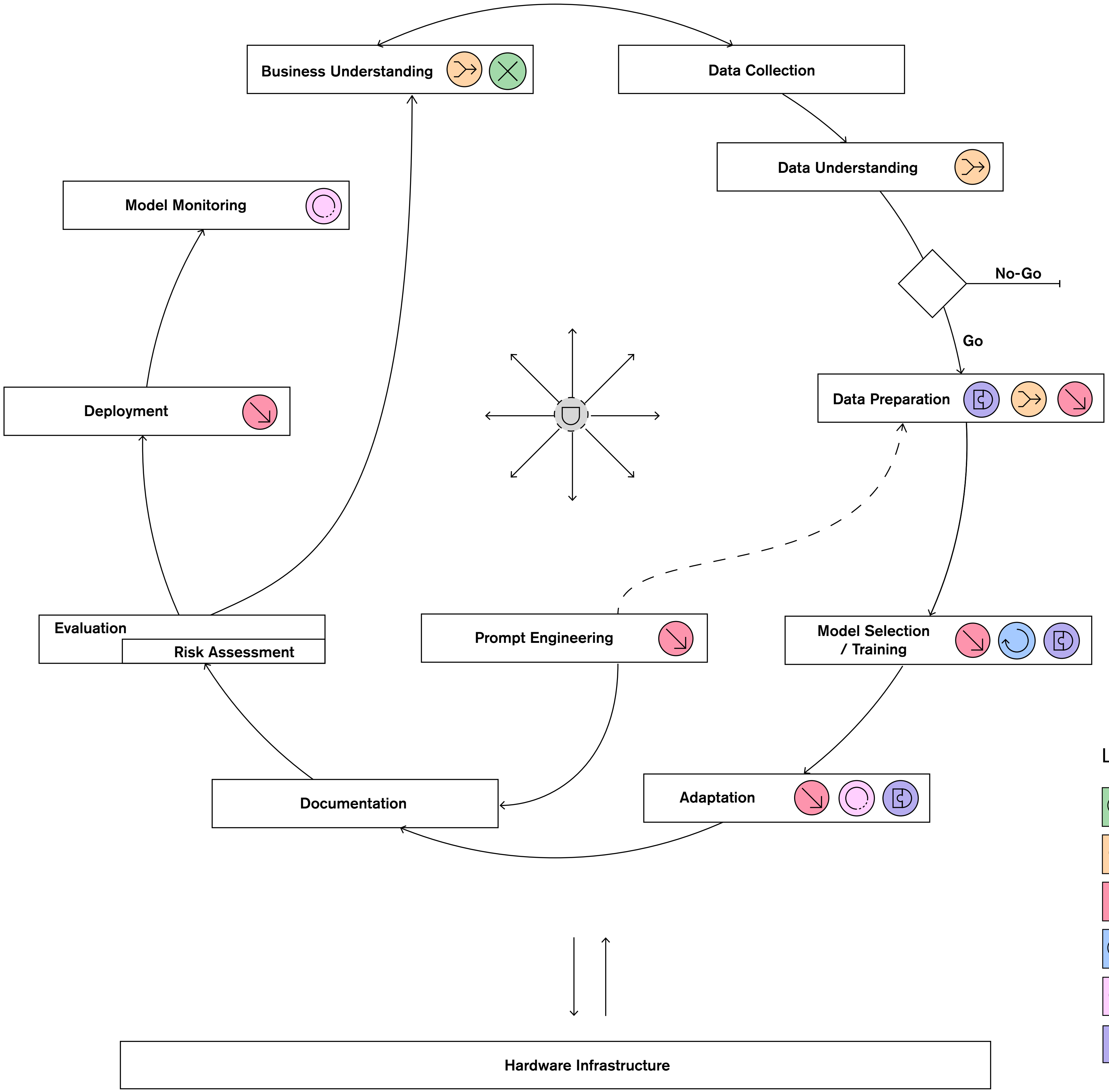
Approaches, that indirectly affect the sustainability of GenAI applications by increasing the acceptance or implementation rate of the other sustainability strategies.

Examples: Reporting of environmental impacts, forming research consortia on sustainable AI, open-sourcing.

Framework

Across the various phases of the GenAI lifecycle the different strategy types can be allocated. This allows to understand, which **strategy types can be applied to which lifecycle stage.**

The proposed framework allows to identify specific approaches in a project across the different strategy types to **create a holistic, multi-layered sustainability strategy.**



Business Understanding

Data Collection

Data Understanding

No-Go

Go

Data Preparation

Model Monitoring

Deployment

Evaluation

Risk Assessment

Prompt Engineering

Model Selection / Training

Documentation

Adaptation

Hardware Infrastructure

Legend

- Ⓧ R0 **Refuse**
- R1 **Reframe**
- R2 **Reduce**
- R3 **Re-use**
- R4 **Release**
- R5 **Revise**
- **Support**

Integration Steps

In order to integrate Green GenAI measures into the development of GenAI based applications, various steps need to be undertaken.

The nature of these steps will depend on the maturity of preexisting AI governance structures and greenIT capabilities. Some enterprises might be able to slightly adjust the responsibilities and nature of preexisting roles and processes, while others need newly introduce them.

To harvest the benefits, it is important, that these measures get introduced as a standardised and mandatory process.

1 Define Roles

Define and assign roles. It must be clear who executes, who advises, and who monitors, depending on the organisational structure. Leverage preexisting roles, e.g. GreenIT experts, AI Ethics Officer or AI Governance Committees, and define clear actions for them.

2 Apply Framework

Directly apply the **R-strategies** (R0–R5) by contextualising them within agile development workflows. Within each sprint, identify the addressed lifecycle stages and through this applicable strategy types. Apply these strategies to your context to derive at an applicable selection of sustainability approaches. Evaluate and implement them, stack them across the sprints to build a multi-layered selection of approaches for each project, across the R-strategies.

3 Assess Impacts

Before key go-/no-go decisions, assess the sustainability risks and impacts of the concept / MVP. While measuring the impact can be difficult, due to limited insights in resource use of cloud providers and software vendors, a robust process to estimate this impact is a good start.

4 Formulate Guardrails

Based on the multi-layered selection of approaches identified, formulate a selection of sustainability guardrails, which will get implemented for the deployment of the application. Ensure, that the guardrails adequately address the impacts identified.

5 Go-/No-Go Points

Introduce a quality gate before the deployment of an application, in which the proposed sustainability guardrails are assessed on their ability to sufficiently address the sustainability impacts identified. If approved, the application gets deployed with the guardrails in place, if not approved, the guardrails get revised or the application adapted.

6 Monitor & Improve

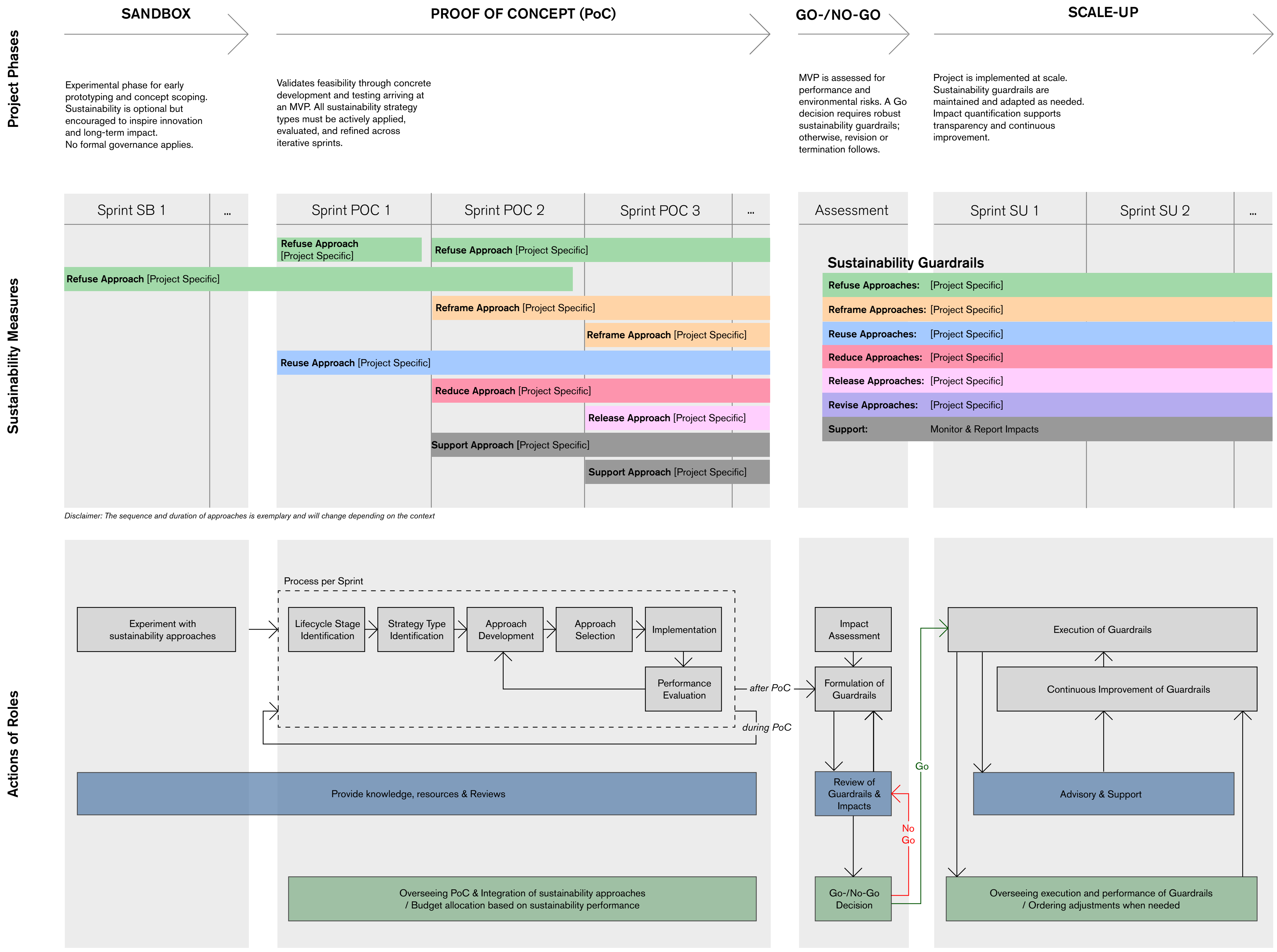
Once deployed, the performance of the application and its environmental impacts are continuously monitored (e.g. estimated emission per user * N° of users, …). The performance is reported and based on the newly created insights the GreenAI governance structure is improved.

Integration Example

This blueprint provides an exemplary overview of how sustainability measures can be integrated into the development of GenAI based applications within the enterprise.

It leverages iterations of approach identification, implementation and evaluation within each sprint to derive at a set of targeted sustainability guardrails, that get evaluated and integrated before deployment

TUDelft



Experimental phase for early prototyping and concept scoping. Sustainability is optional but encouraged to inspire innovation and long-term impact. No formal governance applies.

Validates feasibility through concrete development and testing arriving at an MVP. All sustainability strategy types must be actively applied, evaluated, and refined across iterative sprints.

MVP is assessed for performance and environmental risks. A Go decision requires robust sustainability guardrails; otherwise, revision or termination follows.

Project is implemented at scale. Sustainability guardrails are maintained and adapted as needed. Impact quantification supports transparency and continuous improvement.

Disclaimer: The sequence and duration of approaches is exemplary and will change depending on the context

Project Phases
SANDBOX
PROOF OF CONCEPT (PoC)
GO-/NO-GO
SCALE-UP
Sustainability Measures
Sprint SB 1
...
Sprint POC 1
Sprint POC 2
Sprint POC 3
...
Assessment
Sprint SU 1
Sprint SU 2
...
Refuse Approach [Project Specific]
Refuse Approach [Project Specific]
Refuse Approach [Project Specific]
Reframe Approach [Project Specific]
Reframe Approach [Project Specific]
Reuse Approach [Project Specific]
Reduce Approach [Project Specific]
Release Approach [Project Specific]
Support Approach [Project Specific]
Support Approach [Project Specific]
Sustainability Guardrails
Refuse Approaches: [Project Specific]
Reframe Approaches: [Project Specific]
Reuse Approaches: [Project Specific]
Reduce Approaches: [Project Specific]
Release Approaches: [Project Specific]
Revise Approaches: [Project Specific]
Support: Monitor & Report Impacts
Actions of Roles
Experiment with sustainability approaches
Process per Sprint
Lifecycle Stage Identification
Strategy Type Identification
Approach Development
Approach Selection
Implementation
Performance Evaluation
after PoC
during PoC
Impact Assessment
Formulation of Guardrails
Review of Guardrails & Impacts
Go-/No-Go Decision
Go
No Go
Execution of Guardrails
Continuous Improvement of Guardrails
Advisory & Support
Provide knowledge, resources & Reviews
Overseeing PoC & Integration of sustainability approaches / Budget allocation based on sustainability performance
Overseeing execution and performance of Guardrails / Ordering adjustments when needed

Roles

Owner & Project Teams	Project owners and involved teams, which develop and execute the project. They lead the practical execution of sustainability integration, consisting of the selection, development and implementation of sustainability approaches, development of and compliance to the sustainability guardrails, as well as conducting the impact assessment of their project.
Center of Excellence	The Center of Excellence (CoE) consists of domain experts. It provides methodological guidance, review processes, and support materials to ensure consistent and effective sustainability integration. It supports project teams throughout the process by providing materials, knowledge, support and reviews. Further, the CoE develops materials and best practices to improve the integration of sustianability.
Steering Commitee	The Steering Committee consits of leadership overseeing the trajectory of the project and the sustainability measures. It makes the central decisions in the project, such as the Go-/No-Go decision in which it ensures, that the proposed sustainability guardrails sufficiently adress the environmental risks of the project. During roll-off and scaling, it oversees the execution and performance of sustainability guardrails and can demand a revision.

Key Insights

Environmental sustainability must be embedded in GenAI development from the start

Sustainability needs to move from an afterthought to a core design principle. Only then are sustainability strategies (especially R0 Refuse and R1 Reframe strategies effective.

A sustainability strategy must contain approaches from various strategy types - efficiency alone won't suffice!

While Reduce strategies dominate due to their cost-alignment, they risk triggering rebound effects (Jevons Paradox). True sustainability demands combining efficiency with sufficiency, reuse, and governance-based interventions.

Mapping strategy types to the GenAI lifecycle enables targeted sustainability interventions

By aligning seven distinct strategy types (Refuse–Support) with the GenAI lifecycle, organisations can identify actionable sustainability levers at each stage - from model selection to deployment and infrastructure.

Embedding GreenAI creates clear business value

Sustainability measures enhance regulatory readiness, reduce resource costs, and enable innovation. Enterprises can differentiate their offering, reduce dependencies, and strengthen their ESG credibility - with limited trade-offs.

Environmental impacts of GenAI go far beyond CO_2

Most sustainability discourse focuses on emissions. But GenAI affects six of nine planetary boundaries — including water use, biodiversity loss, and novel pollution — requiring broader sustainability metrics

Most companies ignore GenAI's environmental footprint - at their own risk

Despite bold ESG commitments, only a fraction of firms actively mitigate its environmental impact. This disconnect exposes companies to reputational, regulatory, and strategic blind spots.

Ready to make GenAI sustainable in your organisation? Let’s connect.

Validated in practice

Industry leaders are already putting this framework to use. Here’s what they’re saying:

> “[The framework] incorporates environmental sustainability factors that do not typically appear in AI governing processes. (…) It has helped inform our broad AI governance model for IT leaders, encompassing such critical requirements as upstream and downstream data management, model selection, business case fit-for-purpose, and infrastructure suitability. (…)“
>
> *- Rick Pastore, Principle Researcher of SustainableIT.org & Senior Director of The Hackett Group*

Author

Raphael Jung
Delft University of Technology

Appendix C: Project Brief

Personal Project Brief – IDE Master Graduation Project

Name student Raphael Jung

Student number 5626056

PROJECT TITLE, INTRODUCTION, PROBLEM DEFINITION and ASSIGNMENT

Complete all fields, keep information clear, specific and concise

Project title Reducing the Environmental Impact of GenAI: A Framework

Please state the title of your graduation project (above). Keep the title compact and simple. Do not use abbreviations. The remainder of this document allows you to define and clarify your graduation project.

Introduction

Describe the context of your project here; What is the domain in which your project takes place? Who are the main stakeholders and what interests are at stake? Describe the opportunities (and limitations) in this domain to better serve the stakeholder interests. (max 250 words)

The rapid expansion of generative artificial intelligence (GenAI) has made it a crucial tool for business organizations, with the market projected to grow by 212%, reaching $206.95 billion by 2030 [1]. This growth comes at a time in which the world faces significant environmental challenges, such as climate change and resource depletion [2].

Proponents of AI-driven innovation highlight GenAI's potential to solve complex business problems and improve operations, including contributions to sustainability efforts. Yet, these claims are tempered by increasing concerns from researchers and environmental advocates, who point to the high environmental costs of GenAI technologies [2,3,4,5]. Training and using GenAI models consumes vast amounts of energy, and operating AI infrastructure significantly increases the carbon footprint and resource consumption [2,5].

In the business context, this creates a critical tension: how to balance the operational benefits GenAI offers—like improved efficiency, creativity, and customer services [6,7] - against its environmental impact. The key stakeholders in this debate include AI developers advancing the technology, businesses leveraging it for performance gains, policymakers (such as those behind the EU's AI Act 8) regulating the industry, and society at large, which bears the environmental costs of widespread GenAI adoption.

Despite GenAI's potential to drive sustainable innovation in business, current policies and practices do not adequately address its environmental impact [9,4].

There is a growing need for a clear framework to inform decision-making in the integration and use of GenAI in business organizations, by classifying existing approaches and guiding the development of new ones.

➔ *space available for images / figures on next page*

introduction (continued): space for images

click to add picture

image / figure 1

image / figure 2

Problem Definition

What problem do you want to solve in the context described in the introduction, and within the available time frame of 100 working days? (= Master Graduation Project of 30 EC). What opportunities do you see to create added value for the described stakeholders? Substantiate your choice.
(max 200 words)

The rapid adoption of GenAI in business organizations is transforming operations and driving innovation. However the uptake of this technology results in severe environmental impacts [2], a development that lead to necessity for action. Besides the overarching problem of the sustainability crises, other factors will exert pressure on businesses: Regulatory pressure will likely increase, such as hinted by the American Artificial Intelligence Environmental Impact [10] or the EU's AI Act [8], companies already struggling to meet their scope 3 emission goals [11], and customers switching companies for the benefit of environmental sustainability [11].

With pressure rising, the need for approaches to improve the environmental sustainability of GenAI increases. Businesses currently lack in-depth structures and tools needed to assess and improve the sustainability of their GenAI usage, creating a critical gap between technological progress and environmental responsibility.

This thesis aims to develop a framework that identifies the dimensions in which measures to improve the environmental sustainability of GenAI can be implemented. The framework will then be populated with existing approaches - aimed at enhancing the environmental sustainability of GenAI. The value of such a framework lies in both the collection and contextualization of existing approaches, as well as the identification of gaps and, therefore, the inspiration for novel approaches. This theoretical framework will be supported by a practical tool aimed to be utilised in the context of GenAI

Assignment

This is the most important part of the project brief because it will give a clear direction of what you are heading for. Formulate an assignment to yourself regarding what you expect to deliver as result at the end of your project. (1 sentence) As you graduate as an industrial design engineer, your assignment will start with a verb (Design/Investigate/Validate/Create), and you may use the green text format:

Designing a framework which collects and contextualises approaches for improving the environmental impacts of generative artificial intelligence (GenAI), to inform decision-making in the integration and moderation of GenAI models in business organizations.

Then explain your project approach to carrying out your graduation project and what research and design methods you plan to use to generate your design solution (max 150 words)

After a contextual literature research, I will adapt the preexisting and proven „10R framework“ [12] to the development, implementation and use of GenAI. This will be done by abstracting the dimensions of the existing 10R framework and, in the second step, translating the abstractions to the new case. The resulting framework will be validated through expert interviews.

Once a suitable framework has beed developed, I will conduct a scoping study [13] to populate the developed framework with existing approaches for managing the environmental impacts of GenAI. The resulting populated framework will then both present an overview of existing methods, and showcase knowledge gaps. The value this brings is to both inform the application of GenAI and to inspire the development of future approaches.

Based on the theoretical framework, I will design a practical tool, that fits the context of application within businesses. Context, medium and requirements will be determined based on field research / case studies (depending on resources) and the tool will be tested and validated with stakeholders and experts.

Project planning and key moments

To make visible how you plan to spend your time, you must make a planning for the full project. You are advised to use a Gantt chart format to show the different phases of your project, deliverables you have in mind, meetings and in-between deadlines. Keep in mind that all activities should fit within the given run time of 100 working days. Your planning should include a ***kick-off meeting***, ***mid-term evaluation meeting***, ***green light meeting*** *and* ***graduation ceremony***. *Please indicate periods of part-time activities and/or periods of not spending time on your graduation project, if any (for instance because of holidays or parallel course activities).*

Make sure to attach the full plan to this project brief.
The four key moment dates must be filled in below

Kick off meeting 18.11.2024

Mid-term evaluation 11.02.2025

Green light meeting 28.04.2025

Graduation ceremony 30.05.2025

In exceptional cases (part of) the Graduation Project may need to be scheduled part-time. Indicate here if such applies to your project

Part of project scheduled part-time	✔
For how many project weeks	all
Number of project days per week	4

Comments:
Due to other work besides thesis, I intent to work on 4 days per week on thesis project. (See Project Plan attached)

Motivation and personal ambitions

Explain why you wish to start this project, what competencies you want to prove or develop (e.g. competencies acquired in your MSc programme, electives, extra-curricular activities or other).

Optionally, describe whether you have some personal learning ambitions which you explicitly want to address in this project, on top of the learning objectives of the Graduation Project itself. You might think of e.g. acquiring in depth knowledge on a specific subject, broadening your competencies or experimenting with a specific tool or methodology. Personal learning ambitions are limited to a maximum number of five.
(200 words max)

As sustainability challenges intensify, I am driven to explore how technological progress, particularly Generative AI, can be aligned with environmental responsibility in business organizations. I see this project as an opportunity to design a practical solution that will help companies detect and reduce the environmental impact of their AI systems, addressing a critical gap in current business practices.

Through this project, I aim to develop key competencies in building decision support tools, deepening my understanding of AI technologies, and crafting strategies for sustainable business transformation. My background in sustainability, from my minor in "Designing for Sustainability Transitions" at TU Delft to my Bachelor thesis on resource-sensitive electronics, has equipped me with a solid foundation. In my Master’s program, courses like AI & Society expanded my understanding of AI's role in society, while independent learning in data literacy and Python has enhanced my technical skillset.

I aim to gain deeper insights into business organizations, build expertise in sustainability approaches, sustainable business transformation, and understand the real-world applications of GenAI in corporate settings. This project aligns well with both my professional aspirations and personal ambitions.

References:

1 Statista. (2024, February 9). Global generative AI market size from 2020 to 2030. https://www.statista.com/forecasts/1449838/generative-ai-market-size-worldwide

2. Falk, S., van Wynsberghe, A., & Biber-Freudenberger, L. (2024). The attribution problem of a seemingly intangible industry. *Environmental Challenges*, *16*, 101003. https://doi.org/10.1016/J.ENVC.2024.101003

3 Robbins, S., & van Wynsberghe, A. (2022). Our New Artificial Intelligence Infrastructure: Becoming Locked into an Unsustainable Future. *Sustainability (Switzerland)*, *14*(8). https://doi.org/10.3390/su14084829

4 van Wynsberghe, A. (2021). Sustainable AI: AI for sustainability and the sustainability of AI. *AI and Ethics*, *1*(3), 213–218. https://doi.org/10.1007/s43681-021-00043-6

5 Wu, C.-J., 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., Melnikov, A., Candido, S., Brooks, D., Chauhan, G., Lee, B., Lee, H.-H. S., … Hazelwood, K. (2021). *Sustainable AI: Environmental Implications, Challenges and Opportunities*. http://arxiv.org/abs/2111.00364

6 Chui, M., Robinson, K., & Singla, A. (2023, August 24). *Don't wait—create, with generative AI*. McKinsey & Company. https://www.mckinsey.com/mgi/our-research/dont-wait-create-with-generative-ai

7 Thiruneelakandan, A., & Umamageswari, A. (2024, January 4). *Generative AI: a transformative force in business intelligence*. IEEE Conference Publication I IEEE Xplore. https://ieeexplore.ieee.org/document/10467477

8 Artificial Intelligence Act, COM/2021/206 final. (2023, July 12). Official Journal of the European Union, L 169/1.

9 Bashir, Noman, Priya Donti, James Cuff, Sydney Sroka, Marija Ilic, Vivienne Sze, Christina Delimitrou, and Elsa Olivetti. 2024. "The Climate and Sustainability Implications of Generative AI." An MIT Exploration of Generative AI, March. https://doi.org/10.21428/e4baedd9.9070dfe7.

10 Artificial Intelligence Environmental Impacts Act of 2024, S.3732, 118th Congress (2024). https://www.congress.gov/bill/118th-congress/senate-bill/3732/text

11 Bain & Co. (2024). The Visionary CEO's Guide to Sustainability. Bain. Retrieved September 17, 2024, from https://www.bain.com/insights/topics/ceo-sustainability-guide/

12 Potting, J., Hekkert, M., Worrell, E., Hanemaaijer, A., & PBL Netherlands Environmental Assessment Agency. (2017). *Circular Economy: Measuring innovation in the product chain*. https://www.pbl.nl/sites/default/files/downloads/pbl-2016-circular-economy-measuring-innovation-in-product-chains-2544.pdf

13 Arksey, H., & O'Malley, L. (2005). Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology*, *8*(1), 19–32. https://doi.org/10.1080/1364557032000119616

Conflict of Interest

The author declares no conflict of interest.

This thesis was conducted with support of a company. The company did not influence the results or conclusions presented in this work. All presented views reflect those of the author and involved individuals. The thesis does not claim to reflect the views of the company

Statement on AI use

This thesis benefited from the use of AI tools (e.g. ChatGPT) for language refinement, idea structuring, and feedback generation. All content and conclusions are the author's own, and critical analysis was performed independently.

Raphael Jung
Nr. 5626056

Delft University of Technology
Faculty of Industrial Design Engineering